nature portfolio

Peer Review File

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

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have attempted to provide an explanation for an ongoing question of significant importance where we don't understand the underlying mechanism. They have explored an area where previous work has not demonstrated any explanation for the mechanism behind why deep learning models can identify demographics and why there is persistent underdiagnosis for minority groups.

However, there are several concerns with their methodology and presentation of results. Firstly, the data set used to evaluate their hypothesis is flawed. The use of MIMIC chest X-ray and CheXpert datasets, which are released in JPEG format rather than the original DICOM format, makes it difficult to assess their results and conclusions. This is because preprocessing has already been performed on these images – and there is known variation of minor differences of even 8 bit versus 16 bit.

Moreover, the processing approaches suggested by the authors include zooming and windowing. These would not typically be considered image preprocessing but rather factors that are usually changed on the view/display by the end user – usually a radiologist . Therefore, their selection of preprocessing tasks is misleading.

Secondly, there is concern for sensitivity to the labels that were provided in their data. For example, if you look at extended data table two in MIMIC chest X-ray dataset, most images are acquired using portable method as expected for ICU images. However, when you look at distribution between AP and lateral views quite a large number of lateral views in ICU which would not make sense. This may indicate a need for further cleaning of datasets by authors.

I recommend that the authors try a different modality such as brain MRI to demonstrate this process of preprocessing and underdiagnosis. The reason for this recommendation is because there is more harmonization and standardization of brain imaging. If there is a desire to work on chest X-rays, there is an opportunity to access and curate a dataset to make sure that it's applicable for this task.

There are also other minor comments including some papers referenced in their work have been published but an archive link is provided instead (e.g., reference 27). I recommend that the authors go through their references and make sure they're referencing the most recent publication.

In terms of information presentation, I thought figures two and three were very dense and difficult to follow. I recommend better visualization of results.

While I do want to commend the authors for mentioning that training-based data augmentation did not reduce underdiagnosis (and increased it), presenting an area that can be further researched if we can understand the robustness of underlying methods proposed by authors; overall my recommendation is that the paper is flawed in its methodology

Reviewer #2 (Remarks to the Author):

The paper posits that acquisition parameters such as the field of view, exposure, and view (AP vs PA vs lateral) may influence the detection of race by AI on chest xray images, and possibly related race-based biases such as the underdiagnosis bias reported in the literature previously. They use windowing intensity levels to stand in for exposure and varying the cropping (during preprocessing) to simulate field of view. There are two key results claimed in the paper: 1. Increasing the contrast (by decreasing the window width) and reducing the field of view tended to push the output probability scores (for classifying white vs black vs asian) towards the dominant class: white; PA views tended to decrease the prediction probability for the "white" class. 2. Based on the above observation, using different decision thresholds based on different views can reduce underdiagnosis bias; using data augmentation based on random windowing or cropping did not reduce the bias.

This is certainly interesting work on an important topic. However, I feel there are major concerns about the validity of the work, which I enumerate below:

This work builds upon and replicates the results presented in reference [1]. Unfortunately, several important factors were not accounted for in the analysis of [1] as pointed out by multiple "matters arising" responses to that paper, (see https://www.nature.com/articles/s41591-022-01847-7, https://www.nature.com/articles/s41591-022-01846-8). The current work also fails to account for multiple confounding effects which raises doubts about the validity of the current work. A key paper that talks about these factors is [a]

[a] Algorithmic encoding of protected characteristics in chest X-ray disease detection models
Glocker, Ben et al. eBioMedicine, Volume 89, 104467
https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(23)00032-4/fulltext

This reference is missing in the current paper. I strongly suggest the authors refer to and discuss their current paper in the context of [a]. Let me also elaborate further on the potential pitfalls of the current work:
1. the intensity/contrast of a chest x-ray images depends on the disease and the severity of the disease. Thus, to discern the impact of intensity/contrast on race identification, one needs to control for the distribution of diseases across the different races.
2. another confounding factor is the age distribution: younger patients may have less severe presentation of a disease (higher contrast in x-ray images, perhaps) and if there are more younger patients in a certain subgroup during training, the model will be likely to prefer that subgroup with higher contrast images. Without controlling for such confounding effects, it is very hard to interpret the effect of windowing/contrast (or exposure).
3. similar confounding factors may exist for the field of view: younger patients, shorter patients, very sick patients may have larger effective fields of view.

I strongly urge the authors to carefully think about the potential confounding effects and account for them in their analysis -- in its present form I am not convinced about the validity of the presented results.

Response to Reviewers
We thank the Reviewers for their comments, which have helped us improve our manuscript. Below, we provide a point-by-point reply to these comments and detail corresponding changes in the manuscript. The original Reviewer comments are *italicized*, followed by our reply.

Reviewer #1 (Remarks to the Author):
The authors have attempted to provide an explanation for an ongoing question of significant importance where we don't understand the underlying mechanism. They have explored an area where previous work has not demonstrated any explanation for the mechanism behind why deep learning models can identify demographics and why there is persistent underdiagnosis for minority groups.

However, there are several concerns with their methodology and presentation of results. Firstly, the data set used to evaluate their hypothesis is flawed. The use of MIMIC chest X-ray and CheXpert datasets, which are released in JPEG format rather than the original DICOM format, makes it difficult to assess their results and conclusions. This is because preprocessing has already been performed on these images – and there is known variation of minor differences of even 8 bit versus 16 bit.

Response: We thank the Reviewer for expressing this point. While our primary goal is to better understand and help mitigate bias of standard AI approaches and datasets, which rely on preprocessed images in an "AI-ready" format, we agree that it is useful to understand if this preprocessing itself causes our observed results. Thus, we have now also included evaluation on the original DICOM images for the MIMIC dataset. While the JPEG version is typically in AI work, these DICOM images are publicly available for MIMIC, though not for CheXpert. We have specifically tested our original models, which were trained on the JPEG images, directly on the MIMIC DICOM images to more robustly test the generalization and sensitivity of our results. Following the DICOM Standard (National Electrical Manufacturers Association, VA, USA), we extract and process the MIMIC images directly from the DICOM files using the default parameters contained in the DICOM headers before AI evaluation. This analysis is contained in a new section in the manuscript titled "Analysis of potential confounding factors" with accompanying new figures in the Extended Data Figures 1-3. Despite the models having been trained on images with different preprocessing, we find that the overall patterns regarding the technical parameter analysis and underdiagnosis bias remain when testing on the DICOM images. For the racial identity prediction task, the AI predictions are still influenced by the technical parameters, and on the disease classification task, the baseline model still shows underdiagnosis bias which is reduced by the view-specific threshold approach. These analyses suggest that these results are not simply caused by the preprocessing used to create the AI-ready datasets. Nonetheless, we agree that this preprocessing is an important consideration and we have expanded on this consideration in the Discussion (lines 342-350).

Moreover, the processing approaches suggested by the authors include zooming and windowing. These would not typically be considered image preprocessing but rather factors that *are usually changed on the view/display by the end user – usually a radiologist . Therefore, their selection of preprocessing tasks is misleading.*

Response: Thank you for raising this important point. We agree that it is challenging to define a precise delineation of what is considered image "preprocessing". Related to the previous point, there are a number of processing stages that take place from initial x-ray exposure through to when the image is actually viewed by the end user (i.e., radiologist) or even AI model. For instance, most DICOM viewers implement image processing steps according to the DICOM Standard by default, including windowing according to the default values in the DICOM header before display to the end user, which can then be adjusted as desired as rightfully stated in the comment above. Similarly, preprocessing stages such as windowing, normalization, resizing, and even bit depth conversion are commonly used in AI approaches. Ultimately, our goal was to simulate technical variations that are important in chest x-ray image acquisition and processing, including overall contrast/exposure and the relative size of the field of view, which can be changed through collimation. To study the effect of contrast, we perform windowing with different window widths to produce different levels of overall contrast. To study variations in collimation, we effectively perform 'electronic collimation' (Tsalafoutas et al., reference 37; Bomer et al., reference 38) to modify the relative field of view for the image. We had used the term "zoom" as an intuitive, non-technical way to explain this phenomenon, but we recognize that this term can have different meanings in different contexts and can cause confusion. Thus, we now refer to this aspect as the "field of view" parameter instead of the "zoom" parameter and have updated its use throughout the manuscript. Altogether, we believe that the window width and field of view parameters represent highly relevant transformations for chest x-ray acquisition and preprocessing, especially in the context of AI development. In addition to updating the field of view terminology, we have now expanded upon these points in lines 70-71, 87-88, 342-350.

Secondly, there is concern for sensitivity to the labels that were provided in their data. For example, if you look at extended data table two in MIMIC chest X-ray dataset, most images are acquired using portable method as expected for ICU images. However, when you look at distribution between AP and lateral views quite a large number of lateral views in ICU which would not make sense. This may indicate a need for further cleaning of datasets by authors.

Response: It is an important point that dataset cleaning is critical for AI applications. According to the initial paper describing the MIMIC dataset (Johnson et al., 2019), the authors state the following: "We queried the BIDMC EHR for chest radiograph studies acquired in the emergency department between 2011–2016, and extracted only the set of patient identifiers associated with these studies. A collection of images associated with a single report is referred to as a study, identified by a unique identifier, the study ID. We then extracted all chest radiographs and radiology reports available in the RIS for this set of patients between 2011–2016."

Thus, while the set of patients was identified based on studies acquired in the emergency department, it is our interpretation that all studies for these patients were then extracted regardless of whether they originated from the emergency department or not. We have now expanded on the description of the MIMIC dataset in the Methods section (lines 493-497) to more fully reflect the original description of the dataset. Additionally, we realize that our original presentation of Extended Data Table 2 may be confusing in that we displayed the breakdown of Standard vs Portable views for the AP view position, since this position is used for the vast majority of Portable views (as stated in the comment above). For improved clarity, we have added an additional row in this table showing the total proportion of Portable views amongst all images (33.8%) and updated the table legend with further description.

We have additionally pursued further data cleaning using three strategies to ensure that the AP and Lateral views are properly labeled as the Reviewer suggests. First, we plotted 100 random AP views and 100 random Lateral views according to the MIMIC metadata. We manually reviewed these 200 images and determined that they are all correctly labeled (e.g., there were no Lateral views that were labeled as AP and vice versa). Second, we extracted the View Position directly from the DICOM files that were obtained for the prior comment. We compared these extracted views to the metadata and there was a 100% correspondence. Third, we reviewed the code used by the dataset creators to create this dataset, which is publicly available on Github. We did not find any apparent issues through this review. While some amount of noise is likely to be expected in any clinical dataset, we believe that these efforts support the validity and proper curation of the MIMIC dataset, which is further supported by its popularity and the initial publication describing its curation.

I recommend that the authors try a different modality such as brain MRI to demonstrate this process of preprocessing and underdiagnosis. The reason for this recommendation is because there is more harmonization and standardization of brain imaging. If there is a desire to work on chest X-rays, there is an opportunity to access and curate a dataset to make sure that it's applicable for this task.

Response: We agree that assessing AI bias and sources thereof are important directions in all medical imaging domains. Given recent high-attention work in AI-based racial identity prediction and performance bias in chest x-rays, we focused on better understanding and addressing these findings rather than identifying similar biases in other domains. With this focus, we have thus aimed to ensure proper dataset curation as suggested by the Reviewer in this comment and the comments above. These important comments raised pertain to the use of DICOM images and sufficient curation of view information. We have now performed evaluation directly from the DICOM images and ensured proper curation of view information in the MIMIC dataset, as described above. We also note that identifying preprocessing and underdiagnosis effects in a less standardized modality could actually be a “feature” and not a “bug” in that it points to the potential of improved harmonization and standardization that could potentially reduce these issues.

There are also other minor comments including some papers referenced in their work have been published but an archive link is provided instead (e.g., reference 27). I recommend that the authors go through their references and make sure they're referencing the most recent publication.

Response: Thank you for pointing this out. We have now reviewed all references and updated to the most recent publications we could find, including removing arXiv links when appropriate.

In terms of information presentation, I thought figures two and three were very dense and difficult to follow. I recommend better visualization of results.

Response: Thank you for this feedback. We have now updated Figures 2 and 3 to help improve interpretability, including increasing spacing and font sizes and better harmonizing the presentation of results across CXP and MXR.

While I do want to commend the authors for mentioning that training-based data augmentation did not reduce underdiagnosis (and increased it), presenting an area that can be further researched if we can understand the robustness of underlying methods proposed by authors; overall my recommendation is that the paper is flawed in its methodology

Response: Thank you for your thoughtful consideration of our manuscript and for mentioning the significance of the underlying problem and potential implications of our results. We hope that our updated manuscript supports the robustness of our findings.

