nature communications

Article

https://doi.org/10.1038/s41467-024-52003-3

Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias

Received: 7 February 2023

Accepted: 22 August 2024

Published online: 29 August 2024

Check for updates

William Lotter [1,2,3]

A core motivation for the use of artificial intelligence (AI) in medicine is to reduce existing healthcare disparities. Yet, recent studies have demonstrated two distinct findings: (1) AI models can show performance biases in underserved populations, and (2) these same models can be directly trained to recognize patient demographics, such as predicting self-reported race from medical images alone. Here, we investigate how these findings may be related, with an end goal of reducing a previously identified underdiagnosis bias. Using two popular chest x-ray datasets, we first demonstrate that technical parameters related to image acquisition and processing influence AI models trained to predict patient race, where these results partly reflect underlying biases in the original clinical datasets. We then find that mitigating the observed differences through a demographics-independent calibration strategy reduces the previously identified bias. While many factors likely contribute to AI bias and demographics prediction, these results highlight the importance of carefully considering data acquisition and processing parameters in AI development and healthcare equity more broadly.

As applications of artificial intelligence (AI) in medicine extend beyond initial research studies to widespread clinical use, ensuring equitable performance across populations is essential. There remains much room for improvement towards this goal, with several studies demonstrating evidence of bias in underserved populations in particular[1–4]. Adjacent recent work has also shown that these same algorithms can be directly trained to recognize patient demographic information[5–7], such as predicting self-reported race from medical images alone[7]. These results are significant because it is unclear how these algorithms identify this information given it is not a task clinicians perform, and critically, it provides further means for the potential for bias[7].

Instead of creating additional risks of bias, a core motivation for the use of AI in healthcare is to reduce disparities that are already known to exist[8–10]. Disparities across different demographic subgroups have been identified in many areas of medicine[11,12], including medical imaging[13,14]. These disparities span the full care continuum, from access to imaging to patient outcomes and even the image acquisition process itself[14]. For instance, in breast cancer screening where disparities have been heavily studied, Black women have a higher breast cancer mortality rate than white women and are less likely to undergo screening mammography at centers with breast imaging specialists[15–17]. Regarding image acquisition, several studies have shown evidence of bias in image and positioning quality and in access to newer breast imaging technology[18–20].

Such disparities in technical data acquisition and processing factors may exist in many imaging domains[14,21–23] and are of particular concern from an AI perspective. AI algorithms have been shown to be

[1]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. [2]Department of Pathology, Brigham & Women's Hospital, Boston, MA, USA. [3]Harvard Medical School, Boston, MA, USA. e-mail: lotterb@ds.dfci.harvard.edu

sensitive to low-level statistics in images and can even learn 'shortcut' connections between irrelevant confounders and disease labels[24–27], thus risking the amplification of statistical biases that may be present in the training data for medical applications. These risks are further exacerbated by the common practice of adapting AI approaches from natural image tasks, which may not fully take advantage of the acquisition and processing parameters unique to medical images. Thus, it is paramount to study the influence of medical image acquisition factors on AI behavior, especially in the context of bias.

Here, the aims of our study were twofold. First, we sought to better understand the factors that influence AI-based prediction of patient race in medical images, focusing specifically on technical aspects related to image acquisition and processing. Second, we aimed to use the knowledge gained to reduce bias in AI diagnostic performance. As a domain which has been heavily studied in both AI performance bias and patient race prediction, we focus on chest X-ray interpretation using two popular public datasets. We first show that AI models are indeed influenced by technical acquisition and processing factors when learning to predict patient race, and this at least partly reflects underlying biases in the original clinical datasets. Based on these findings, we devise two strategies to reduce a previously identified performance bias[1]. We find that a strategy which calibrates the algorithm's score threshold based on the view position of the chest X-ray significantly reduces this bias by upwards of 50%. While there are many sources of potential bias and this strategy alone is not sufficient, these results emphasize the importance of carefully considering image acquisition and processing parameters when developing medical AI algorithms and in achieving healthcare equity more broadly.

Results

An overview of our approach is contained in Fig. 1. We train and analyze two sets of AI models. The first set of models are trained to predict self-reported race based on chest X-ray images (Fig. 1a). We then examine how the predictions of these models change when varying several technical parameters. We use the resulting knowledge to inform the development of a second set of models. This second set is trained to detect 12 types of pathological findings (e.g., pneumonia, fracture, pneumothorax, etc.) and is then evaluated in a binary task of predicting whether or not there are any pathological findings present (Fig. 1b). In this task, Seyyed-Kalantari et al. identified an underdiagnosis bias for underserved populations, where, for instance, Black patients were more likely to have a false negative result by the AI algorithm compared to white patients[1].

The technical factors that we investigate are illustrated in Fig. 1a, which were chosen based on their relevance to chest X-ray imaging and data availability. There are many parameters involved in chest X-ray acquisition and processing, some of which depend on the technologist performing the procedure, some that are automatically set by the X-ray machine, and some that relate to image processing once the radiograph has been acquired. One important set of parameters centers around X-ray exposure, dictating the energy and quantity of X-rays emitted by the machine[28,29]. The appropriate level of exposure and the effects of differing exposures on image statistics such as contrast and noise are complex topics that depend on patient and machine-specific characteristics[28–33]. In modern digital radiography, additional image processing takes place that can compensate for some of these effects, such as 'windowing' the image to help normalize overall brightness and contrast[28,29]. While it is not possible to retrospectively alter the X-ray exposure in the images used here, we can still perform windowing modifications to simulate changes in the image processing and, to some extent, exposure. Here, we specifically explore modifying the window width used in processing the image (Fig. 1a). While subtle, this effectively changes the overall contrast within the image, such as the relative difference in intensity between lung and bone regions.

The other technical factors we explore relate to the positioning of the patient. One important aspect of chest X-ray positioning is the area of the X-ray field relative to the patient's chest[34,35]. During acquisition, this area may be 'collimated' in order to cover the relevant anatomy while limiting unnecessary X-ray exposure to other regions[34–36]. After acquisition, the image may also be 'electronically collimated' via cropping[37,38]. As the image preprocessing used in AI algorithms typically consists of center-cropping and then resizing to a fixed size, it is relatively straightforward to analogously simulate different X-ray field sizes by adjusting these resizing and cropping steps (Fig. 1a, see "Methods"). These adjustments effectively alter the field of view of the image, and this parameter is the second factor we consider. For a third factor, we consider the view position of the chest X-ray. The view position indicates the position of the patient with respect to the X-ray source. Typical view positions used in chest X-rays are anterior-posterior (AP), posterior-anterior (PA), and lateral (Fig. 1a). In addition, the X-ray equipment itself may be a standard, stationary machine or a portable device that can be moved as necessary to image the patient.

In exploring the effects of these technical factors, we train and evaluate AI algorithms using two popular public datasets: CheXpert[39] and MIMIC-CXR[40]. CheXpert (CXP) consists of 224,316 chest X-ray images from 65,240 patients collected from the Stanford Hospital. MIMIC-CXR (MXR) consists of 377,110 chest X-ray images from 65,379 patients collected from the Beth Israel Deaconess Medical Center. We split each dataset randomly at the patient level into training, validation, and testing splits with percentages of 70/10/20%, respectively. In analysis by self-reported race, we consider subgroups of Asian, Black, and white patients following Gichoya et al., who demonstrated that AI models can be trained to recognize self-reported race in both CXP and MXR[7]. We note that these subgroups were used because there is inconsistent and incomplete information regarding ethnicity (e.g., Hispanic or non-Hispanic) in the datasets and insufficient numbers of images from patients of other self-reported races to effectively analyze[7].

Effects of technical factors on AI-based racial identity prediction

We train the first set of AI models to predict self-reported race in each of the CXP and MXR datasets. The models were trained and assessed separately on each dataset to assess the consistency of results across datasets. For model architecture, we use the high-performing convolutional neural network known as DenseNet121[41]. The model was trained to output scores between 0 and 1 for each patient race, indicating the model's confidence that a given image came from a patient of that self-reported race. Consistent with Gichoya et al.[7], the DenseNet121 model achieved high accuracy in predicting patient race in both datasets, with an average area under the receiver operating characteristic curve (AUROC) of 0.918 for CXP and 0.944 for MXR (see Supplementary Table 1).

We next characterized the predictions of the AI-based racial identity prediction models as a function of the described technical factors. For window width and field of view, the AI models were evaluated on copies of the test set that were preprocessed using different parameter values. Figure 2 illustrates how each model's average score per race varies according to these parameters. For CXP, decreasing the window width and field of view increases the model's average score corresponding to white patients and decreases the average score corresponding to Asian and Black patients. In other words, simulating increases in image contrast (decreases in window width) and increases in collimation (decreases in field of view) caused the AI model to predict the image was more likely to come from a white patient on average than an Asian or Black patient. The factors are combinatorial, with changes of $-68.3 \pm 0.6\%$, $-58.0 \pm 1.0\%$, $+33.0 \pm 0.5\%$ in the average Asian, Black, and white prediction scores, respectively, when changing each parameter by 20%. These patterns are similar in the MXR dataset for Asian and white prediction scores (Fig. 2b). However, the model's

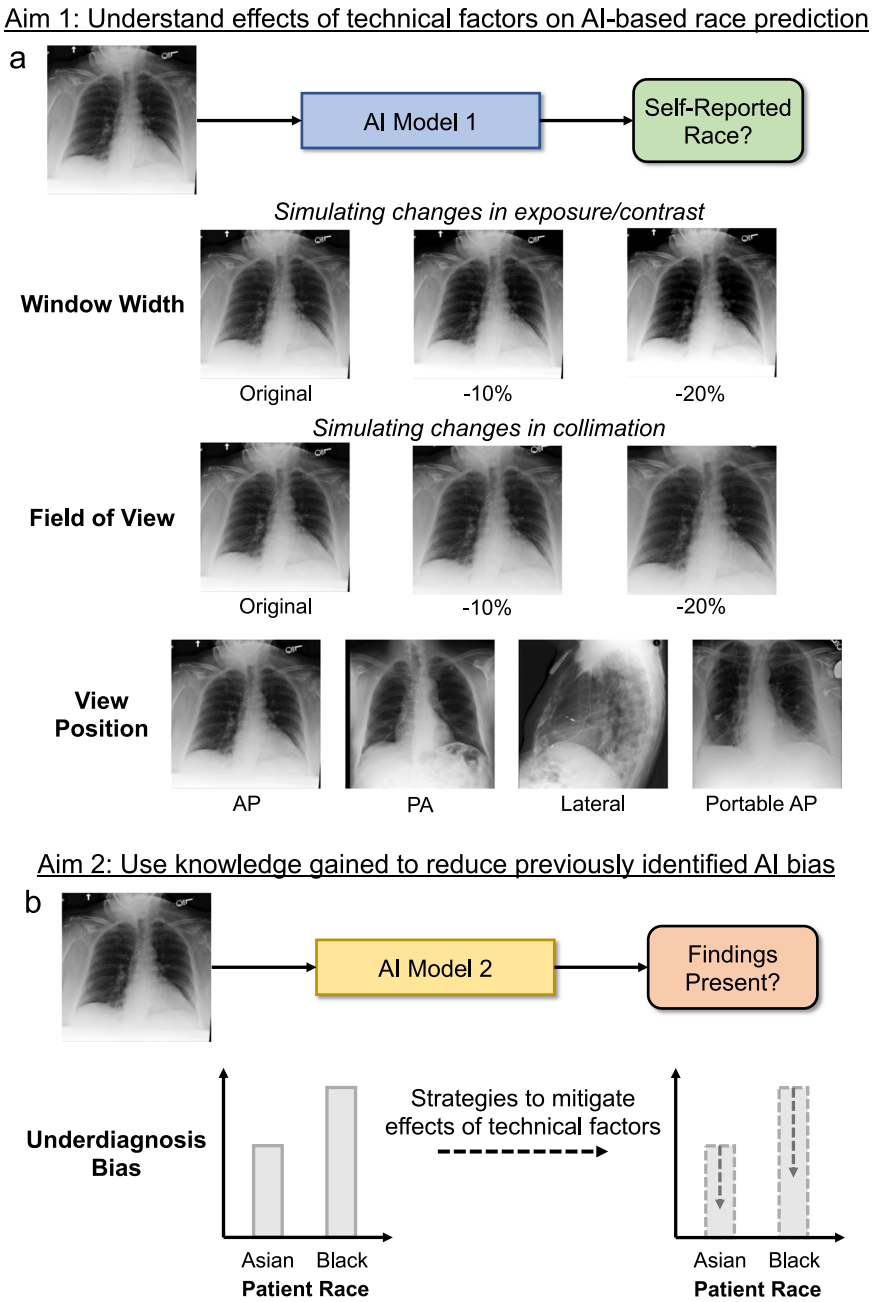

Fig. 1 | Overview of approach. Our study aims to (1) better understand the effects of technical parameters on AI-based racial identity prediction, and (2) use the resulting knowledge to implement strategies to reduce a previously identified AI performance bias. **a** We first train AI models to predict self-reported race in chest X-rays. We then assess how the models' predictions change as a function of factors relating to image acquisition and processing. These factors include the window width, field of view, and view position. **b** We next train AI models to predict the presence of pathological findings, where an underdiagnosis bias for underrepresented patients has been previously identified[1]. Based on the results of the technical factor analysis, we devise strategies with a goal of reducing this bias. PA posterior-anterior, AP anterior-posterior.

score corresponding to Black patients shows a different pattern in MXR, demonstrating much smaller variation by window width and field of view. Thus, while there is some variation across datasets, varying the window width and field of view parameters can generate relatively large changes in the average predictions of the AI model by patient race.

We next assessed the impact of the view position on AI-based race prediction. We quantified effects by comparing the average scores per view to the composite average score across views. Since the view position is a discrete parameter that is available in each dataset, we can additionally compare the per view scores to the empirical prevalence of views for each race. Figure 3 contains the results of this analysis, with the raw view counts per patient race also provided in Supplementary Table 2. We again observe variations in the AI predictions, where the AI models output higher scores on average for certain patient race and view position combinations than others. For instance, both the CXP and MXR models show increased average Asian and Black prediction scores on PA views and a decreased white prediction score. There are also notable effects of the equipment type (standard or portable). In the MXR dataset where this data is available, portable views show an increased average white prediction score but lower average Asian and Black prediction scores. In examining the empirical frequencies per view, we also observe differences by patient race (orange bars in Fig. 3). For instance, Asian and Black patients had relatively higher