Reviewer #2 (Remarks to the Author):
The paper posits that acquisition parameters such as the field of view, exposure, and view (AP vs PA vs lateral) may influence the detection of race by AI on chest xray images, and possibly related race-based biases such as the underdiagnosis bias reported in the literature previously. They use windowing intensity levels to stand in for exposure and varying the cropping (during preprocessing) to simulate field of view. There are two key results claimed in the paper: 1. Increasing the contrast (by decreasing the window width) and reducing the field of view tended to push the output probability scores (for classifying white vs black vs asian) towards the dominant class: white; PA views tended to decrease the prediction probability for the "white" class. 2. Based on the above observation, using different decision thresholds based on different views can reduce underdiagnosis bias; using data augmentation based on random windowing or cropping did not reduce the bias.

This is certainly interesting work on an important topic. However, I feel there are major concerns about the validity of the work, which I enumerate below:

This work builds upon and replicates the results presented in reference [1]. Unfortunately, several important factors were not accounted for in the analysis of [1] as pointed out by multiple "matters arising" responses to that paper, (see https://www.nature.com/articles/s41591-022-01847-7, https://www.nature.com/articles/s41591-022-01846-8). The current work also fails to account for multiple confounding effects which raises doubts about the validity of the current work. A key paper that talks about these factors is [a]

[a] Algorithmic encoding of protected characteristics in chest X-ray disease detection models

Glocker, Ben et al.
eBioMedicine, Volume 89, 104467
https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(23)00032-4/fulltext

This reference is missing in the current paper. I strongly suggest the authors refer to and discuss their current paper in the context of [a]. Let me also elaborate further on the potential pitfalls of the current work:
1. the intensity/contrast of a chest x-ray images depends on the disease and the severity of the disease. Thus, to discern the impact of intensity/contrast on race identification, one needs to control for the distribution of diseases across the different races.
2. another confounding factor is the age distribution: younger patients may have less severe presentation of a disease (higher contrast in x-ray images, perhaps) and if there are more younger patients in a certain subgroup during training, the model will be likely to prefer that subgroup with higher contrast images. Without controlling for such confounding effects, it is very hard to interpret the effect of windowing/contrast (or exposure).
3. similar confounding factors may exist for the field of view: younger patients, shorter patients, very sick patients may have larger effective fields of view.

I strongly urge the authors to carefully think about the potential confounding effects and account for them in their analysis -- in its present form I am not convinced about the validity of the presented results.

Response: Thank you for emphasizing the importance of considering confounders when studying AI bias and its potential underlying causes. We certainly agree with this viewpoint and in fact it was a motivation for the current manuscript in studying how “technical” confounders may help explain some of the previously observed results. Thank you also for pointing to reference [a] Glocker et al. We note that this manuscript was published after the current manuscript was submitted, but we certainly agree it’s highly relevant and important work. As such, we have now referenced it (and the “Matters Arising” references above) and included a discussion of our work in the context of this publication (lines 327 to 335 in the Discussion). We have additionally included a new section in the Results titled “Analysis of potential confounding factors” that leverages the test set resampling approach proposed by Glocker et al. to create more balanced test sets according to age, sex, and disease labels across patient race. We additionally include an assessment of this approach when applied to the training set, where we perform both training and test set resampling. As illustrated in new Extended Data Figures 1 and 2, we find similar patterns regarding the effects of the window width, field of view, and view position parameters in the resampling approaches compared to the original results for both CXP and MXR. For instance, in both the original results 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. While there is some quantitative variation, particularly when performing training set resampling, the overall trends are robustly present, suggesting that our observations cannot be solely explained by age, sex, or disease shifts in the training or testing sets alone. For the underdiagnosis bias analysis, we observe a similar pattern to Glocker et al. for the baseline approach, where the overall sensitivity disparities are reduced when performing balanced resampling, as illustrated in new Extended Data Figure 3. Nonetheless, we find that our view-specific threshold approach can reduce these differences even further. Thus, our findings are synergistic with Glocker et al., where we identify underappreciated differences related to image acquisition and processing in addition to the population and prevalence shifts emphasized by Glocker et al.

As now expanded upon in the Discussion (lines 360 to 362), we note that there may be other unmeasured population shifts that could relate to our results or the results of Seyyed-Kalantari et al. (reference [1]) or Gichoya et al. (reference [7]). Our intent in the current manuscript is not to elucidate all of these potential biases, but instead focus on underexplored "technical" parameters, which could potentially be addressed in a demographics-independent way from an AI perspective, where the knowledge of e.g. patient race, age, or sex/gender is not needed during training or model inference. We believe the new confounder analyses enhance the robustness of our results that were initially shown in the original, "AI-ready" MXR and CXP datasets.

We thank the Reviewer for their insightful comments which have helped us improve our manuscript. We hope that our work adds to the growing list of important efforts in better understanding AI behavior and potential sources of bias, as expanded upon in the Discussion.

Reviewers' comments:

Reviewer #2 (Remarks to the Author):

I thank the authors for performing the additional experiments, and adding the extra discussion -- I think it has strengthened the paper considerably, and I feel they have addressed my comments adequately. However, I would like to point out that it may be very difficult to disentangle the various confounders in the setup studied in the paper. For example, the CT acquisition parameters are often set based on the weight and BMI of the patient, and the image quality also varies based on that. Thus, if weight of patients in the dataset is associated with race, the technical parameters (such as contrast, zoom, etc) themselves may be associated with race and it is quite likely that the model can learn and exploit those associations. Also, the paper does not discuss possible underlying reasons for the results they find: preference of the model towards a particular race when factors such as zoom or contrast are changed in particular ways. Another weakness of the paper is this: the per-view thresholds are specific to each dataset and is is likely that the thresholds will not generalize to other centers. Also, the reduction in disparities is relatively modest and it is not clear if the improved fairness in sensitivity comes at the cost of fairness in other important measures such as precision or specificity.

Response to Reviewers

We thank the Reviewers for their consideration of our manuscript. Below, we provide a point-by-point reply to these comments and detail corresponding changes in the manuscript. The original Reviewer comments are in **bold**, followed by our reply.

Reviewer #2:

I thank the authors for performing the additional experiments, and adding the extra discussion -- I think it has strengthened the paper considerably, and I feel they have addressed my comments adequately. However, I would like to point out that it may be very difficult to disentangle the various confounders in the setup studied in the paper. For example, the CT acquisition parameters are often set based on the weight and BMI of the patient, and the image quality also varies based on that. Thus, if weight of patients in the dataset is associated with race, the technical parameters (such as contrast, zoom, etc) themselves may be associated with race and it is quite likely that the model can learn and exploit those associations.

We thank the Reviewer for their emphasis on considering confounders and we are glad to hear that the added experiments and discussion has adequately addressed the previous comments. To summarize this prior confounder analysis, we performed the test set resampling approach proposed by Glocker et al. (2023) across age, sex, and disease prevalence. We additionally extended this analysis to perform a combined training and test set resampling to control for both training-time and testing-time effects. We also tested directly on the DICOM images in the MXR dataset to control for potential effects when generating the AI-ready preprocessed data.

In addition to these experiments, we have now additionally performed the training and testing resampling approach for BMI. While BMI is not available for CXP, we were able to calculate BMI for MXR by querying the MIMIC-IV database, where height and weight is available for 39% of the dataset. Given the prior results and non-uniform availability of BMI, we performed the resampling strategy for BMI separately from the other factors, where we resampled across patient race to achieve approximately equal distributions of the WHO's BMI classifications (underweight, normal, overweight, obese). As illustrated in the updated Extended Data Figures 1, 2, and 3, we again find similar patterns regarding the effects of the window width, field of view, and view position parameters when controlling for BMI. Additionally, while BMI is not available in CXP, we note that the window width and field of view parameters demonstrated similar patterns for Asian and Black patients in this dataset compared to White patients, whereas the CDC[1] reports opposite trends in BMI for these patients in the US, which we also observe in MXR.

Thus, while we have acknowledged and discussed that there may be many factors contributing to our findings (e.g. lines 293-306 and 374-389), the fact that the overall results hold when controlling for many factors - age, disease, DICOM processing, BMI - further emphasizes the robustness of the findings. We also would like to note that the comment mentions CT exams, but we would like to clarify that the current datasets consist of x-rays. Whether CT or x-ray, the

[1] https://stacks.cdc.gov/view/cdc/106273

general goal of adjusting acquisition parameters based on patient-specific factors (such as BMI) is to maintain consistent image quality across patients (e.g., refs. 28, 29). Thus, a priori, image quality (e.g., sufficient image contrast and relative field of view) should not be significantly affected if the acquisition parameters are properly adjusted, which is an important consideration that we emphasize in the Discussion (lines 384-389).

Also, the paper does not discuss possible underlying reasons for the results they find: preference of the model towards a particular race when factors such as zoom or contrast are changed in particular ways.

Thank you for the opportunity to clarify this point. We have generally tried to be conservative in our claims and refrain from speculating, but we have discussed these considerations and potential underlying reasons in several parts of the manuscript. We include several of the relevant excerpts below, with additional new text indicated in blue. Overall, there are two general considerations: why do these factors influence the AI models, and where do these factors arise to begin with. We organize the excerpts below according to these considerations.

Influence of view position parameter on AI models

Lines 149-156: *"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 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)."*

Lines 310-319: *"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 correlations and even 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 (Extended Data Figure 4)."*

Influence of window width and field of view parameters on AI models

Lines 340-347: *"While altering the window width was designed to mimic changes in contrast and exposure[28,29,49,50], it is an imperfect simulation and does not cleanly map to a single physical value…. Nonetheless, the fact that the race prediction model did show differences in predictions over these parameters* [window width and field of view] *does suggest that it may have learned intrinsic patterns in the underlying datasets. Comparisons of the average image per race may offer some intuition on the model's behavior (Extended Data Figure 6)."*

Extended Data Figure 6*: "Average images per patient race were computed based on the training data split using the default preprocessing. An aggregate average image was computed as the average over the per-race average images. Images for both the CXP and MXR datasets are shown. Differences between the per-race average images and the aggregate average image within each dataset are also shown. Red areas indicate regions where the per-race average image has higher pixel intensity than the aggregate average image, whereas blue areas correspond to lower intensity. Overall, differences are quite subtle and could be caused by a number of factors. Yet the differences at the edges of images and in the relative contrast between lung and non-lung regions could potentially relate to the effects observed in the window width and field of view parameter analysis. For instance, the average image for White patients has relatively high contrast between lung and non-lung regions, which is qualitatively similar to the observed effect of an increased average White prediction score when the window width is decreased."*

Discussion of how differences in factors potentially arise to begin with

Lines 374-389: *"While we focused on mitigating 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 overall by patient race[21–23,52]. 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 that contribute to the findings. Nonetheless, as optimal patient positioning and x-ray exposure parameters depend on many patient-specific factors[28–30,34,49], 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[53]."*

Thus, we note that the observed effects regarding view position are correlated with differences in the proportions of views by race in the underlying datasets. Differences in the window width and relative field of view do not map to a discrete value like view position, but we offer intuition on how the observed effects could arise at the image-level, where even the aggregate mean images show correlations with the observed effects. The added text indicated in blue above further clarifies this point. Regarding how the effects arise to begin with, we reference several studies that have identified disparities in imaging utilization and quality in related domains (e.g., refs. 18, 20-23, 53), which would provide precedence for similar effects here. However, we feel that it is unwise to overly speculate on these underlying causes in the current manuscript, where our core focus is studying the effects of technical factors from an AI perspective.

Another weakness of the paper is this: the per-view thresholds are specific to each dataset and is is likely that the thresholds will not generalize to other centers.

The generalization of thresholds, i.e., calibration, is indeed a general challenge in AI, which is not specific to our approach. Whether using a single threshold or view-specific thresholds, a set of patients/images must be chosen to calculate these thresholds. A common method to adapt thresholds to a new clinical site is to first pilot the AI model on a set of data from the site (either retrospectively or prospectively), and then adjust the thresholds as needed. It would be straightforward to use this method with per-view thresholds, as done with other threshold strategies.

Regardless of site-specific adaptation, one could ask whether the per-view thresholds are more or less likely to generalize to a new dataset than the standard single threshold approach. To test this consideration, we have performed a calibration experiment of choosing thresholds based on the CXP validation set and evaluating in the MXR testing set (and vice versa). This analysis results in calibration curves that compare the target percentage of images above a given threshold (based on the original dataset) to the actual percentage of images in the other dataset. These calibration curves are contained in Extended Data Figure 5. Consistent with the general challenge of calibration in AI, the models demonstrate miscalibration when assessed in the unseen dataset. However, we find that the average calibration error is actually slightly lower when using the per-view thresholds compared to the single threshold.

Thus, while calibration is a general AI challenge with resolution that is outside the scope of the current work, we do not find evidence that the per-view threshold approach is any worse than the standard approach, and could in fact potentially help in some scenarios. We hypothesize that such improvements could result from helping to control for differences in the frequencies of views across sites, which would not be controlled for with a single threshold. This hypothesis could be explored in future work. Regardless, the common approach of adapting thresholds to new sites based on a local pilot evaluation could still be used for the per-view method. In addition to adding Extended Data Figure 5, we have now commented on the general challenge of calibration in the Discussion (lines 324-326).

Also, the reduction in disparities is relatively modest and it is not clear if the improved fairness in sensitivity comes at the cost of fairness in other important measures such as precision or specificity.

Thank you for the opportunity to clarify and for emphasizing an important point. We would argue that the observed ~50% reduction in disparities described in lines 205-210 is indeed meaningful, especially given the source of the reduction. Regardless, we also note that we have generally tried to be conservative in our language, such as lines 321-324: "*We note, however, that this strategy* [per-view thresholds] *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 different score distributions and assessing overall performance and fairness*[42,46-48]". Thus, we acknowledge that there is still room for improvement, but we believe that a ~50% reduction in bias through a straightforward, inference-based strategy that mitigates observed differences in view position utilization is quite significant.

We certainly agree, however, that it would be useful to more explicitly show whether the improved fairness in sensitivity comes at the cost of fairness in specificity. We have now included such analysis in the main text, which complements the Extended Data Table 3 that contains both sensitivity and specificity values. We quantify fairness in specificity as the standard deviation in specificity across races, where a higher standard deviation would indicate more variation and decreased fairness. We have now included these specificity-based fairness calculations in lines 212-218 in the main text, which we copy below for convenience:

"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: -0.98, 0.33). 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: -0.99, 0.19)."

Thus, we find that the improved fairness in sensitivity does not come at the cost of fairness in specificity, with the observed variation in specificity actually slightly decreasing as well.

Reviewers' comments:

Reviewer #2 (Remarks to the Author):

I thank the authors for adding additional experiments to strengthen the paper. To summarize, the manuscript first posits that technical factors such as field of view, exposure and view (AP vs lateral, for example) of chest x-rays affect the detection of race in chest x-rays. For example, higher contrast in the scan makes it more likely for the race-detection model to predict "white"; on the other hand, PA views tend to increase prediction scores for black and Asian categories. Next, the authors argue that underdiagnosis bias presented in Seyyed-Kalantari et al. can be mitigated (at least partially) by controlling for these technical factors. The authors study two approaches: 1. data augmentation with contrast (by windowing) and field of view variations, and 2. setting per-view thresholds after training the model. The conclusion is that the first approach of data augmentation did not work well, while the second approach of setting decision thresholds based on view helped reduce the underdiagnosis bias to some extent.

I believe the paper still has several weaknesses:
1. First, the observed effects of the technical factors on race detection are likely reflective of the biases inherent in the training data. Indeed, the authors seem to agree and they state:
"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." (lines 149-156). This is expected behavior of an AI model. Now, the question is: is the underdiagnosis bias reported in Seyyed-Kalantari et al. due (at least partially) to these confounding technical factors? The authors perform an experiment where they perform data augmentation by randomly changing the contrast (as proxy for exposure) and zoom (as proxy for field of view) while training the model. Assuming that windowing and zooming faithfully mimics the effects of exposure and field of view, respectively, if exposure and field of view were responsible for the observed underdiagnosis bias, one would think that the underdiagnosis bias would be reduced by this data augmentation strategy since done correctly, the data augmentation would ensure that the distribution of zoom and contrast were independent of the race. (As a sanity check, the authors should check if this removes the effect of these technical parameters in the race detection AI algorithm). Instead, the underdiagnosis bias persists and even slightly increases, suggesting that the technical factors are likely not responsible for the underdiagnosis bias observed in Seyyed-Kalantari et al. and reproduced here. In my mind, this undercuts the main stated takeaway of the paper: the importance of considering technical factors as they relate to race-based bias in AI models. The authors do not consider the view (AP vs PA) in this experiment -- so we are not sure if view has a role in the underdiagnosis bias -- the authors could have resampled the training data so that all races were equally represented in each view to determine its effect, but the authors do not perform that experiment.

2. I do not understand the motivation of using a race-blind factor to tune thresholds. Why not just use different thresholds for the different races in order to combat underdiagnosis bias? In fact, looking at Extended data table 3, it seems that in most cases, the higher sensitivity for whites comes at the cost of lower specificity (irrespective of the approach), suggesting that the performance in terms of metrics such as AUROC are similar across every race. The simplest approach would be choose the operating point

based on race and mitigate the underdiagnosis bias.

3. Before using the view as the basis of setting the decision thresholds, the authors should check if it indeed has any effect on the underdiagnosis bias (see comment 1). It seems to me that if one stratifies the patients into multiple subgroups and then optimizes the thresholds in each subgroup to minimize race-based bias, one would observe less bias overall, even if the factor on which the subgroups were made was completely unrelated. The fact that using per-view thresholds seems to show less underdiagnosis bias

4. The authors claim that their method reduces bias by 50%. This is technically true, but we should remember that the bias in the baseline was only a few percent in sensitivity (at the cost of lower specificity), and also that most of these disappear when controlled for age, sex, BMI, etc (Extended data table 4). As I mentioned in comment 1, the results suggest that the technical factors have little influence/effect on the underdiagnosis bias (if there is one in the first place.)

Reviewer #3 (Remarks to the Author):

CONTEXT:
I have been brought in as an additional reviewer to replace Reviewer 1, that was no longer able to comment on the paper. Thus, my first set of comments aim to assess whether the previous reviewer's comments have been appropriately taken into account. The authors' "Response to Reviewers" didn't actually include any responses to Reviewer 1, which seems odd -- if these were indeed supposed to be there, I was not able to find them. I also was not able to find the review from Reviewer 2, but I have seen those parts of it that were addressed in the Response to Reviewers.

I will, moreover, add two major concerns regarding the authors' modelling objective and the used datasets.

SUMMARY:
This paper studies the extent to which acquisition choices are the source of racial biases observed in state-of-the-art chest X-ray diagnostic AI. While this is a very important question, I find the authors' analysis too superficial and making conclusions that the analysis does not sufficiently justify. In particular, as partially also pointed out by Reviewer 1, the used datasets are notorious for their hidden biases, and I don't think confounders are taken into account to the degree necessary to support conclusions that could actually affect how X-ray acquisition is done in the clinic.

Thus: While I find the studied problem extremely important, I don't think the concerns of Reviewer 1 are taken sufficiently into account, and I unfortunately have to add some concerns of my own. As a result of these concerns, I don't think the drawn conclusions -- which are what brings the paper to the level of interest of Nature Communications -- are sufficiently supported to be published at present.

*** FOLLOW-UP ON COMMENTS BY THE PREVIOUS REVIEWER ***

The previous reviewer had 4 main concerns regarding the appropriateness of the study. I will list these 4 main concerns of this reviewer along with my assessment of the authors' adaptations to the concerns.

1. The use of JPEG image format in place of the original DICOM, which is particularly important when you want to address image acquisition.
The authors repeated the experiments using DICOM images and saw similar results.

2. The branding of zooming and windowing as preprocessing
I don't see any changes in this regard, but to me this also isn't an important concern.

3. Whether there are hidden biases in the dataset selection, as there is an unexpectedly large proportion of lateral images in ICU.
I don't see any comment on this, and I do think hidden biases are an important concern -- please see my further concerns below.

4. The reviewer recommended including experiments on brain MRI to validate the method in a modality where the entire process is more controlled
The authors have not included such an analysis, and also do not seem to comment on it. To me, if the flaws of the chest X-ray analysis could be brought down, that would make the need for another dataset less prominent. But this is difficult.

*** FURTHER CONCERNS ***

I have two important further concerns regarding the potential hidden biases in the used datasets, as well as with the paper's motivation.

1. Potential hidden biases

Chest X-ray datasets are notorious for their potential built-in hidden biases, which include but are not limited to:
a) Known errors in diagnostic labels, which are inferred using NLP tools [1]. These errors might have a racial bias -- which could happen e.g. if one group has more follow-up scans (associated with higher error) than another group. Such biases could automatically recalibrate the algorithm towards over- or underdiagnosis. Such biases would also make a proper assessment of group-wise performance -- as carried out in this paper -- impossible.
b) Potential group-wise differences in disease prevalence or -severity.
c) Potential group-wise differences in the use of support devices, which are known for their ability to act as shortcuts for algorithms [2].
d) Potential group-wise differences in the effective dataset size -- if there are generally more views included for one group than another, its effective size goes down, which could affect both training and testing.

The authors don't even provide group-wise numbers that allow us as readers to assess whether such hidden biases might be affecting the algorithm, which leaves me concerned. The label errors are particularly problematic -- I don't think this dataset is suitable for doing any analyses that inform actual real-world choices unless the disease labels are revisited and performed manually by a qualified clinician, at least on the test set.

2. The paper's motivation

The paper is motivated by AI algorithms' ability to recognize race from chest X-ray images. While I was, as the rest of the community, surprised to see this, I disagree with the narrative that paints this as a problem that you want to remove. Consider for a second that disease X has different prevalences between different groups. If this is the case, then the diagnosis label itself will be enough to predict race above chance. Which means that an algorithm that is *unable* to predict race, necessarily has to predict equal disease prevalance across races. If the true prevalence is different across races, the algorithm has no choice but to have a racial performance bias. In other words, reducing the algorithm's ability to predict race is not necessarily good for its ability to predict disease with equal performance across race. Please see [3] for further details. In their study, the authors do actually verify that their performance does not go down -- but their discussion does not reflect this potential challenge. If this paper is to be published in Nature Communications, I think it needs to make sure that this motivating factor is not misrepresented -- otherwise, they risk inspiring the development of more biased methods in our community.

3. Details

I think the authors are sometimes interpreting too much from their quantitative results. An example is the discussion of Extended Data Figure 6, where the authors write "For instance, the average image for White patients has relatively high contrast between lung and non-lung regions, which is qualitatively similar to the observed effect of an increased average White prediction score when the window width is decreased." I don't see any racial differences in the contrast of the average images. However, it seems very likely that the differences observed in Extended Data Figure 6 could be caused by Asian patients on average having a lower BMI than Black and White patients. Also, there is no colorbar, which makes it very hard to interpret the scale of the shown differences. I don't myself see any visual difference between the different average images.

References:
[1] https://laurenoakdenrayner.com/2019/02/25/half-a-million-x-rays-first-impressions-of-the-stanford-and-mit-chest-x-ray-datasets/
[2] Oakden-Rayner, Luke, et al. "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. arXiv." (2019).
{3] Petersen, Eike, et al. "Are demographically invariant models and representations in medical imaging

fair?." arXiv preprint arXiv:2305.01397 (2023).

Reviewer #4 (Remarks to the Author):

I am reviewing this paper for this first time and have been asked to consider whether the authors have satisfactorily responded to the concerns of the last review.

I found the paper to be interesting, well-written and scientifically sound. The issues of AI bias and prediction of race from medical images are certainly topical and important, and I believe that the investigation performed in this paper makes a useful contribution to this area.

The previous reviewer's comments focus on the following points:
(1) Can an association between patient weight/BMI and both race and acquisition parameters be having a confounding effect on the results?
(2) Lack of discussion of possible underlying reasons for results found.
(3) Lack of generalization of per-view thresholding approach.
(4) Modest reduction in sensitivity disparity, and how is specificity affected?

In response to (1), the authors have added an extra experiment which controls for BMI and found similar results to their main analysis. This is a satisfactory response in my opinion.

In response to (2), the authors highlighted several parts of the paper in which such discussion was included, and slightly expanded this discussion. I agree with the authors that they have now sufficiently discussed this issue.

In response to (3), the authors pointed out that the calibration of models to other domains is a common issue in AI and not specific to their work. I agree with the authors that this is the case and that it does not significantly impact their findings. Other work has shown that fairness metrics do not always generalize well in the presence of other forms of domain shift and this is an open research question, but beyond the scope of this paper. If they wanted to, the authors could cite https://doi.org/10.48550/arXiv.2202.01034 as an example of such work.