https://doi.org/10.1038/s41467-024-52003-3

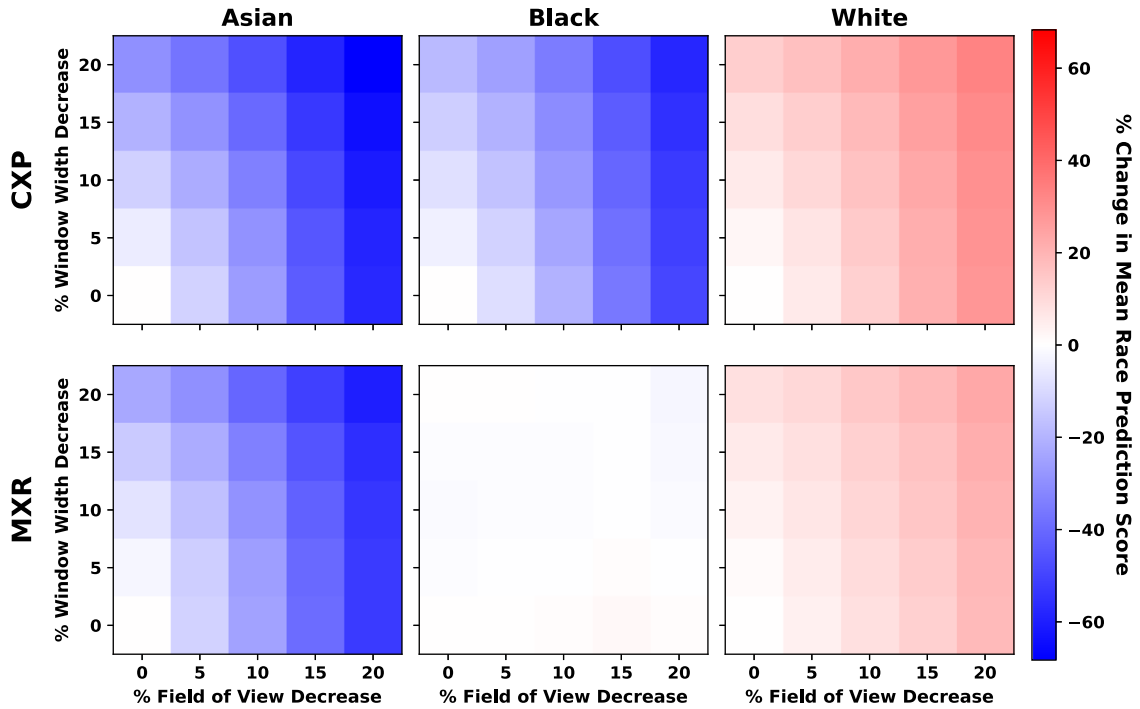

Fig. 2 | Effects of changes in image preprocessing on AI-based racial identity prediction. The average scores of the racial identity prediction model were computed for different window width and field of view values and compared to the default preprocessing used to train the model. The average scores were computed in a weighted fashion to equally weight each patient race across the test dataset (see "Methods"). The percent change in average score per race is plotted. A positive change (red) indicates an increase in the average score for the corresponding race and preprocessing combination across the entire test set. Results are shown separately for the CXP and MXR datasets. Source data are provided as a Source data file.

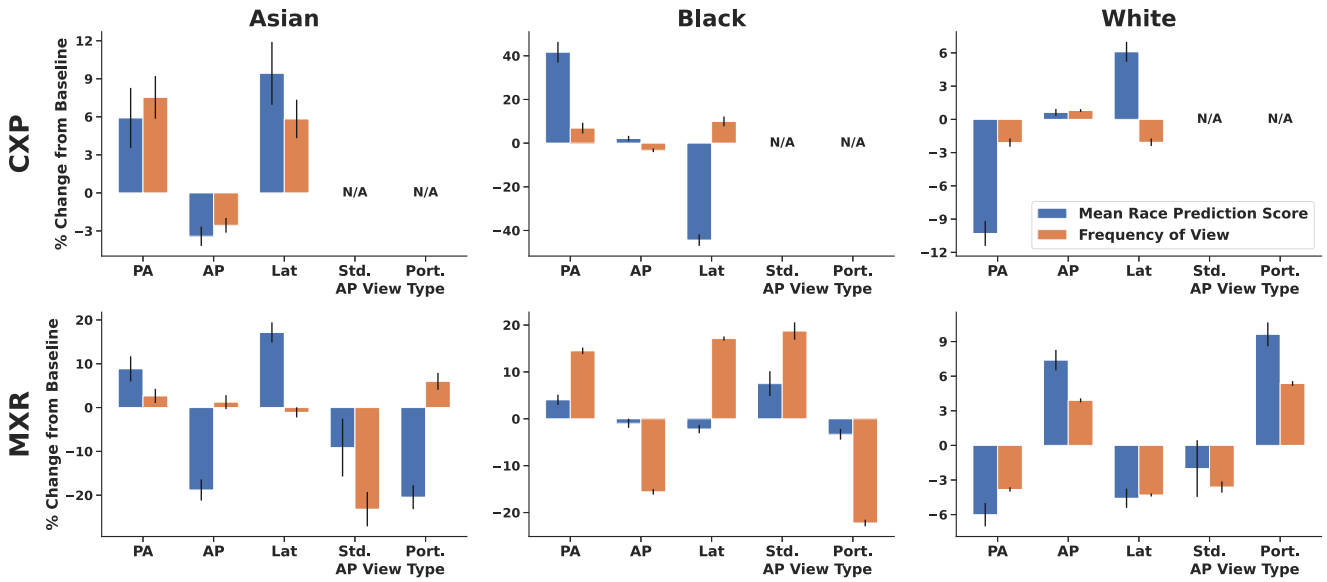

Fig. 3 | Effects of view position on AI-based racial identity prediction. The average scores of the racial identity prediction model were computed for each view position and compared to the average scores across all views. The average scores were computed in a weighted fashion to equally weight each patient race across the test dataset (see "Methods"). The percent change in average score per race is plotted (blue bars) compared to the differences in the frequencies of views by race (orange bars). The view frequencies are also plotted as percent changes compared to the frequency across all views. Results are shown separately for the CXP and MXR datasets. PA posterior-anterior, AP anterior-posterior, Lat lateral, Std standard, Port portable. The results are derived from the following number of images for Asian, Black, and white patients respectively: CXP-PA: 2376, 1202, 11399; CXP-AP: 11780, 5952, 64211; CXP-Lat: 2589, 1369, 12624; MXR-PA: 1904, 10867, 35855; MXR-AP: 3154, 13460, 65068; MXR-Lat: 2322, 14070, 45161; MXR-Std AP: 389, 3077, 9813; MXR-Port AP: 2765, 10383, 55255. Error bars correspond to standard deviation computed via bootstrapping. Source data are provided as a Source data file.

percentages of PA views than white patients in both the CXP and MXR datasets, which is also consistent with the behavior of the AI model for this view. In other words, PA views were relatively more frequent in Asian and Black patients, and the AI model trained to predict patient race was relatively more likely to predict PA images as coming from Asian and Black patients. Out of the 24 possible view-race combinations, 17 (71%) showed patterns in the same direction (i.e., a higher average score and a higher view frequency). Overall, the largest magnitude of differences in both AI score and view frequencies occurred for Black patients. For instance, the average Black prediction score varied by upwards of 40% in the CXP dataset and the difference in view frequencies varied by upwards of 20% in MXR.

Reducing underdiagnosis bias by mitigating bias in technical factors

The technical factor analysis above suggests that certain parameters related to image acquisition and processing significantly influence AI models trained to predict self-reported race from chest X-rays in two popular AI datasets. Given these findings, we next asked if mitigating the observed differences could reduce a previously identified AI bias by developing a second set of AI models. Following Seyyed-Kalantari et al.[1], we trained models to predict the presence of 12 pathologies that are labeled in the CXR and MXR datasets, and then evaluated the models in the binary task of predicting if any pathological findings are present ("Findings Present") or not ("No Findings"). Example findings include pneumonia and pneumothorax, with a full list included in the "Methods". In this task, Seyyed-Kalantari et al. discovered that underserved populations tended to be underdiagnosed by AI algorithms, meaning a lower sensitivity at a fixed operating point. In the context of race, this bias was especially apparent for Black patients in the MXR dataset[1].

We explored two approaches motivated by the results above to reduce the underdiagnosis bias. We specifically sought to develop strategies that were relatively easy to implement, could be adapted to other domains, and did not require knowledge of patient demographics during training or testing. The first approach consists of a data augmentation strategy based on varying the window width and field of view parameters during model training. This strategy aims to create a model that is robust to variations in these factors, for which the race prediction model exhibited patterns across different races. The second approach does not involve changes to model training, but instead to score threshold selection at model inference. As the AI model outputs a continuous score from 0 to 1 for the "No Findings" vs. "Findings Present" task, a threshold must be chosen to generate binary outputs. This threshold is typically chosen based on a target metric and the model's predictions across a validation set. Given the view position results above, we asked if separate thresholds for each view could help mitigate the underdiagnosis bias. The thresholds would again be calculated in the validation set, but separately for each view instead of having one single threshold across all views.

As a baseline approach, we again use a DenseNet121 model architecture that was trained separately on each of the CXP and MXR datasets. The baseline models were trained without data augmentation and used a single score threshold across all views for inference that was chosen to achieve approximately equal sensitivity and specificity in the validation split (see "Methods"). Consistent with Seyyed-Kalantari et al., we find that this baseline approach exhibits an underdiagnosis bias in the MXR dataset for Black and Asian patients compared to white patients[1]. The sensitivity of the model was 80.5 ± 1.4%, 75.8 ± 0.8%, 83.5 ± 0.4% for Asian, Black, and white patients respectively, for a difference of −7.7% (95% CI: −6.2%, −9.3%) between Black and white patients and −3.0% (95% CI: −0.4%, −5.8%) between Asian and white patients. We find that this bias is not as pronounced in the CXP dataset (Supplementary Table 3), which is also consistent with Seyyed-Kalantari et al.[1].

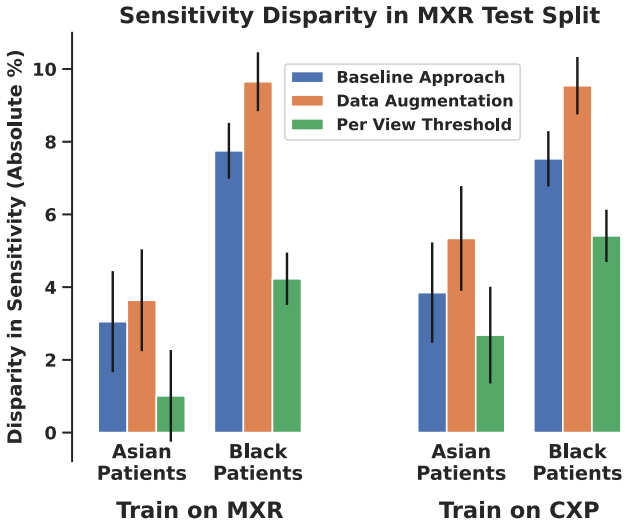

Fig. 4 | Comparison of underdiagnosis bias in MXR test split by AI approach. The disparity in sensitivity of the AI diagnostic model was quantified as the sensitivity of the model for white patients minus the sensitivity of the model for patients of other races. This disparity is plotted by AI approach and training dataset. Error bars correspond to standard deviation computed via bootstrapping and are plotted with respect to the point estimate in the MXR test split. The results are derived from 1992, 10,335, and 38,282 images for Asian, Black, and white patients respectively. Source data are provided as a Source data file.

Having reproduced the previously identified underdiagnosis bias, we next assessed the effects of altering the baseline approach with the two methods proposed above. Figure 4 compares the results for all three approaches on the MXR test split. We include results from a model trained on MXR as well as one trained on CXP to assess robustness. We find that the training-based data augmentation approach does not reduce the underdiagnosis bias. However, the per-view threshold approach reduces this disparity in both Black and Asian patients. For the MXR-trained model, the sensitivity was 82.3 ± 1.3, 79.1 ± 0.7, 83.3 ± 0.4 for Asian, Black, and white patients respectively when using the per-view thresholds, corresponding to an absolute reduction in sensitivity disparity of 3.5% (95% CI: 2.4%, 4.6%; $p < 0.001$) for Black patients and 2.0% (95% CI: −0.04%, 4.2%; $p = 0.03$) for Asian patients. These values represent 46% and 67% relative reductions in disparities for Black and Asian patients, respectively. The CXP model with the per-view thresholds also showed reductions in bias for both Black patients (28% decrease, $p < 0.001$) and Asian patients (29% decrease, $p = 0.08$) despite not being trained on MXR. Importantly, we find that the reduced disparities in sensitivity do not come at the cost of decreased fairness in specificity, as quantified by the variation in specificity across races. For the MXR-trained model, the standard deviation in specificity across races was 3.36 for the per-view threshold approach compared to 3.81 for the baseline approach, for a difference of −0.45 (95% CI: −1.0, 0.26). For the CXP-trained model, the standard deviation was 4.83 for the per-view threshold approach compared to 4.89 for the baseline approach, for a difference of −0.06 (95% CI: −0.71, 0.60).

Analysis of potential confounding factors

Our results above suggest that technical factors related to image acquisition and processing can influence the subgroup behavior of AI models trained on popular chest X-ray datasets. While biases in technical acquisition and image processing parameters have generally been underexplored, recent work has called attention to other potential sources for bias in the CXP and MXR datasets, including differences in distributions of age and disease prevalence by patient race[42–44]. In addition, the versions of the CXP and MXR datasets

used by the AI community consist of JPEG images that were converted and preprocessed from the original DICOM format used in medical practice. While our primary goal is to better understand and mitigate bias of standard AI approaches, it is useful to assess how these potential confounders relate to our observed results. To do so, we have pursued three strategies. For the first strategy, we follow Glocker et al.[42] in creating resampled test sets with approximately equal distributions of age, sex, and disease labels within each race subgroup (see "Methods" and Supplementary Table 4). This strategy aims to control for differences in distributions across these confounders during model testing. For a second strategy, we additionally perform this resampling during model training. Finally, to explore the impact of DICOM conversion and dataset-specific preprocessing, we evaluate on the images extracted directly from the original DICOM files. We specifically perform this evaluation for MXR, as the original DICOM files are publicly available for this dataset but not for CXP. For this strategy and the test set resampling approach, we evaluate the originally trained AI models without modification. The training set resampling approach requires training new models, which we then evaluate on the resampled test sets.

First analyzing the racial identity prediction task, we find that the results for each of the confounder mitigation strategies are consistent with the original findings. While there are slight decreases in the race prediction performance, the performance remains high in all approaches described above (see Supplementary Table 1), with average AUROCs between 0.922–0.934 for MXR and 0.894–0.903 for CXP, compared to 0.944 and 0.918 in the original evaluations for MXR and CXP, respectively. We also find that the window width, field of view, and view position parameters show similar patterns in all conditions, as illustrated in Supplementary Figs. 1 and 2. For both CXP and MXR, test set resampling alone has little effect on the observed results. Combining training and test set resampling leads to more quantitative variation, but the overall trends across these technical parameters remain similar. For instance, in both the original and the training & test resampling results, decreases in the window width and field of view parameters lead to lower Asian prediction scores in CXP and MXR, lower Black predictions scores in CXP, and higher white prediction scores in CXP and MXR. Similarly, there are some quantitative differences when performing the DICOM-based evaluation in MXR, but the core trends are preserved with the models again showing changes in behavior across the factors.

With respect to reducing the underdiagnosis bias, the results are again consistent as the view-specific threshold approach reduces this bias in MXR across all strategies (Supplementary Fig. 3). In the resampled test set, we observe that the overall underdiagnosis bias is lower at baseline, as recently demonstrated by Glocker et al.[42]. Nonetheless, we find that the bias can be further reduced when using the per-view thresholds, with similar results also observed when performing training set resampling. For the DICOM-based evaluation, both the baseline disparity magnitude and its decrease with view-specific thresholds are similar to the original results. Thus, we observe variations in the baseline underdiagnosis bias, but the view-specific threshold approach reduces this bias for each confounder strategy, patient race (Asian and Black), and model training set (CXP and MXR).