In response to (4), the authors have now reported results which show that disparities in specificity were not significant. I am satisfied with their response to this point.

Overall, I believe that the authors have responded well to all concerns and I would be happy for the paper to be published.

I just found one minor typo in the caption of Figure 2 – "racial identify" should be "racial identity".

Response to Reviewers

We thank the Reviewers for their comments and consideration of our manuscript. Below, we provide a point-by-point reply to these comments. Specifically, as requested by the editors, this response pertains to the most recent comments for each reviewer, which were provided in relation to the initial submission for Reviewer 1, and revision 2 for Reviewers 2, 3, and 4. Reviewer comments are in **bold**, followed by our reply.

Reviewer #1

The authors have attempted to provide an explanation for an ongoing question of significant importance where we don't understand the underlying mechanism. They have explored an area where previous work has not demonstrated any explanation for the mechanism behind why deep learning models can identify demographics and why there is persistent underdiagnosis for minority groups.

However, there are several concerns with their methodology and presentation of results. Firstly, the data set used to evaluate their hypothesis is flawed. The use of MIMIC chest X-ray and CheXpert datasets, which are released in JPEG format rather than the original DICOM format, makes it difficult to assess their results and conclusions. This is because preprocessing has already been performed on these images – and there is known variation of minor differences of even 8 bit versus 16 bit.

Response: We thank the Reviewer for expressing this point. While our primary goal is to better understand and help mitigate bias of standard AI approaches and datasets, which rely on preprocessed images in an "AI-ready" format, we agree that it is useful to understand if this preprocessing itself causes our observed results. Thus, we have also included evaluation on the original DICOM images for the MIMIC dataset. While the JPEG version is typically used in AI work, these DICOM images are publicly available for MIMIC, though not for CheXpert. We have specifically tested our original models, which were trained on the JPEG images, directly on the MIMIC DICOM images to more robustly test the generalization and sensitivity of our results. Following the DICOM Standard (National Electrical Manufacturers Association, VA, USA), we extract and process the MIMIC images directly from the DICOM files using the default parameters contained in the DICOM headers before AI evaluation. This analysis is contained in a section in the manuscript titled "Analysis of potential confounding factors" with accompanying Extended Data Figures 1-3. Despite the models having been trained on images with different preprocessing, we find that the overall patterns regarding the technical parameter analysis and underdiagnosis bias remain when testing on the DICOM images. For the racial identity prediction task, the AI predictions are still influenced by the technical parameters, and on the disease classification task, the baseline model still shows underdiagnosis bias which is still reduced by our view-specific threshold approach. These analyses suggest that these results are not simply caused by the preprocessing used to create the AI-ready datasets. Nonetheless, we agree that this preprocessing is an important consideration and we have expanded on this consideration in the Discussion (lines 367-375).

Moreover, the processing approaches suggested by the authors include zooming and windowing. These would not typically be considered image preprocessing but rather

factors that are usually changed on the view/display by the end user – usually a radiologist . Therefore, their selection of preprocessing tasks is misleading.

Response: Thank you for raising this important point. We agree that it is challenging to define a precise delineation of what is considered image "preprocessing". Related to the previous point, there are a number of processing stages that take place from initial x-ray exposure through to when the image is actually viewed by the end user (i.e., radiologist) or even AI model. For instance, most DICOM viewers implement image processing steps according to the DICOM Standard by default, including windowing according to the default values in the DICOM header before display to the end user, which can then be adjusted as desired as rightfully stated in the Reviewer's comment above. Similarly, preprocessing stages such as windowing, normalization, resizing, and even bit depth conversion are commonly used in AI approaches. Ultimately, our goal was to simulate technical variations that are important in chest x-ray image acquisition and processing, including overall contrast/exposure and the relative size of the field of view, which can be changed through collimation. To study the effect of contrast, we perform windowing with different window widths to produce different levels of overall contrast. To study variations in collimation, we effectively perform 'electronic collimation' (Tsalafoutas et al., reference 37; Bomer et al., reference 38) to modify the relative field of view for the image. We had used the term "zoom" as an intuitive, non-technical way to explain this phenomenon in the initial submission, but we recognize that this term can have different meanings in different contexts and can cause confusion. Thus, we now refer to this aspect as the "field of view" parameter instead of the "zoom" parameter and have updated its use throughout the manuscript. Altogether, we believe that the window width and field of view parameters represent highly relevant transformations for chest x-ray acquisition and preprocessing, especially in the context of AI development. In addition to updating the field of view terminology since the initial submission, we have expanded upon these points in lines 70-71, 87-88, 367-375.

Secondly, there is concern for sensitivity to the labels that were provided in their data. For example, if you look at extended data table two in MIMIC chest X-ray dataset, most images are acquired using portable method as expected for ICU images. However, when you look at distribution between AP and lateral views quite a large number of lateral views in ICU which would not make sense. This may indicate a need for further cleaning of datasets by authors.

Response: Thank you for the opportunity to clarify. According to the initial paper describing the MIMIC dataset (Johnson et al., 2019), the authors state the following: "*We queried the BIDMC EHR for chest radiograph studies acquired in the emergency department between 2011–2016, and extracted only the set of patient identifiers associated with these studies. A collection of images associated with a single report is referred to as a study, identified by a unique identifier, the study ID. We then extracted all chest radiographs and radiology reports available in the RIS for this set of patients between 2011–2016*."

Thus, while the set of patients was identified based on studies acquired in the emergency department, it is our interpretation that all studies for these patients were then extracted regardless of whether they originated from the emergency department or not. We have since expanded on the description of the MIMIC dataset in the Methods section (lines 556-560) to more fully reflect the original description of the dataset. Additionally, we realize that our original presentation of Extended Data Table 2 may have been confusing in that we displayed the breakdown of Standard vs Portable views for the AP view position, since this position is used for the vast majority of Portable views (as stated in the comment above). For improved clarity, we have added an additional row in this table since the initial submission showing the total proportion of Portable views amongst all images (33.8%) and updated the table legend with further description.

We have additionally pursued further data cleaning using three strategies to ensure that the AP and Lateral views are properly labeled as the Reviewer suggests. First, we plotted 100 random AP views and 100 random Lateral views according to the MIMIC metadata. We manually reviewed these 200 images and determined that they are all correctly labeled (e.g., there were no Lateral views that were labeled as AP and vice versa). Second, we extracted the View Position directly from the DICOM files that were obtained for the prior comment. We compared these extracted views to the metadata and there was a 100% correspondence. Third, we reviewed the code used by the dataset creators to create this dataset, which is publicly available on Github. We did not find any apparent issues through this review. While some amount of noise is likely to be expected in any clinical dataset, we believe that these efforts support the validity and proper curation of the MIMIC dataset, which is further supported by its popularity and the initial publication describing its curation.

I recommend that the authors try a different modality such as brain MRI to demonstrate this process of preprocessing and underdiagnosis. The reason for this recommendation is because there is more harmonization and standardization of brain imaging. If there is a desire to work on chest X-rays, there is an opportunity to access and curate a dataset to make sure that it's applicable for this task.

Response: The goal for the current work is to better understand and address biases identified in recent high-profile work using highly-popular chest x-ray datasets, rather than identifying new biases in other domains altogether. In particular, we are unaware of AI results showing that race can be predicted from brain MRI, and we believe that training models to do so in this work would be counterproductive and may be deemed especially controversial given the modality (i.e., discriminating between brain images of different races). With our focus thus on popular chest x-ray datasets, we have aimed to ensure sufficient handling of potential confounders as suggested by the Reviewer. These important comments raised pertain to the use of DICOM images and sufficient curation of view information. Beyond these factors, we have additionally now controlled for potential confounders of age, disease prevalence, sex, and BMI, all of which resulted in similar core findings. We also note that studying preprocessing in a highly popular yet less standardized modality than brain MRI may actually be more beneficial in that it points to the potential of improved harmonization and standardization that could potentially reduce these issues. Thus, while we certainly agree that identifying potential AI bias and sources thereof is important in all medical imaging domains, it is essential to study existing findings, especially those involving high-profile work and prominent AI datasets.

There are also other minor comments including some papers referenced in their work have been published but an archive link is provided instead (e.g., reference 27). I recommend that the authors go through their references and make sure they're referencing the most recent publication.

Response: Thank you for pointing this out. We have reviewed all references and updated to the most recent publications we could find, including removing arXiv links when appropriate.

In terms of information presentation, I thought figures two and three were very dense and difficult to follow. I recommend better visualization of results.

Response: Thank you for this feedback. We have updated Figures 2 and 3 since the initial submission to help improve interpretability, including increasing spacing and font sizes and better harmonizing the presentation of results across CXP and MXR.

While I do want to commend the authors for mentioning that training-based data augmentation did not reduce underdiagnosis (and increased it), presenting an area that can be further researched if we can understand the robustness of underlying methods proposed by authors; overall my recommendation is that the paper is flawed in its methodology

Response: Thank you for your consideration of our manuscript and for mentioning the significance of the underlying problem and potential implications of our results. We believe that the additional experiments and discussion since the original submission confirm the validity of our methods and support the robustness of our findings.

Reviewer #2

I thank the authors for adding additional experiments to strengthen the paper. To summarize, the manuscript first posits that technical factors such as field of view, exposure and view (AP vs lateral, for example) of chest x-rays affect the detection of race in chest x-rays. For example, higher contrast in the scan makes it more likely for the race-detection model to predict "white"; on the other hand, PA views tend to increase prediction scores for black and Asian categories. Next, the authors argue that underdiagnosis bias presented in Seyyed-Kalantari et al. can be mitigated (at least partially) by controlling for these technical factors. The authors study two approaches: 1. data augmentation with contrast (by windowing) and field of view variations, and 2. setting per-view thresholds after training the model. The conclusion is that the first approach of data augmentation did not work well, while the second approach of setting decision thresholds based on view helped reduce the underdiagnosis bias to some extent.

I believe the paper still has several weaknesses:

1. First, the observed effects of the technical factors on race detection are likely reflective of the biases inherent in the training data. Indeed, the authors seem to agree and they state: "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." (lines 149-156). This is expected behavior of an AI model. Now, the question is: is the underdiagnosis bias reported in Seyyed-Kalantari et al. due (at least partially) to these confounding technical factors? The authors perform an experiment where they perform data augmentation by randomly changing the contrast (as proxy for exposure) and zoom (as proxy for field of view) while training the model. Assuming that windowing and zooming faithfully mimics the effects of exposure and field of view, respectively, if exposure and field of view were responsible for the observed underdiagnosis bias, one would think that the underdiagnosis bias would be reduced by this data augmentation strategy since done correctly, the data augmentation would ensure that the distribution of zoom and contrast were independent of the race. (As a sanity check, the authors should check if this removes the effect of these technical parameters in the race detection AI algorithm). Instead, the underdiagnosis bias persists and even slightly increases, suggesting that the technical factors are likely not responsible for the underdiagnosis bias observed in Seyyed-Kalantari et al. and reproduced here. In my mind, this undercuts the main stated takeaway of the paper: the importance of considering technical factors as they relate to race-based bias in AI models. The authors do not consider the view (AP vs PA) in this experiment -- so we are not sure if view has a role in the underdiagnosis bias -- the authors could have resampled the training data so that all races were equally represented in each view to determine its effect, but the authors do not perform that experiment.

Response: We are encouraged that the Reviewer seems to agree that biases exist regarding the technical parameters and that this could contribute to the effects observed on race prediction, where previously it was suggested that other confounders such as age, sex, disease labels, and BMI might instead explain the observed effects. It is certainly not obvious that this would be the case, and we believe it supports a core message of our work on the underappreciation of technical preprocessing and acquisition parameters. In terms of the potential contribution of these factors to the underdiagnosis bias observed by Seyyed-Kalantari et al., the Reviewer rightfully suggests that there could be many possible approaches to try to mitigate these biases. We certainly do not claim that our methods are the only ones, but we were motivated to choose practical approaches that encompass both training and inference-time strategies. Inference-time strategies such as our threshold approach are particularly attractive compared to e.g. training data resampling given that they don't require training a new model and can more easily be adapted to new sites. As the Reviewer mentions in a later comment, we do indeed show that this method reduces underdiagnosis bias reported in Seyyed-Kalantari et al.. Regarding the data augmentation strategy, we commented on possible explanations in the Discussion (lines 337-340) and note that even if this strategy did not reduce the underdiagnosis bias, the reduction in bias through our per-view threshold strategy still supports the importance of considering technical factors as they relate to race-based bias in AI models.