Finally, as body mass index (BMI) is a relevant factor in setting X-ray acquisition parameters, we additionally perform the combined training & testing set resampling strategy based on BMI. We perform this experiment using MXR as BMI is available for 39% of this dataset but is not available for CXP. In this MXR subset, we generate resampled training and testing sets to achieve approximately equal distributions of BMI across patient race (see "Methods"). With a lower amount of training data, the performance of the racial identity prediction model decreases, but remains significantly higher than random chance (Supplementary Table 1). Nonetheless, we find that the trends observed across the technical parameters remain, where, for instance, decreases in the window width and field of view parameters still lead to lower Asian prediction scores and higher white prediction scores (Supplementary Figs. 1 and 2). For the diagnostic task, we again find that the per-view threshold strategy reduces the underdiagnosis bias (Supplementary Fig. 3). Thus, while there is some quantitative variation when performing resampling based on BMI, the core patterns are again preserved.

Discussion

Recent important work has demonstrated two distinct findings: (1) AI models trained for medical tasks can show biases in performance for underrepresented populations, and (2) these same models can be trained to directly predict patient demographics like self-reported race. We investigated connections between these two findings with an end goal of reducing a previously identified performance bias. We find that AI models trained to predict self-reported race in chest X-rays from two popular datasets are influenced by several technical factors related to image acquisition and processing. These factors include the view position of the chest X-ray, where we identify disparities by patient race in the original datasets themselves. Through a practical strategy of choosing score thresholds per view, we find that a previously reported underdiagnosis bias in underrepresented populations can be significantly reduced. Altogether, we present a synergistic approach of using AI to elucidate underlying biases in clinical AI datasets to then reduce AI performance bias itself.

While many features may be involved in AI-based racial identity prediction and performance bias[7], including other demographic confounders[42,45], we focused on image acquisition and processing factors for several reasons. First, it is known that biases related to such factors already exist in several medical imaging domains[14,19,20,22,23] and may be more widespread. Therefore, AI models risk learning and perpetuating these biases in the training data. Second, compared to natural images, medical images are more structured and controlled in their acquisition and processing, resulting in rich yet complex metadata. As AI models and even preprocessing techniques are often borrowed from natural image tasks, these technical parameters may be underappreciated and underutilized in AI medical imaging applications. Lastly, we focus on such parameters from a goal-oriented perspective—image preprocessing and the handling of readily available parameters can be adjusted during AI development and deployment. Thus, these parameters offer a means for mitigating bias from an AI standpoint. As such, our goal was not to elucidate all of the features enabling AI-based race prediction, but instead focus on those that could lead to straightforward AI strategies for reducing AI diagnostic performance bias. To this end, our analysis is not intended to advocate for the removal of the ability to predict race from medical images, rather to better understand potential technical dataset factors that influence this behavior and improve AI diagnostic fairness.

Through this approach, we identified that AI models trained to recognize race in chest X-rays exhibit significant changes in predictions by the view position of the X-ray and by image preprocessing parameters related to contrast/exposure and the field of view. As the view position is a discrete, interpretable parameter, it is straightforward to compare the behavior of the AI model by this parameter to its empirical statistics in the dataset. We indeed find differences in the relative frequencies of views across races in both the CXP and MXR datasets. Overall, the largest discrepancies were observed for Black patients in the MXR dataset, which also corresponds to where the largest AI-based underdiagnosis bias was observed. These differences in view proportions are problematic from an AI development perspective, in part because the AI model may learn shortcut connections between the view type and the presence of pathological findings[24,25]. Indeed, we do find that AI models trained to predict pathological findings exhibit different score distributions for different views

(Supplementary Fig. 4). This observation can help explain why choosing score thresholds per view can help mitigate the underdiagnosis bias. We note, however, that this strategy did not completely eliminate the performance bias, leaving room for improvement in future work. Furthermore, it is important to consider both sensitivity and specificity when calibrating score distributions and assessing overall performance and fairness[42,46–48]. Calibration and the generalization of fairness metrics across datasets is indeed an unsolved, general challenge in AI regardless of how thresholds are chosen[49] (see also Supplementary Fig. 5).

In contrast to the score threshold strategy, we did not find that a training-based data augmentation strategy reduced the underdiagnosis bias. This strategy involved randomly applying different window width and field of view parameters to images during training, designed to make the AI model more robust to these parameters. Though the race prediction models exhibited changes in predicted race over these parameters, this strategy did not translate to lower underdiagnosis bias. There are several reasons why this may be the case. The intra-race variation across these parameters may already be sufficiently larger than the inter-race variation, or perhaps the data augmentation approach or its implementation were simply not effective. It is also possible that these parameters influence the race prediction models but are not the main drivers of bias in the diagnostic models. The fact that the data augmentation approach did not help, and actually seemed to slightly increase the underdiagnosis bias, does raise an important question of whether current standard data augmentation techniques have any contribution to AI bias. We also note that it is much more challenging to assess the "true" underlying distribution of the factors represented by the window width and field of view parameters. While altering the window width was designed to mimic changes in contrast and exposure[28,29,50,51], it is an imperfect simulation such as not precisely capturing higher signal-to-noise ratios that result from higher exposures and does not cleanly map to a single physical value. The field of view parameter is also an imperfect simulation of changing the collimation and relative size of the X-ray field with respect to the patient. Nonetheless, the fact that the race prediction model did show differences in predictions over these parameters does suggest that it may have learned intrinsic patterns in the underlying datasets (Supplementary Fig. 6).

Our work adds to the growing attention towards better understanding the underlying causes of AI bias and behavior across protected subgroups[1,2,7,8,42,45,52]. In the current context, it has been suggested that factors ranging from demographic confounders to label bias[42–44] could contribute to the performance differences observed by Seyyed-Kalantari et al.[1]. In fact, in important concurrent work, Glocker et al.[42] proposed several strategies for exploring this behavior, including the use of test set resampling to better control for demographic and prevalence shifts amongst racial subgroups. The authors found that this resampling reduced racial performance differences in CXP and MXR, suggesting that these factors (e.g., age, disease prevalence) may at least partially underlie the previously observed bias. We observe similar results when performing this resampling, where, interestingly, we find that using view-specific thresholds may be synergistic with this resampling to reduce the bias even further. Importantly, our view-specific threshold approach operates in a demographics and disease-independent fashion, providing a practical strategy for real-world use. Furthermore, we find that the effects of the window width, field of view, and view position parameters are present even when performing training and test set resampling, suggesting that these effects are not simply the product of age, sex, disease label, or BMI shifts alone. We also examined whether the specific preprocessing used to create the "AI-ready" MXR dataset can explain our findings by evaluating on the images extracted directly from their original DICOM format. We again observe similar results across the racial identity prediction and underdiagnosis analyses. This perspective of AI-ready vs. source data does raise important considerations, however, such as ensuring that commonly used image preprocessing techniques (e.g., normalization) are optimized to perform consistently across populations and data characteristics. The precise delineation of what is considered image (pre)processing is also unclear when considering the full path from initial X-ray exposure through to input to an AI model or even presentation on a viewing workstation for clinician review. Beyond the image preprocessing used to create AI-ready datasets, the optimal way to generate "ground-truth" labels is an important open question in terms of both overall diagnostic performance and fairness, where natural language processing (NLP)-based extraction of labels from clinical records as performed for the studied datasets offers enhanced scalability but also room for label noise and bias[42]. In addition to the use of NLP-based labeling, the CXP and MXR datasets have several known and possibly unknown limitations[24,43,44], including limited overall diversity in patient race/ethnicity and possible hidden differences in disease severity across subgroups.

While we focused on studying differences in technical factors from an AI perspective, understanding how these differences arise to begin with is a critical area of research. The differences in view position utilization rates observed here are qualitatively similar to recent findings of different utilization rates of thoracic imaging by patient race[21–23,53]. As different views and machine types (e.g., fixed or portable) may be used for different procedures and patient conditions, it is important to understand if the observed differences underlie larger disparities. The effects regarding the other preprocessing parameters are more challenging to directly compare to clinical practice given the complexity of the X-ray acquisition process and its relationship to statistical image properties. While controlling for age, sex, disease prevalence, and BMI did not resolve these effects, there may be other unmeasured population shifts or hidden biases in the studied datasets that contribute to the findings. Thus, as our analysis and conclusions focus on AI efforts using popular datasets, they should not be interpreted as directly informing how X-ray acquisition should be done in the clinic. Nonetheless, as optimal patient positioning and X-ray exposure parameters depend on many patient-specific factors[28–30,34,50], where some of these parameters are set by the technologist and some are set by the machine itself, it is important to consider which populations these settings are optimized for and if the effects observed here have any relationship to image and/or positioning quality. Indeed, the subject of X-ray dosage and race has a complex and controversial history[54].

We studied bias in the context of chest X-rays and certain racial subgroups given important recent work in this domain and the popularity of the studied AI tasks and datasets, but we envision that similar analysis can fruitfully be applied to other domains and other demographic subgroups. This analysis emphasizes the importance of carefully considering technical acquisition and processing parameters, but also the importance of carefully choosing score thresholds. Threshold selection involves optimizing a tradeoff between sensitivity and specificity, and it is critical to understand the factors that influence score distributions and ultimately this tradeoff. Altogether, a detail-oriented approach is necessary towards the effective and equitable integration of AI systems in clinical practice.

Methods

Ethics statement

This study utilized two public chest X-ray datasets, CheXpert[39] and MIMIC-CXR[40], which are de-identified in accordance with the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) Safe Harbor requirements. The study is classified as not-human subjects research as determined by the Dana-Farber/Harvard Cancer Center Institutional Review Board.

Dataset descriptions

CheXpert (CXP) consists of 224,316 chest X-ray images from 65,240 patients collected from Stanford Hospital. The exams were performed between 2002-2017 in both inpatient and outpatient settings. MIMIC-CXR (MXR) consists of 377,110 chest X-ray images from 65,379 patients collected from the Beth Israel Deaconess Medical Center (BIDMC). The MIMIC-CXR dataset was constructed by first querying the BIDMC electronic health record (EHR) to obtain a list of patients who received a chest radiograph in the emergency department from 2011 to 2016. All chest radiographs available in the BIDMC Radiology Information System (RIS) for this set of patients from 2011 to 2016 were then retrieved.

Both the CXP and MXR datasets consist of chest X-ray images in a JPEG format with accompanying de-identified patient metadata. The metadata includes patient demographic information and labels for each image corresponding to 14 different types of observations that were automatically parsed from radiology reports. The observations consist of 12 pathological findings (Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture) as well as "No Finding" and "Support Devices". In the CXP dataset, race and ethnicity are available as separate data elements, whereas in the MXR dataset, they are combined together in one element. For AI-based racial identity prediction, we follow Gichoya et al. in including self-reported Asian, Black, and white patients given this inconsistency in handling ethnicity between the two datasets and the insufficient number of studies from patients of other races to effectively analyze[7]. The percentages of race by patient in CXP are 63.6% white, 12.2% Asian, 5.4% Black, and 18.8% other race or unknown. At an X-ray image level, the percentages are 63.8% white, 11.8% Asian, 6.1% Black, and 18.3% other race or unknown. For MXR, the patient percentages are 67.2% white, 17.4% Black, 3.8% Asian, and 11.6% other or unknown; with X-ray image percentages of 68.4% white, 17.4% Black, 3.5% Asian, and 10.6% other or unknown. Race was self-reported in both datasets and is missing or unknown for 12% and 15% of X-rays in the CXP and MXR datasets respectively. For the underdiagnosis analysis, we follow Seyyed-Kalantari et al. in training the AI models using all available labels and then evaluating binary classification performance based on the "No Finding" label[1]. For both datasets, each X-ray image is assigned one of four possible values for each label class: 1 (positive for that class), 0 (negative for that class), −1 (uncertain for that class), or nan (missing). To ensure robust training and testing, we treated uncertain labels (−1) as missing. We additionally populated missing values whenever possible based on the following rules: (a) if a pathology finding was present and the "No Finding" label was missing, the "No Finding" label was set to 0; (b) if the "No Finding" label was 1, we set each of the labels corresponding to pathologies to 0. A breakdown of the view position counts for each dataset is contained in Supplementary Table 2. We note that the distinction between standard and portable views is only available in the MXR dataset. A view was determined to have come from a portable machine if the text "port" was present in the "Performed Procedure Step Description". For portable views, the AP position is almost always used. In CXP, 94.1% of X-ray images are labeled as having a support device present (93.9%, 94.3%, and 94.0% for Asian, Black, and white patients, respectively). In MXR, 93.8% of X-ray images are labeled as having a support device present (91.9%, 93.3%, and 94.2% for Asian, Black, and white patients, respectively).

Each dataset was split randomly into training, validation, and testing with percentages of 70/10/20%. The splits were created at the patient level such that all X-rays from a given patient were included in the same split. The training split was used for AI model training. The validation split was used for model selection during training and for the calculation of score thresholds. The testing split was only used for model testing.

AI model development

A DenseNet121[41] architecture was used for both the race prediction and underdiagnosis analyses. For the diagnostic task, weights were initialized from ImageNet[55] pre-training, whereas the weights were randomly initialized for the race prediction task to limit any implicit assumptions about the features used for this task. Models were trained separately on each dataset for each task in order to assess the consistency of results across datasets. The Adam optimizer[56] with a learning rate of 1e−3 and a weight decay of 1e−5 was used for all model training. Models were trained for 40 epochs with the final weights selected based on AUROC performance in the validation set. Image preprocessing consisted of the following steps in order: center cropping to a square image, resizing to 224 × 224 pixels, and normalizing to a range of −1024 to 1024. Data augmentation was not performed for the race prediction task to again limit any implicit assumptions of the features used for this task. For the diagnostic task, no data augmentation was used for the baseline approach. For the data augmentation approach, augmentation was used during training by varying the window width and field of view parameters by up to 20%. These factors are explained in more detail below.

Technical factor analysis

The technical factor analysis consisted of assessing the influence of three parameters on the racial identity prediction models. The window width and field of view parameters involved changing the preprocessing of input images, whereas the view position involved subgroup analysis by this parameter without changing the preprocessing. For the window width and field of view parameters, the changes in preprocessing were performed on the 8-bit JPEG images from the datasets, except for the DICOM confounder analysis (detailed below). A linear mapping was used for windowing with a fixed window center of 128 and a varying window width. Explicitly, for a given original image im_o and a window width of w, the resulting windowed image im_w was computed using the following function: $\mathrm{im}_w = 225 \times (\frac{\mathrm{im}_o - 128}{w} + 0.5)$. To simulate changes in contrast and exposure, the window width was decreased by increments of 5% up to 20%. For modifications of the field of view parameter, an additional step was added to the default preprocessing process. As described above, the default preprocessing consisted of center cropping to a square image, resizing to 224 × 224, followed by normalization. Changing the field of view by a scale $k \in (1, 1.2]$ consisted of: center cropping to a square image, resizing to a size of $(224k) \times (224k)$, center cropping to 224 × 224, followed by normalization. Values of $k<1$ were not considered as doing so would involve unnatural padding to the image.

The effects of altering the window width and field of view parameters were quantified in terms of the percent change in average prediction score compared to the original images. For a given combination of window width and field of view, the racial identity prediction model was run on each image in the test set to produce three scores per image (corresponding to Asian, Black, and white). An average score across all images was then computed for each of the three outputs, where this average was computed in an inverse weighted fashion by patient race based on the empirical proportions of each patient race in the test set. This weighting was performed to balance the contribution of images from each race in the results. The results presented in Fig. 2 then represent the percent change in average prediction scores per race for each preprocessing combination compared to the original processing.