2. I do not understand the motivation of using a race-blind factor to tune thresholds. Why not just use different thresholds for the different races in order to combat underdiagnosis bias? In fact, looking at Extended data table 3, it seems that in most

cases, the higher sensitivity for whites comes at the cost of lower specificity (irrespective of the approach), suggesting that the performance in terms of metrics such as AUROC are similar across every race. The simplest approach would be choose the operating point based on race and mitigate the underdiagnosis bias.

Response: Race-based medicine has a very controversial history with many recent efforts pointing to its flaws. Beyond the ethical concerns, such an approach would not be practical or likely effective. Indeed, we've shown for instance that there are biases in views by race in the studied clinical datasets, making this a confounder that could lead to changes in performance if thresholds were simply chosen based on race and these view distributions changed over time or across clinical sites. Race, as opposed to view, is also not readily available in the DICOM images which would be used for inference in clinical practice; furthermore, regulatory agencies would likely express strong concern over approving products with race-based thresholds.

3. Before using the view as the basis of setting the decision thresholds, the authors should check if it indeed has any effect on the underdiagnosis bias (see comment 1). It seems to me that if one stratifies the patients into multiple subgroups and then optimizes the thresholds in each subgroup to minimize race-based bias, one would observe less bias overall, even if the factor on which the subgroups were made was completely unrelated. The fact that using per-view thresholds seems to show less underdiagnosis bias

Response: It appears that this comment may have been cut off or incomplete, but the start of the last sentence seems to concur that the per-view thresholds show less underdiagnosis bias. For the suggestion of optimizing thresholds for each patient subgroup, we assume this is referring to the previous suggestion of using different thresholds for different race subgroups, which has a number of drawbacks as detailed above.

4. The authors claim that their method reduces bias by 50%. This is technically true, but we should remember that the bias in the baseline was only a few percent in sensitivity (at the cost of lower specificity), and also that most of these disappear when controlled for age, sex, BMI, etc (Extended data table 4). As I mentioned in comment 1, the results suggest that the technical factors have little influence/effect on the underdiagnosis bias (if there is one in the first place.)

Response: We apologize if it was unclear, but the values in Extended Data Table 4 reflect the distribution of the original and resampled test sets and not the AI model's predictions. When controlling for age, sex, and BMI (such as in the resampled test sets), we do indeed still observe sensitivity disparities that are reduced by the per-view threshold approach (e.g., Extended Data Figure 3).

Reviewer #3
CONTEXT:
I have been brought in as an additional reviewer to replace Reviewer 1, that was no longer able to comment on the paper. Thus, my first set of comments aim to assess

whether the previous reviewer's comments have been appropriately taken into account. The authors' "Response to Reviewers" didn't actually include any responses to Reviewer 1, which seems odd -- if these were indeed supposed to be there, I was not able to find them. I also was not able to find the review from Reviewer 2, but I have seen those parts of it that were addressed in the Response to Reviewers.

Response: Thank you for reviewing our manuscript. It is our understanding that a system glitch caused our prior response to Reviewer 1 to not be available to you. Please find this response above. We welcome your comments to this response.

I will, moreover, add two major concerns regarding the authors' modelling objective and the used datasets.

SUMMARY:
This paper studies the extent to which acquisition choices are the source of racial biases observed in state-of-the-art chest X-ray diagnostic AI. While this is a very important question, I find the authors' analysis too superficial and making conclusions that the analysis does not sufficiently justify. In particular, as partially also pointed out by Reviewer 1, the used datasets are notorious for their hidden biases, and I don't think confounders are taken into account to the degree necessary to support conclusions that could actually affect how X-ray acquisition is done in the clinic.

Response: Thank you for the opportunity to clarify the goals and specific conclusions of our work. We agree that potential hidden biases in highly-popular AI datasets is of critical importance. In fact, this consideration was a core motivation for our work in studying if underappreciated "technical" biases exist in these popular datasets and if these factors may help partially explain previous high-profile AI work using these datasets, as we describe in lines 294-307). Thus, our focus is driven from an AI perspective and we further comment on hidden biases below. Separately, we agree that there are important considerations for how x-ray acquisition is done in the clinic as the Reviewer specifically mentions, but we do not intend for our conclusions to directly inform clinical x-ray acquisition and have not made such claims. We have added the following text in lines 389-393 to make this more explicit (updates highlighted in blue):

"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."

Instead our conclusions are summarized in the first paragraph of the Discussion as follows, with updates highlighted in blue which were made to further emphasize the focus on AI and popular datasets:

"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."

Thus, explicitly we conclude from our results that a) "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" and b) "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". We believe that these conclusions are directly supported by our analysis showing that the race prediction models show significant changes in predictions when varying the studied technical factors (e.g., Figures 2 and 3) and that our per-view threshold strategy reduces the underdiagnosis bias (e.g., Figure 4) reported by Seyyed-Kalantari et al. (Nature Medicine, 2021). Importantly, as detailed in the "Analysis of potential confounder factors" results section, these core findings persist even when controlling for numerous potential confounders/hidden biases of age, sex, disease prevalence, BMI, and DICOM-based processing using multiple strategies.

Thus: While I find the studied problem extremely important, I don't think the concerns of Reviewer 1 are taken sufficiently into account, and I unfortunately have to add some concerns of my own. As a result of these concerns, I don't think the drawn conclusions -- which are what brings the paper to the level of interest of Nature Communications -- are sufficiently supported to be published at present.

***** FOLLOW-UP ON COMMENTS BY THE PREVIOUS REVIEWER *****

The previous reviewer had 4 main concerns regarding the appropriateness of the study. I will list these 4 main concerns of this reviewer along with my assessment of the authors' adaptations to the concerns.

Response: Please find our complete response to Reviewer 1 above. We additionally respond to the specific comments regarding datasets in the section below. We also note that Reviewer 1's comments were made prior to much of the existing confounder analysis, as detailed in our response to their comments above.

1. The use of JPEG image format in place of the original DICOM, which is particularly important when you want to address image acquisition.
The authors repeated the experiments using DICOM images and saw similar results.

2. The branding of zooming and windowing as preprocessing

I don't see any changes in this regard, but to me this also isn't an important concern.

3. Whether there are hidden biases in the dataset selection, as there is an unexpectedly large proportion of lateral images in ICU.
I don't see any comment on this, and I do think hidden biases are an important concern -- please see my further concerns below.

4. The reviewer recommended including experiments on brain MRI to validate the method in a modality where the entire process is more controlled
The authors have not included such an analysis, and also do not seem to comment on it. To me, if the flaws of the chest X-ray analysis could be brought down, that would make the need for another dataset less prominent. But this is difficult.

***** FURTHER CONCERNS *****

I have two important further concerns regarding the potential hidden biases in the used datasets, as well as with the paper's motivation.

1. Potential hidden biases

Chest X-ray datasets are notorious for their potential built-in hidden biases, which include but are not limited to:
a) Known errors in diagnostic labels, which are inferred using NLP tools [1]. These errors might have a racial bias -- which could happen e.g. if one group has more follow-up scans (associated with higher error) than another group. Such biases could automatically recalibrate the algorithm towards over- or underdiagnosis. Such biases would also make a proper assessment of group-wise performance -- as carried out in this paper -- impossible.

Response: We agree that there may be several sources of bias contributing to the group-wise performance differences observed by Seyyed-Kalantari et al., who reported an underdiagnosis bias when using standard AI approaches and standard AI datasets. To reiterate, our goal was to assess if potential hidden biases related to technical parameters, which are underappreciated by the AI community and offer a means for reducing AI diagnostic performance bias, partially contribute to these results. We indeed find that our strategy to address a previously-unreported technical bias significantly reduces the underdiagnosis bias reported by Seyyed-Kalantari et al., even when controlling for a number of possible confounders. However, as we acknowledge in lines 331-332, this strategy did not completely eliminate the performance bias, leaving room for improvement for other strategies and for identifying other sources of bias that may contribute, including potentially label noise/bias. We have now expanded upon this consideration regarding NLP-based labeling by including the following sentence in line 375: "*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.*" We note also that the diagnostic labels are not used for the racial identity prediction analysis (e.g., Figures 2 and 3) and thus only apply to the diagnostic performance analysis (e.g., Figure 4).

We note also that the original CheXpert paper (Irvin et al., 2019) included a validation of their NLP tool compared to radiologists, which demonstrated high performance (average F1 scores > 0.90) that exceeded prior NLP tools, including that of the NIH dataset mentioned in the referenced blog post [1] by Dr. Oakden-Rayner. In this blog post, Dr. Oakden-Rayner states "*Unlike the earlier NIH paper, where they did not initially test the performance of their labeller against human labelled cases in their dataset, the Stanford team produced a dataset of 1000 hand labelled reports, and the MIT team produced a dataset of 687 hand labelled reports. These results are believable, and show very good performance across the board*." While they certainly acknowledge limitations, they also state "*This is a first impressions post, so I don't want to make a strong judgement too early, but I think it is pretty safe to say that this dataset is the best quality chest x-ray dataset we currently have*." Thus, while no AI dataset is perfect, this dataset is highly used by the AI community including in the referenced high-profile work, supporting the importance of analysis on these datasets.

b) Potential group-wise differences in disease prevalence or -severity.

Response: Potential group-wise differences in disease prevalence is certainly an important consideration. We have accounted for this in our confounder analyses, where we find that the observed effects remain even when performing training and test set resampling to control for group-wise differences in this factor (please refer to lines 221-266 and Extended Figures 1 and 2).

c) Potential group-wise differences in the use of support devices, which are known for their ability to act as shortcuts for algorithms [2].

Response: Thank you for this suggestion. We have now examined the frequency of support devices per subgroup and observe similar frequencies across subgroups for both datasets. For CXP, the frequencies are 94.0%, 94.3%, and 93.9% per image for White, Black, and Asian patients respectively. For MXR, the frequencies are 94.2%, 93.3%, and 91.9% for White, Black, and Asian patients respectively. We have now added these numbers in lines 592-595. While the support devices frequencies are similar across subgroups, the notion of shortcut connections is indeed an important consideration and was a motivating factor for the per-view threshold approach, as we describe in lines 314-325:

"*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 correlations and even shortcut connections between the view type and the presence of pathological findings. Indeed, we do find that AI models trained to predict pathological findings exhibit different score distributions for* *different views (Extended Data Figure 4). This observation can help explain why choosing score thresholds per view can help mitigate the underdiagnosis bias.*"

d) Potential group-wise differences in the effective dataset size -- if there are generally more views included for one group than another, its effective size goes down, which could affect both training and testing.

Response: The relative amount of data available for each subgroup is certainly an important question, where public AI datasets, including those studied, are notoriously skewed towards White patients. Regarding effective dataset size, we observe similar subgroup proportions whether calculating by patient or view for both datasets. For MXR, the percentages are 67.2% White, 17.4% Black, 3.8% Asian by patient and 68.4% White, 17.4% Black, 3.5% Asian by view. For the CXP dataset, the subgroup percentages are 63.6% White, 12.2% Asian, and 5.4% Black by patient and 63.8% White, 11.8% Asian, 6.1% Black by view. Thus, the "effective" and "absolute" dataset sizes are similar. We have now added these numbers to lines 573-577 in the Methods to complement the per view numbers reported in Extended Data Table 2. Regarding absolute dataset sizes, we note that the observed effects of the technical parameters remained when performing resampling to equalize subgroup proportions, as described in the "Analysis of potential confounding factors" section.

The authors don't even provide group-wise numbers that allow us as readers to assess whether such hidden biases might be affecting the algorithm, which leaves me concerned. The label errors are particularly problematic -- I don't think this dataset is suitable for doing any analyses that inform actual real-world choices unless the disease labels are revisited and performed manually by a qualified clinician, at least on the test set.

Response: In terms of group-wise numbers, we note that such values are reported across views in Extended Data Table 2 and by age, sex, BMI, and disease prevalence in Extended Data Table 4. Based on the prior comment, we have now also added numbers for supported devices and relative data sizes which show similar patterns across subgroups. This adds to our prior analysis controlling for potential hidden biases in age, sex, BMI, disease prevalence, and DICOM processing using multiple approaches, where this level of confounder/hidden bias analysis exceeds that of the referenced related work in the field.

We additionally highlight that the studied datasets are actively being used to develop AI to inform real-world choices and are supported by thousands of citations. Thus, the concern of other potential hidden biases and diagnostic label errors is not specific to our work, but any work using these and related datasets which have become benchmarks in the field. To this end, our efforts align with this core message of carefully considering the construction and subsequent use of AI datasets, where we specifically focus on technical factors for the reasons described above. We certainly agree that manual labeling of these datasets would be useful for the entire AI community, but this would require clinician review of tens of thousands of images (even for the test sets) which we believe is not within the scope of the current work. Given these considerations, in addition to the changes previously described, we have made text changes in lines 48, 163, 286, 291, and 401 to reiterate the focus and importance of studying popular AI datasets and their potential hidden biases.

2. The paper's motivation

The paper is motivated by AI algorithms' ability to recognize race from chest X-ray images. While I was, as the rest of the community, surprised to see this, I disagree with the narrative that paints this as a problem that you want to remove. Consider for a second that disease X has different prevalences between different groups. If this is the case, then the diagnosis label itself will be enough to predict race above chance. Which means that an algorithm that is *unable* to predict race, necessarily has to predict equal disease prevalance across races. If the true prevalence is different across races, the algorithm has no choice but to have a racial performance bias. In other words, reducing the algorithm's ability to predict race is not necessarily good for its ability to predict disease with equal performance across race. Please see [3] for further details. In their study, the authors do actually verify that their performance does not go down -- but their discussion does not reflect this potential challenge. If this paper is to be published in Nature Communications, I think it needs to make sure that this motivating factor is not misrepresented -- otherwise, they risk inspiring the development of more biased methods in our community.

Response: We agree that removing the ability to recognize race does not ensure fairness. We certainly would like to be clear that this was not the goal of our work, and we were careful not to make this claim at any point throughout the manuscript. To avoid any potential misinterpretation, we have made this more explicit in lines 307-310 in the Discussion, which we include below (updates are in blue). Thank you for the opportunity to clarify.

"As such, our goal was not to elucidate all of the features enabling AI-based race prediction, but instead focusing 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."

3. Details

I think the authors are sometimes interpreting too much from their quantitative results. An example is the discussion of Extended Data Figure 6, where the authors write "For instance, the average image for White patients has relatively high contrast between lung and non-lung regions, which is qualitatively similar to the observed effect of an increased average White prediction score when the window width is decreased." I don't see any racial differences in the contrast of the average images. However, it seems very likely that the differences observed in Extended Data Figure 6 could be caused by Asian patients on average having a lower BMI than Black and White patients. Also, there is no colorbar, which makes it very hard to interpret the scale of the shown differences. I don't myself see any visual difference between the different average images.

Response: Thank you for this input. The referenced sentence was added after the last round of review based on a comment from another reviewer, but we agree that it is not necessary. We have now removed this line. We have also added a colorbar and a version of the plot based on the resampled MXR test set that controls for BMI shifts. The differences across the mean images are indeed subtle but reach relative magnitudes of ~10-20%, and similar overall patterns are observed when controlling for BMI.

References:
[1] https://laurenoakdenrayner.com/2019/02/25/half-a-million-x-rays-first-impressions-of-the-stanford-and-mit-chest-x-ray-datasets/
[2] Oakden-Rayner, Luke, et al. "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. arXiv." (2019).
{3] Petersen, Eike, et al. "Are demographically invariant models and representations in medical imaging fair?." arXiv preprint arXiv:2305.01397 (2023).

Reviewer #4
I am reviewing this paper for this first time and have been asked to consider whether the authors have satisfactorily responded to the concerns of the last review.

Response: Thank you for reviewing our manuscript. It is our understanding that a system glitch may have caused only our response to Reviewer 2 to be available to you and not our response to Reviewer 1. Please also find this response above. We welcome your comments to this response.

I found the paper to be interesting, well-written and scientifically sound. The issues of AI bias and prediction of race from medical images are certainly topical and important, and I believe that the investigation performed in this paper makes a useful contribution to this area.

The previous reviewer's comments focus on the following points:
(1) Can an association between patient weight/BMI and both race and acquisition parameters be having a confounding effect on the results?
(2) Lack of discussion of possible underlying reasons for results found.
(3) Lack of generalization of per-view thresholding approach.
(4) Modest reduction in sensitivity disparity, and how is specificity affected?

In response to (1), the authors have added an extra experiment which controls for BMI and found similar results to their main analysis. This is a satisfactory response in my opinion.

In response to (2), the authors highlighted several parts of the paper in which such discussion was included, and slightly expanded this discussion. I agree with the authors that they have now sufficiently discussed this issue.

In response to (3), the authors pointed out that the calibration of models to other domains is a common issue in AI and not specific to their work. I agree with the authors that this is the case and that it does not significantly impact their findings. Other work has shown that fairness metrics do not always generalize well in the presence of other forms of domain shift and this is an open research question, but beyond the scope of this paper. If they wanted to, the authors could cite https://doi.org/10.48550/arXiv.2202.01034 as an example of such work.

Response: Thank you for pointing us to this interesting reference, which we agree is useful to include and we have now done so (it appears as reference 49).

In response to (4), the authors have now reported results which show that disparities in specificity were not significant. I am satisfied with their response to this point.

Overall, I believe that the authors have responded well to all concerns and I would be happy for the paper to be published.

I just found one minor typo in the caption of Figure 2 – “racial identify” should be “racial identity”.

Response: Thank you very much for your positive feedback of the manuscript and our response to Reviewer 2. Thank you also for catching the typo in Figure 2, which we have now fixed.

REVIEWER COMMENTS

Reviewer #2 (Remarks to the Author):

I thank the authors for their responses to my earlier comments. Unfortunately, I still think the paper's claims are not adequately justified by the experiments.

1. It is not clear to me that zoom and contrast are the main factors driving the observed biases -- if they were, the data augmentation strategy (which should, in theory, make the zoom and contrast independent of race) should have reduced bias -- but it did not. This leads me to believe that other confounding factors are likely responsible.
2. It is not clear how well the cropping and windowing simulate the actual field of view and exposure -- for example, higher radiation dose results in higher SNR, which is not possible to get with windowing
3. Some of the figures are really not convincing -- for example, Extended Data Figure 4 -- on the CXP dataset, the score distribution is heavily tilted towards the "findings present" class. This likely suggests the scores are heavily miscalibrated (even compared to MXR). As such I am not sure how to interpret the claims about the scores increasing or decreasing when a particular parameter such as zoom or window is changed.

Reviewer #4 (Remarks to the Author):

I have read the authors' response carefully as well as the other reviewers' comments. My previous opinion was that the paper was scientifically sound and a useful contribution to the literature in what I believe to be an important area. Based on what I have read in this resubmission I see no reason to change this opinion. I recommend acceptance of the paper without further revision.

I have included below the full response letter of the authors, which also includes the comments of the reviewers. I have added my opinions on the points made and the authors' responses into the below text, preceded by "*****".

Reviewer #1

The authors have attempted to provide an explanation for an ongoing question of significant importance where we don't understand the underlying mechanism. They have explored an area where previous work has not demonstrated any explanation for the mechanism behind why deep learning models can identify demographics and why there is persistent underdiagnosis for minority groups.

However, there are several concerns with their methodology and presentation of results. Firstly, the data set used to evaluate their hypothesis is flawed. The use of MIMIC chest X-ray and CheXpert datasets, which are released in JPEG format rather than the original DICOM format, makes it difficult to assess their results and conclusions. This is because preprocessing has already been performed on these images – and there is known variation of minor differences of even 8 bit versus 16 bit.

Response: We thank the Reviewer for expressing this point. While our primary goal is to better understand and help mitigate bias of standard AI approaches and datasets, which rely on preprocessed images in an "AI-ready" format, we agree that it is useful to understand if this preprocessing itself causes

our observed results. Thus, we have also included evaluation on the original DICOM images for the MIMIC dataset. While the JPEG version is typically used in AI work, these DICOM images are publicly available for MIMIC, though not for CheXpert. We have specifically tested our original models, which were trained on the JPEG images, directly on the MIMIC DICOM images to more robustly test the generalization and sensitivity of our results. Following the DICOM Standard (National Electrical Manufacturers Association, VA, USA), we extract and process the MIMIC images directly from the DICOM files using the default parameters contained in the DICOM headers before AI evaluation. This analysis is contained in a section in the manuscript titled "Analysis of potential confounding factors" with accompanying Extended Data Figures 1-3. Despite the models having been trained on images with different preprocessing, we find that the overall patterns regarding the technical parameter analysis and underdiagnosis bias remain when testing on the DICOM images. For the racial identity prediction task, the AI predictions are still influenced by the technical parameters, and on the disease classification task, the baseline model still shows underdiagnosis bias which is still reduced by our view-specific threshold approach. These analyses suggest that these results are not simply caused by the preprocessing used to create the AI-ready datasets. Nonetheless, we agree that this preprocessing is an important consideration and we have expanded on this consideration in the Discussion (lines 367-375).
***** Reviewer #1 expressed concern about the impact of the data format of the images. The authors have satisfactorily addressed this point by adding an extra experiment on the original DICOM images of the MIMIC dataset, showing results consistent with their original findings.

Moreover, the processing approaches suggested by the authors include zooming and windowing. These would not typically be considered image preprocessing but rather factors that are usually changed on the view/display by the end user – usually a radiologist . Therefore, their selection of preprocessing tasks is misleading.

Response: Thank you for raising this important point. We agree that it is challenging to define a precise delineation of what is considered image "preprocessing". Related to the previous point, there are a number of processing stages that take place from initial x-ray exposure through to when the image is actually viewed by the end user (i.e., radiologist) or even AI model. For instance, most DICOM viewers implement image processing steps according to the DICOM Standard by default, including windowing according to the default values in the DICOM header before display to the end user, which can then be adjusted as desired as rightfully stated in the Reviewer's comment above. Similarly, preprocessing stages such as windowing, normalization, resizing, and even bit depth conversion are commonly used in AI approaches. Ultimately, our goal was to simulate technical variations that are important in chest x-ray image acquisition and processing, including overall contrast/exposure and the relative size of the field of view, which can be changed through collimation. To study the effect of contrast, we perform windowing with different window widths to produce different levels of overall contrast. To study variations in collimation, we effectively perform 'electronic collimation' (Tsalafoutas et al., reference 37; Bomer et al., reference 38) to modify the relative field of view for the image. We had used the term "zoom" as an intuitive, non-technical way to explain this phenomenon in the initial submission, but we recognize that this term can have different meanings in different contexts and can cause confusion. Thus, we now refer to this aspect as the "field of view" parameter instead of the "zoom" parameter and have updated its use throughout the manuscript. Altogether, we believe that the window width and field of view parameters represent highly relevant transformations for chest x-ray acquisition and preprocessing, especially in the context of AI development. In addition to updating the field of view terminology since

the initial submission, we have expanded upon these points in lines 70-71, 87-88, 367-375.
***** This point relates to the description of the zooming and windowing operations as “preprocessing”. The authors have adjusted their terminology so I think this point has been satisfactorily addressed.

Secondly, there is concern for sensitivity to the labels that were provided in their data. For example, if you look at extended data table two in MIMIC chest X-ray dataset, most images are acquired using portable method as expected for ICU images. However, when you look at distribution between AP and lateral views quite a large number of lateral views in ICU which would not make sense. This may indicate a need for further cleaning of datasets by authors.
Response: Thank you for the opportunity to clarify. According to the initial paper describing the MIMIC dataset (Johnson et al., 2019), the authors state the following: “We queried the BIDMC EHR for chest radiograph studies acquired in the emergency department between 2011–2016, and extracted only the set of patient identifiers associated with these studies. A collection of images associated with a single report is referred to as a study, identified by a unique identifier, the study ID. We then extracted all chest radiographs and radiology reports available in the RIS for this set of patients between 2011–2016.”
Thus, while the set of patients was identified based on studies acquired in the emergency department, it is our interpretation that all studies for these patients were then extracted regardless of whether they originated from the emergency department or not. We have since expanded on the description of the MIMIC dataset in the Methods section (lines 556-560) to more fully reflect the original description of the dataset. Additionally, we realize that our original presentation of Extended Data Table 2 may have been confusing in that we displayed the breakdown of Standard vs Portable views for the AP view position, since this position is used for the vast majority of Portable views (as stated in the comment above). For improved clarity, we have added an additional row in this table since the initial submission showing the total proportion of Portable views amongst all images (33.8%) and updated the table legend with further description.
We have additionally pursued further data cleaning using three strategies to ensure that the AP and Lateral views are properly labeled as the Reviewer suggests. First, we plotted 100 random AP views and 100 random Lateral views according to the MIMIC metadata. We manually reviewed these 200 images and determined that they are all correctly labeled (e.g., there were no Lateral views that were labeled as AP and vice versa). Second, we extracted the View Position directly from the DICOM files that were obtained for the prior comment. We compared these extracted views to the metadata and there was a 100% correspondence. Third, we reviewed the code used by the dataset creators to create this dataset, which is publicly available on Github. We did not find any apparent issues through this review. While some amount of noise is likely to be expected in any clinical dataset, we believe that these efforts support the validity and proper curation of the MIMIC dataset, which is further supported by its popularity and the initial publication describing its curation.
***** This point relates to the concerns about the data labelling. The authors have made changes to the descriptions of the data and performed some further verification of label fidelity. In my opinion, this is a satisfactory response.