The effects of view position were quantified in a similar fashion by comparing the average racial identity prediction scores for each view position compared to the average scores across all views. Figure 3 additionally compares these values to differences in the empirical frequencies of the view positions across patient race. Namely, for each view position, the proportions of patient race across images with that view position were compared to the patient race proportions across

the entire dataset. This difference was then quantified as a percent change, enabling a normalized comparison to the score changes per view. As an example, if 10% of images in the dataset came from Black patients, whereas 15% of Lateral views are from Black patients, this would correspond to a 50% relative increase.

Underdiagnosis bias analysis

Following Seyyed-Kalantari et al., we evaluate the diagnostic AI models on the binary task of classifying if findings are present using the "No Findings" label available in each dataset[1]. As the AI model outputs a continuous 0–1 score, a threshold must be chosen to binarize the model's outputs, which is described further below. Once the model's outputs have been binarized, the underdiagnosis bias can be assessed by quantifying differences in sensitivity between patient races. Sensitivity is defined as the percentage of chest X-rays with findings that are identified as such by the AI model, whereas specificity is defined as the percentage of chest X-rays with no findings that are identified as such. The underdiagnosis bias identified by Seyyed-Kalantari et al. and reproduced here manifests in a higher sensitivity for white patients than for Asian and Black patients[1].

For score threshold selection, we targeted a 'balanced' threshold computed to achieve approximately equal sensitivity and specificity in the validation set. Such a selection strategy is invariant to the empirical prevalence of findings in the dataset used to choose the threshold, allowing more consistent comparisons across datasets and different subgroups. For the baseline model, a single score threshold was used per model. For the per-view threshold strategy, a separate threshold was computed for each view position. To facilitate consistency in selection criteria across views, the threshold for each view was chosen to target the same sensitivity in the validation split, namely the sensitivity of the balanced threshold across all views. At inference time, the threshold used for a given image then corresponds to the threshold for the view position of that image. In CXP, the view positions consisted of PA, AP, and Lateral; whereas the AP view was treated separately for portable and non-portable views in MXR as this information is available in MXR.

Confounder analysis

The confounder analysis included three strategies: test set resampling, training set resampling, and DICOM-based evaluation. For the test set resampling, we follow Glocker et al.[42] by resampling the original test sets with replacement to create more similar distributions of age, sex, and disease labels across patient race. Specifically, we implement a hierarchical sampling strategy that first samples a possible value independently for each confounder, and then randomly samples an image to be included in the test set with this combination of values. To do so, we first compute the aggregate empirical distribution of each factor in the dataset. For age, we create bins in increments of 10 years from 30 to 80 years. For each diagnosis class (including "No Findings"), we compute the percentage of images with a "1" label and consider all other labels as "0" for these sampling purposes. We then sample an image to be included in the resampled test set as follows: (1) sample a patient race with equal probability, (2) sample a patient age bin based on the aggregate probability distribution, (3) sample a patient sex based on the aggregate probability distribution, (4) sample a diagnosis class (out of the 14 included in the dataset) with equal probability, (5) sample a 0 or 1 label for this diagnosis class using the aggregate probability distribution for this class, and (6) sample an image from a study where each of the sampled factors are met. This process is performed until a resampled test set is created that is three times the size of the original test set to help reduce sampling variation. Ultimately, this process aims to create a test set that is balanced across patient race where the distributions of age, sex, and disease labels are approximately equivalent for each patient race. We note that the described diagnosis sampling process was chosen in order to equally consider all diagnosis classes in the sampling and because of the challenging combinatorial problem of jointly considering multiple diagnosis classes at once. For training set resampling, we perform a similar process but perform the sampling 'on-the-fly' when creating each training batch and sample patient race based on the empirical proportions in each dataset, as the balanced sampling would lead to severe overfitting in underrepresented patient races. We evaluate the resulting models on the resampled test sets. For MXR, we additionally perform the training & testing resampling procedure separately based on BMI. We first query the MIMIC-IV database[57] using the "chartevents" table to obtain height and weight for the patients in MXR and use these values to compute BMI, with values available for patients corresponding to 39% of images. For resampling, we bin BMI values according to the World Health Organization (WHO) classifications of underweight (<18.5), normal weight (18.5–25), overweight (25–30), and obese (≥30). As BMI is not available for CXP and missing for many patients in MXR, we perform the BMI-based resampling analysis separately from the analysis by age, sex, and disease prevalence.

For the DICOM-based evaluation, we use the same list of images as the original MXR test set but extract the pixel data from the corresponding DICOM files instead of using the preprocessed JPEG files. We restrict this evaluation to MXR because the original DICOM files are not publicly available for CXP. When evaluating the AI models on the DICOM images, we first extract and process the pixel data according to the DICOM Standard[58] using code based on the pydicom library[59]. This processing includes using the default windowing parameters in the DICOM header, as would be done by standard DICOM viewers. When performing the window width experiments (i.e., Supplementary Fig. 1), we modify this process by changing the window width from its default value by increments of 5%. Ultimately, after the pixel data is processed from the DICOM images using the described steps, the originally trained AI models are evaluated in a similar fashion to the JPEG-based images, including using the same AI preprocessing steps described in the "AI model development" section above.

Statistical analysis

Confidence intervals and standard deviations for AUROC were computed via the Delong method[60]. All other confidence intervals, standard deviations, and p-values were computed via bootstrapping with 2000 samples. The p-values for the underdiagnosis analysis are one-sided. All analysis was performed on a per-image basis. Throughout the text, '95% CI' was used when representing the 95% confidence interval and '±' was used when representing standard deviation.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The two chest X-ray datasets used in this study are publicly available. The CheXpert[39] dataset can be obtained after signing a data use agreement by following the instructions at https://stanfordmlgroup.github.io/competitions/chexpert/. The MIMIC-CXR[40] dataset can be obtained after signing a data use agreement and completing a credentialing process by following the instructions at https://physionet.org/content/mimic-cxr/2.0.0/. For the MIMIC-CXR dataset, the patient race information can be obtained via the admissions table in the MIMIC-IV[57] dataset: https://physionet.org/content/mimiciv/2.2/. Source data are provided with this paper.

Code availability

All analyses were performed using the Python programming language (version 3.9). The AI models were trained using Pytorch[61] and a modified version of the TorchXRayVision library[62]. The model training and analysis code is available at https://github.com/lotterlab/cxr_tech_bias.