I recommend that the authors try a different modality such as brain MRI to demonstrate this process of preprocessing and underdiagnosis. The reason for this recommendation is because there is more harmonization and standardization of brain imaging. If there is a desire to work on chest X-rays, there is an opportunity to access and curate a dataset to make sure that it's applicable for this task.

Response: The goal for the current work is to better understand and address biases identified in recent high-profile work using highly-popular chest x-ray datasets, rather than identifying new biases in other domains altogether. In particular, we are unaware of AI results showing that race can be predicted from brain MRI, and we believe that training models to do so in this work would be counterproductive and may be deemed especially controversial given the modality (i.e., discriminating between brain images of different races). With our focus thus on popular chest x-ray datasets, we have aimed to ensure sufficient handling of potential confounders as suggested by the Reviewer. These important comments raised pertain to the use of DICOM images and sufficient curation of view information. Beyond these factors, we have additionally now controlled for potential confounders of age, disease prevalence, sex, and BMI, all of which resulted in similar core findings. We also note that studying preprocessing in a highly popular yet less standardized modality than brain MRI may actually be more beneficial in that it points to the potential of improved harmonization and standardization that could potentially reduce these issues. Thus, while we certainly agree that identifying potential AI bias and sources thereof is important in all medical imaging domains, it is essential to study existing findings, especially those involving high-profile work and prominent AI datasets.
***** The reviewer is suggesting to add extra experiments on brain MRI to strengthen the conclusions of the paper. I agree with the authors that this would be out of the scope of the paper and would not benefit it in terms of the clarity of the findings.

There are also other minor comments including some papers referenced in their work have been published but an archive link is provided instead (e.g., reference 27). I recommend that the authors go through their references and make sure they're referencing the most recent publication.
Response: Thank you for pointing this out. We have reviewed all references and updated to the most recent publications we could find, including removing arXiv links when appropriate.
In terms of information presentation, I thought figures two and three were very dense and difficult to follow. I recommend better visualization of results.
Response: Thank you for this feedback. We have updated Figures 2 and 3 since the initial submission to help improve interpretability, including increasing spacing and font sizes and better harmonizing the presentation of results across CXP and MXR.
***** The above are minor points which have been addressed by the authors.

While I do want to commend the authors for mentioning that training-based data augmentation did not reduce underdiagnosis (and increased it), presenting an area that can be further researched if we can understand the robustness of underlying methods proposed by authors; overall my recommendation is that the paper is flawed in its methodology
Response: Thank you for your consideration of our manuscript and for mentioning the significance of the underlying problem and potential implications of our results. We believe that the additional experiments and discussion since the original submission confirm the validity of our methods and support the robustness of our findings.
Reviewer #2
I thank the authors for adding additional experiments to strengthen the paper. To summarize, the manuscript first posits that technical factors such as field of view, exposure and view (AP vs lateral, for example) of chest x-rays affect the detection of race in chest x-rays. For example, higher contrast in the scan makes it more likely for the race-detection model to predict "white"; on the other hand, PA views

tend to increase prediction scores for black and Asian categories. Next, the authors argue that underdiagnosis bias presented in Seyyed-Kalantari et al. can be mitigated (at least partially) by controlling for these technical factors. The authors study two approaches: 1. data augmentation with contrast (by windowing) and field of view variations, and 2. setting per-view thresholds after training the model. The conclusion is that the first approach of data augmentation did not work well, while the second approach of setting decision thresholds based on view helped reduce the underdiagnosis bias to some extent.

I believe the paper still has several weaknesses:

1. First, the observed effects of the technical factors on race detection are likely reflective of the biases inherent in the training data. Indeed, the authors seem to agree and they state: "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." (lines 149-156). This is expected behavior of an AI model. Now, the question is: is the underdiagnosis bias reported in Seyyed-Kalantari et al. due (at least partially) to these confounding technical factors? The authors perform an experiment where they perform data augmentation by randomly changing the contrast (as proxy for exposure) and zoom (as proxy for field of view) while training the model. Assuming that windowing and zooming faithfully mimics the effects of exposure and field of view, respectively, if exposure and field of view were responsible for the observed underdiagnosis bias, one would think that the underdiagnosis bias would be reduced by this data augmentation strategy since done correctly, the data augmentation would ensure that the distribution of zoom and contrast were independent of the race. (As a sanity check, the authors should check if this removes the effect of these technical parameters in the race detection AI algorithm). Instead, the underdiagnosis bias persists and even slightly increases, suggesting that the technical factors are likely not responsible for the underdiagnosis bias observed in Seyyed-Kalantari et al. and reproduced here. In my mind, this undercuts the main stated takeaway of the paper: the importance of considering technical factors as they relate to race-based bias in AI models. The authors do not consider the view (AP vs PA) in this experiment -- so we are not sure if view has a role in the underdiagnosis bias -- the authors could have resampled the training data so that all races were equally represented in each view to determine its effect, but the authors do not perform that experiment.

Response: We are encouraged that the Reviewer seems to agree that biases exist regarding the technical parameters and that this could contribute to the effects observed on race prediction, where previously it was suggested that other confounders such as age, sex, disease labels, and BMI might instead explain the observed effects. It is certainly not obvious that this would be the case, and we believe it supports a core message of our work on the underappreciation of technical preprocessing and acquisition parameters. In terms of the potential contribution of these factors to the underdiagnosis bias observed by Seyyed-Kalantari et al., the Reviewer rightfully suggests that there could be many possible approaches to try to mitigate these biases. We certainly do not claim that our methods are the only ones, but we were motivated to choose practical approaches that encompass both training and inference-time strategies. Inference-time strategies such as our threshold approach are particularly attractive compared to e.g. training data resampling given that they don’t require training a new model and can more easily be adapted to new sites. As the Reviewer mentions in a later comment, we do indeed show that this method reduces underdiagnosis bias reported in Seyyed-Kalantari et al.. Regarding the data augmentation strategy, we commented on possible explanations in the Discussion (lines 337-340) and note that even if this strategy did not reduce the underdiagnosis bias, the reduction in bias through our

per-view threshold strategy still supports the importance of considering technical factors as they relate to race-based bias in AI models.
*****The reviewer makes a good point here about the impact of other factors such as view angle on the results. I agree that the results of this paper are not completely clear-cut and unambiguous. But the authors also acknowledge this and discuss possible reasons for the results in their paper. So personally I do not see this ambiguity as a flaw in the methodology, just a slightly surprising finding that has been satisfactorily discussed in the text.
-----
2. I do not understand the motivation of using a race-blind factor to tune thresholds. Why not just use different thresholds for the different races in order to combat underdiagnosis bias? In fact, looking at Extended data table 3, it seems that in most cases, the higher sensitivity for whites comes at the cost of lower specificity (irrespective of the approach), suggesting that the performance in terms of metrics such as AUROC are similar across every race. The simplest approach would be choose the operating point based on race and mitigate the underdiagnosis bias.
Response: Race-based medicine has a very controversial history with many recent efforts pointing to its flaws. Beyond the ethical concerns, such an approach would not be practical or likely effective. Indeed, we've shown for instance that there are biases in views by race in the studied clinical datasets, making this a confounder that could lead to changes in performance if thresholds were simply chosen based on race and these view distributions changed over time or across clinical sites. Race, as opposed to view, is also not readily available in the DICOM images which would be used for inference in clinical practice; furthermore, regulatory agencies would likely express strong concern over approving products with race-based thresholds.
***** The reviewer makes a sensible suggestion about the use of race-specific thresholds. The authors make two points in response: the ethical aspect and the practicality. I accept that there is at least a debate to be had about the ethics of such an approach. The authors' answer about the practicality is also reasonable so overall I am happy with the response.
-----
3. Before using the view as the basis of setting the decision thresholds, the authors should check if it indeed has any effect on the underdiagnosis bias (see comment 1). It seems to me that if one stratifies the patients into multiple subgroups and then optimizes the thresholds in each subgroup to minimize race-based bias, one would observe less bias overall, even if the factor on which the subgroups were made was completely unrelated. The fact that using per-view thresholds seems to show less underdiagnosis bias
Response: It appears that this comment may have been cut off or incomplete, but the start of the last sentence seems to concur that the per-view thresholds show less underdiagnosis bias. For the suggestion of optimizing thresholds for each patient subgroup, we assume this is referring to the previous suggestion of using different thresholds for different race subgroups, which has a number of drawbacks as detailed above.
***** Here, the original comment is not clear to me and does appear to be incomplete.
-----
4. The authors claim that their method reduces bias by 50%. This is technically true, but we should remember that the bias in the baseline was only a few percent in sensitivity (at the cost of lower specificity), and also that most of these disappear when controlled for age, sex, BMI, etc (Extended data table 4). As I mentioned in comment 1, the results suggest that the technical factors have little

influence/effect on the underdiagnosis bias (if there is one in the first place.)

Response: We apologize if it was unclear, but the values in Extended Data Table 4 reflect the distribution of the original and resampled test sets and not the AI model's predictions. When controlling for age, sex, and BMI (such as in the resampled test sets), we do indeed still observe sensitivity disparities that are reduced by the per-view threshold approach (e.g., Extended Data Figure 3).

***** The authors' response to this minor point is satisfactory.

Reviewer #3

CONTEXT:

I have been brought in as an additional reviewer to replace Reviewer 1, that was no longer able to comment on the paper. Thus, my first set of comments aim to assess whether the previous reviewer's comments have been appropriately taken into account. The authors' "Response to Reviewers" didn't actually include any responses to Reviewer 1, which seems odd -- if these were indeed supposed to be there, I was not able to find them. I also was not able to find the review from Reviewer 2, but I have seen those parts of it that were addressed in the Response to Reviewers.

Response: Thank you for reviewing our manuscript. It is our understanding that a system glitch caused our prior response to Reviewer 1 to not be available to you. Please find this response above. We welcome your comments to this response.

I will, moreover, add two major concerns regarding the authors' modelling objective and the used datasets.

SUMMARY:

This paper studies the extent to which acquisition choices are the source of racial biases observed in state-of-the-art chest X-ray diagnostic AI. While this is a very important question, I find the authors' analysis too superficial and making conclusions that the analysis does not sufficiently justify. In particular, as partially also pointed out by Reviewer 1, the used datasets are notorious for their hidden biases, and I don't think confounders are taken into account to the degree necessary to support conclusions that could actually affect how X-ray acquisition is done in the clinic.

Response: Thank you for the opportunity to clarify the goals and specific conclusions of our work. We agree that potential hidden biases in highly-popular AI datasets is of critical importance. In fact, this consideration was a core motivation for our work in studying if underappreciated "technical" biases exist in these popular datasets and if these factors may help partially explain previous high-profile AI work using these datasets, as we describe in lines 294-307). Thus, our focus is driven from an AI perspective and we further comment on hidden biases below. Separately, we agree that there are important considerations for how x-ray acquisition is done in the clinic as the Reviewer specifically mentions, but we do not intend for our conclusions to directly inform clinical x-ray acquisition and have not made such claims. We have added the following text in lines 389-393 to make this more explicit (updates highlighted in blue):

"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."

Instead our conclusions are summarized in the first paragraph of the Discussion as follows, with updates highlighted in blue which were made to further emphasize the focus on AI and popular datasets:

"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."

Thus, explicitly we conclude from our results that a) "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" and b) "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". We believe that these conclusions are directly supported by our analysis showing that the race prediction models show significant changes in predictions when varying the studied technical factors (e.g., Figures 2 and 3) and that our per-view threshold strategy reduces the underdiagnosis bias (e.g., Figure 4) reported by Seyyed-Kalantari et al. (Nature Medicine, 2021). Importantly, as detailed in the "Analysis of potential confounder factors" results section, these core findings persist even when controlling for numerous potential confounders/hidden biases of age, sex, disease prevalence, BMI, and DICOM-based processing using multiple strategies.

*****The reviewer's point here seems to be related to the points made by Reviewer #1, which I have already commented on above. In the above text, I do not think Reviewer #3 adds any further specific criticisms supported by detailed evidence, but rather expresses an opinion on the points of Reviewer #1. Personally, I find the authors' responses to Reviewer #1 and to this point by Reviewer #3 satisfactory.