References

1. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**, 2176–2182 (2021).
2. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
3. Hsu, W. et al. External validation of an ensemble model for automated mammography interpretation by artificial intelligence. *JAMA Netw. Open* **5**, e2242343 (2022).
4. Daneshjou, R. et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci. Adv.* **8**, eabq6147 (2022).
5. Yi, P. H. et al. Radiology 'forensics': determination of age and sex from chest radiographs using deep learning. *Emerg. Radiol.* **28**, 949–954 (2021).
6. Rim, T. H. et al. Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. *Lancet Digit Health* **2**, e526–e536 (2020).
7. Gichoya, J. W. et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health* **4**, e406–e414 (2022).
8. Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S. & Obermeyer, Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat. Med.* **27**, 136–140 (2021).
9. Chen, I. Y., Joshi, S. & Ghassemi, M. Treating health disparities with artificial intelligence. *Nat. Med.* **26**, 16–17 (2020).
10. Chen, I. Y., Szolovits, P. & Ghassemi, M. Can AI help reduce disparities in general medical and mental health care? *AMA J. Ethics* **21**, E167–E179 (2019).
11. Williams, D. R. & Mohammed, S. A. Discrimination and racial disparities in health: evidence and needed research. *J. Behav. Med.* **32**, 20–47 (2009).
12. Fiscella, K. & Sanders, M. R. Racial and ethnic disparities in the quality of health care. *Annu. Rev. Public Health* **37**, 375–394 (2016).
13. Betancourt, J. R., Tan-McGrory, A. & Flores, E. & López, D. Racial and ethnic disparities in radiology: a call to action. *J. Am. Coll. Radiol.* **16**, 547–553 (2019).
14. Waite, S., Scott, J. & Colombo, D. Narrowing the gap: imaging disparities in radiology. *Radiology* **299**, 27–35 (2021).
15. Richardson, L. C., Henley, S. J., Miller, J. W., Massetti, G. & Thomas, C. C. Patterns and trends in age-specific black-white differences in breast cancer incidence and mortality - United States, 1999–2014. *MMWR Morb. Mortal. Wkly Rep.* **65**, 1093–1098 (2016).
16. Rauscher, G. H., Allgood, K. L., Whitman, S. & Conant, E. Disparities in screening mammography services by race/ethnicity and health insurance. *J. Women's. Health* **21**, 154–160 (2012).
17. Rauscher, G. H., Khan, J. A., Berbaum, M. L. & Conant, E. F. Potentially missed detection with screening mammography: does the quality of radiologist's interpretation vary by patient socioeconomic advantage/disadvantage? *Ann. Epidemiol.* **23**, 210–214 (2013).
18. Rauscher, G. H., Conant, E. F., Khan, J. A. & Berbaum, M. L. Mammogram image quality as a potential contributor to disparities in breast cancer stage at diagnosis: an observational study. *BMC Cancer* **13**, 208 (2013).
19. Miles, R. C., Onega, T. & Lee, C. I. Addressing potential health disparities in the adoption of advanced breast imaging technologies. *Acad. Radiol.* **25**, 547–551 (2018).
20. Christensen, E. W. et al. Relationship between race and access to newer mammographic technology in women with medicare insurance. *Radiology* **306**, 221153 (2022).
21. Schrager, J. D. et al. Racial and ethnic differences in diagnostic imaging utilization during adult emergency department visits in the United States, 2005 to 2014. *J. Am. Coll. Radiol.* **16**, 1036–1045 (2019).
22. Ross, A. B. et al. Racial and/or ethnic disparities in the use of imaging: results from the 2015 National Health Interview Survey. *Radiology* **302**, 140–142 (2022).
23. Ross, A. B., Kalia, V., Chan, B. Y. & Li, G. The influence of patient race on the use of diagnostic imaging in United States emergency departments: data from the National Hospital Ambulatory Medical Care survey. *BMC Health Serv. Res.* **20**, 840 (2020).
24. Jabbour, S., Fouhey, D., Kazerooni, E., Sjoding, M. W. & Wiens, J. Deep learning applied to chest X-rays: exploiting and preventing shortcuts. In *Machine Learning for Healthcare Conference* Vol. 126, 750–782 (2020).
25. DeGrave, A. J., Janizek, J. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).
26. Geirhos, R. et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference for Learning Representations* (2019).
27. Zech, J. R. et al. Confounding variables can degrade generalization performance of radiological deep learning models. *PLoS Med.* **15**, 1–17 (2019).
28. Williams, M. B. et al. Digital radiography image quality: image acquisition. *J. Am. Coll. Radiol.* **4**, 371–388 (2007).
29. Huda, W. & Abrahams, R. B. Radiographic techniques, contrast, and noise in X-ray imaging. *Am. J. Roentgenol.* **204**, W126–W131 (2015).
30. Uffmann, M. & Schaefer-Prokop, C. Digital radiography: the balance between image quality and required radiation dose. *Eur. J. Radiol.* **72**, 202–208 (2009).
31. Al-Murshedi, S., Hogg, P. & England, A. Relationship between body habitus and image quality and radiation dose in chest X-ray examinations: A phantom study. *Phys. Med.* **57**, 65–71 (2019).
32. Shepard, S. J. et al. An exposure indicator for digital radiography: AAPM Task Group 116 (executive summary). *Med. Phys.* **36**, 2898–2914 (2009).
33. Seibert, J. A. & Morin, R. L. The standardized exposure index for digital radiography: an opportunity for optimization of radiation dose to the pediatric population. *Pediatr. Radiol.* **41**, 573–581 (2011).
34. Tschauner, S. et al. European Guidelines for AP/PA chest X-rays: routinely satisfiable in a paediatric radiology division? *Eur. Radiol.* **26**, 495–505 (2016).
35. Whaley, J. S. et al. Investigation of the variability in the assessment of digital chest X-ray image quality. *J. Digit. Imaging* **26**, 217–226 (2013).
36. Fauber, T. L. & Dempsey, M. C. X-ray field size and patient dosimetry. *Radiol. Technol.* **85**, 155–161 (2013).
37. Tsalafoutas, I. A. Electronic collimation of radiographic images: does it comprise an overexposure risk? *Br. J. Radiol.* **91**, 20170958 (2018).
38. Bomer, J. et al. Electronic collimation and radiation protection in paediatric digital radiography: revival of the silver lining. *Insights Imaging* **4**, 723–727 (2013).
39. Irvin, J. et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2019).
40. Johnson, A. E. W. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
41. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017).
42. Glocker, B. et al. Algorithmic encoding of protected characteristics in chest X-ray disease detection models. *eBioMedicine* **89**, 104467 (2023).
43. Bernhardt, M. et al. Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. *Nat. Med.* **28**, 1157–1158 (2022).

44. Mukherjee, P. et al. Confounding factors need to be accounted for in assessing bias by machine learning algorithms. *Nat. Med.* **28**, 1159–1160 (2022).
45. Duffy, G. et al. Confounders mediate AI prediction of demographics in medical imaging. *NPJ Digit. Med.* **5**, 188 (2022).
46. Pleiss, G. et al. On fairness and calibration. In *NeurIPS* (2017).
47. Kleinberg, J., Mullainathan, S. & Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)* (2017).
48. Zhang, H. et al. Improving the Fairness of Chest X-ray Classifiers. In *Proceedings of the Conference on Health, Inference, and Learning* (eds. Flores, G., Chen, G. H., Pollard, T., Ho, J. C. & Naumann, T.) Vol. 174, 204–233 (PMLR, 2022).
49. Schrouff, J. et al. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. In *NeurIPS* (2022).
50. Ching, W., Robinson, J. & McEntee, M. Patient-based radiographic exposure factor selection: a systematic review. *J. Med Radiat. Sci.* **61**, 176–190 (2014).
51. Veldkamp, W. J. H., Kroft, L. J. M. & Geleijns, J. Dose and perceived image quality in chest radiography. *Eur. J. Radiol.* **72**, 209–217 (2009).
52. Wu, E. et al. Explaining medical AI performance disparities across sites with confounder Shapley value analysis. In *Machine Learning for Health (ML4H)* (2021).
53. Narayan, A. K. et al. Racial and ethnic disparities in lung cancer screening eligibility. *Radiology* **301**, 712–720 (2021).
54. Bavli, I. & Jones, D. S. Race correction and the X-ray machine—the controversy over increased radiation doses for Black Americans in 1968. *NEJM* **387**, 947–952 (2022).
55. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009).
56. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *International Conference for Learning Representations* (2015).
57. Johnson, A. E. W. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1–9 (2023).
58. National Electrical Manufacturers Association. NEMA PS3/ISO 12052, Digital Imaging and Communications in Medicine (DICOM) Standard (2024).
59. Mason, D. L. et al. pydicom: an open source DICOM library. https://github.com/pydicom/pydicom [Online] (2023).
60. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
61. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *NeurIPS* (2019).
62. Cohen, J. P. et al. TorchXRayVision: a library of chest X-ray datasets and models. In *Medical Imaging with Deep Learning* (2020).

Acknowledgements

We thank Eliezer Van Allen, Katharina Hoebel, and Stephanie McNamara for thoughtful feedback.

Author contributions

W.L. conceived and performed the study and wrote the manuscript.

Competing interests

The author declares no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52003-3.

Correspondence and requests for materials should be addressed to William Lotter.

Peer review information *Nature Communications* thanks Judy Gichoya, Pritam Mukherjee and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.