Thus: While I find the studied problem extremely important, I don't think the concerns of Reviewer 1 are taken sufficiently into account, and I unfortunately have to add some concerns of my own. As a result of these concerns, I don't think the drawn conclusions -- which are what brings the paper to the level of interest of Nature Communications – are sufficiently supported to be published at present.

*** FOLLOW-UP ON COMMENTS BY THE PREVIOUS REVIEWER ***

The previous reviewer had 4 main concerns regarding the appropriateness of the study. I will list these 4 main concerns of this reviewer along with my assessment of the authors' adaptations to the concerns.

Response: Please find our complete response to Reviewer 1 above. We additionally respond to the specific comments regarding datasets in the section below. We also note that Reviewer 1's comments were made prior to much of the existing confounder analysis, as detailed in our response to their comments above.

***** As stated above, I find the authors' responses to Reviewer #1 satisfactory.

-----

1. The use of JPEG image format in place of the original DICOM, which is particularly important when you want to address image acquisition. The authors repeated the experiments using DICOM images and saw similar results.
2. The branding of zooming and windowing as preprocessing I don't see any changes in this regard, but to me this also isn't an important concern.

3. Whether there are hidden biases in the dataset selection, as there is an unexpectedly large proportion of lateral images in ICU. I don't see any comment on this, and I do think hidden biases are an important concern -- please see my further concerns below.
4. The reviewer recommended including experiments on brain MRI to validate the method in a modality where the entire process is more controlled
The authors have not included such an analysis, and also do not seem to comment on it. To me, if the flaws of the chest X-ray analysis could be brought down, that would make the need for another dataset less prominent. But this is difficult.
*** FURTHER CONCERNS ***
I have two important further concerns regarding the potential hidden biases in the used datasets, as well as with the paper's motivation.
1. Potential hidden biases
Chest X-ray datasets are notorious for their potential built-in hidden biases, which include but are not limited to: a) Known errors in diagnostic labels, which are inferred using NLP tools [1]. These errors might have a racial bias -- which could happen e.g. if one group has more follow-up scans (associated with higher error) than another group. Such biases could automatically recalibrate the algorithm towards over- or underdiagnosis. Such biases would also make a proper assessment of group-wise performance -- as carried out in this paper -- impossible.
Response: We agree that there may be several sources of bias contributing to the group-wise performance differences observed by Seyyed-Kalantari et al., who reported an underdiagnosis bias when using standard AI approaches and standard AI datasets. To reiterate, our goal was to assess if potential hidden biases related to technical parameters, which are underappreciated by the AI community and offer a means for reducing AI diagnostic performance bias, partially contribute to these results. We indeed find that our strategy to address a previously-unreported technical bias significantly reduces the underdiagnosis bias reported by Seyyed-Kalantari et al., even when controlling for a number of possible confounders. However, as we acknowledge in lines 331-332, this strategy did not completely eliminate the performance bias, leaving room for improvement for other strategies and for identifying other sources of bias that may contribute, including potentially label noise/bias. We have now expanded upon this consideration regarding NLP-based labeling by including the following sentence in line 375: “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.” We note also that the diagnostic labels are not used for the racial identity prediction analysis (e.g., Figures 2 and 3) and thus only apply to the diagnostic performance analysis (e.g., Figure 4).
We note also that the original CheXpert paper (Irvin et al., 2019) included a validation of their NLP tool compared to radiologists, which demonstrated high performance (average F1 scores > 0.90) that exceeded prior NLP tools, including that of the NIH dataset mentioned in the referenced blog post [1] by Dr. Oakden-Rayner. In this blog post, Dr. Oakden-Rayner states “Unlike the earlier NIH paper, where they did not initially test the performance of their labeller against human labelled cases in their dataset, the Stanford team produced a dataset of 1000 hand labelled reports, and the MIT team produced a dataset of 687 hand labelled reports. These results are believable, and show very good performance across the board.” While they certainly acknowledge limitations, they also state “This is a first impressions post, so I don’t want to make a strong judgement too early, but I think it is pretty safe to say that this dataset is the

best quality chest x-ray dataset we currently have." Thus, while no AI dataset is perfect, this dataset is highly used by the AI community including in the referenced high-profile work, supporting the importance of analysis on these datasets.

***** I think that the authors' response to the reviewer's point about possible errors in diagnostic labels is satisfactory.

-----

b) Potential group-wise differences in disease prevalence or -severity.

Response: Potential group-wise differences in disease prevalence is certainly an important consideration. We have accounted for this in our confounder analyses, where we find that the observed effects remain even when performing training and test set resampling to control for group-wise differences in this factor (please refer to lines 221-266 and Extended Figures 1 and 2).

***** Likewise, the response here is satisfactory.

-----

c) Potential group-wise differences in the use of support devices, which are known for their ability to act as shortcuts for algorithms [2].

Response: Thank you for this suggestion. We have now examined the frequency of support devices per subgroup and observe similar frequencies across subgroups for both datasets. For CXP, the frequencies are 94.0%, 94.3%, and 93.9% per image for White, Black, and Asian patients respectively. For MXR, the frequencies are 94.2%, 93.3%, and 91.9% for White, Black, and Asian patients respectively. We have now added these numbers in lines 592-595. While the support devices frequencies are similar across subgroups, the notion of shortcut connections is indeed an important consideration and was a motivating factor for the per-view threshold approach, as we describe in lines 314-325:

"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 correlations and even shortcut connections between the view type and the presence of pathological findings. Indeed, we do find that AI models trained to predict pathological findings exhibit different score distributions for different views (Extended Data Figure 4). This observation can help explain why choosing score thresholds per view can help mitigate the underdiagnosis bias."

***** In response to this point about groupwise differences in the use of support devices, the authors have performed some further analysis so I think their response is satisfactory.

-----

d) Potential group-wise differences in the effective dataset size -- if there are generally more views included for one group than another, its effective size goes down, which could affect both training and testing.

Response: The relative amount of data available for each subgroup is certainly an important question, where public AI datasets, including those studied, are notoriously skewed towards White patients. Regarding effective dataset size, we observe similar subgroup proportions whether calculating by patient or view for both datasets. For MXR, the percentages are 67.2% White, 17.4% Black, 3.8% Asian by patient and 68.4% White, 17.4% Black, 3.5% Asian by view. For the CXP dataset, the subgroup percentages are 63.6% White, 12.2% Asian, and 5.4% Black by patient and 63.8% White, 11.8% Asian,

6.1% Black by view. Thus, the “effective” and “absolute” dataset sizes are similar. We have now added these numbers to lines 573-577 in the Methods to complement the per view numbers reported in Extended Data Table 2. Regarding absolute dataset sizes, we note that the observed effects of the technical parameters remained when performing resampling to equalize subgroup proportions, as described in the “Analysis of potential confounding factors” section.
***** The authors have also performed some further analysis in response to this point about dataset size, so I find their answer satisfactory.

The authors don't even provide group-wise numbers that allow us as readers to assess whether such hidden biases might be affecting the algorithm, which leaves me concerned. The label errors are particularly problematic -- I don't think this dataset is suitable for doing any analyses that inform actual real-world choices unless the disease labels are revisited and performed manually by a qualified clinician, at least on the test set.
Response: In terms of group-wise numbers, we note that such values are reported across views in Extended Data Table 2 and by age, sex, BMI, and disease prevalence in Extended Data Table 4. Based on the prior comment, we have now also added numbers for supported devices and relative data sizes which show similar patterns across subgroups. This adds to our prior analysis controlling for potential hidden biases in age, sex, BMI, disease prevalence, and DICOM processing using multiple approaches, where this level of confounder/hidden bias analysis exceeds that of the referenced related work in the field.
We additionally highlight that the studied datasets are actively being used to develop AI to inform real-world choices and are supported by thousands of citations. Thus, the concern of other potential hidden biases and diagnostic label errors is not specific to our work, but any work using these and related datasets which have become benchmarks in the field. To this end, our efforts align with this core message of carefully considering the construction and subsequent use of AI datasets, where we specifically focus on technical factors for the reasons described above. We certainly agree that manual labeling of these datasets would be useful for the entire AI community, but this would require clinician review of tens of thousands of images (even for the test sets) which we believe is not within the scope of the current work. Given these considerations, in addition to the changes previously described, we have made text changes in lines 48, 163, 286, 291, and 401 to reiterate the focus and importance of studying popular AI datasets and their potential hidden biases.
***** The response here is also satisfactory in my opinion. The authors pointed to tables where some of the requested is provided. Regarding the suitability of this (widely used) dataset for performing such research, I agree with the reviewer and authors that relabelling would be useful but this is certainly beyond the scope of this paper. There are clearly issues with some of the datasets used in the work but I don’t think this invalidates the work – the issues just need to be acknowledged and the conclusions drawn should be qualified by the limitations of the datasets. I think the authors have done this.
-----
2. The paper's motivation
The paper is motivated by AI algorithms' ability to recognize race from chest X-ray images. While I was, as the rest of the community, surprised to see this, I disagree with the narrative that paints this as a problem that you want to remove. Consider for a second that disease X has different prevalences between different groups. If this is the case, then the diagnosis label itself will be enough to predict race above chance. Which means that an algorithm that is *unable* to predict race, necessarily has to predict

equal disease prevalance across races. If the true prevalence is different across races, the algorithm has no choice but to have a racial performance bias. In other words, reducing the algorithm's ability to predict race is not necessarily good for its ability to predict disease with equal performance across race. Please see [3] for further details. In their study, the authors do actually verify that their performance does not go down -- but their discussion does not reflect this potential challenge. If this paper is to be published in Nature Communications, I think it needs to make sure that this motivating factor is not misrepresented -- otherwise, they risk inspiring the development of more biased methods in our community.
Response: We agree that removing the ability to recognize race does not ensure fairness. We certainly would like to be clear that this was not the goal of our work, and we were careful not to make this claim at any point throughout the manuscript. To avoid any potential misinterpretation, we have made this more explicit in lines 307-310 in the Discussion, which we include below (updates are in blue). Thank you for the opportunity to clarify.
“As such, our goal was not to elucidate all of the features enabling AI-based race prediction, but instead focusing 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.”
*****The reviewer makes a good point that not all bias is bad. But I agree with the authors that they were not claiming this. The additional text added now makes this even clearer.
-----
3. Details
I think the authors are sometimes interpreting too much from their quantitative results. An example is the discussion of Extended Data Figure 6, where the authors write "For instance, the average image for White patients has relatively high contrast between lung and non-lung regions, which is qualitatively similar to the observed effect of an increased average White prediction score when the window width is decreased." I don't see any racial differences in the contrast of the average images. However, it seems very likely that the differences observed in Extended Data Figure 6 could be caused by Asian patients on average having a lower BMI than Black and White patients. Also, there is no colorbar, which makes it very hard to interpret the scale of the shown differences. I don't myself see any visual difference between the different average images.
Response: Thank you for this input. The referenced sentence was added after the last round of review based on a comment from another reviewer, but we agree that it is not necessary. We have now removed this line. We have also added a colorbar and a version of the plot based on the resampled MXR test set that controls for BMI shifts. The differences across the mean images are indeed subtle but reach relative magnitudes of ~10-20%, and similar overall patterns are observed when controlling for BMI.
***** This response is satisfactory.

References:
[1]
https://laurenoakdenrayner.com/2019/02/25/half-a-million-x-rays-first-impressions-of-thestanford-and-mit-chest-x-ray-datasets/
[2] Oakden-Rayner, Luke, et al. "Hidden stratification causes clinically meaningful
failures in machine learning for medical imaging. arXiv." (2019).

{3] Petersen, Eike, et al. "Are demographically invariant models and representations in medical imaging fair?." arXiv preprint arXiv:2305.01397 (2023).

Reviewer #4

I am reviewing this paper for this first time and have been asked to consider whether the authors have satisfactorily responded to the concerns of the last review.

Response: Thank you for reviewing our manuscript. It is our understanding that a system glitch may have caused only our response to Reviewer 2 to be available to you and not our response to Reviewer 1. Please also find this response above. We welcome your comments to this response.

I found the paper to be interesting, well-written and scientifically sound. The issues of AI bias and prediction of race from medical images are certainly topical and important, and I believe that the investigation performed in this paper makes a useful contribution to this area.

The previous reviewer's comments focus on the following points:

(1) Can an association between patient weight/BMI and both race and acquisition parameters be having a confounding effect on the results?

(2) Lack of discussion of possible underlying reasons for results found.

(3) Lack of generalization of per-view thresholding approach.

(4) Modest reduction in sensitivity disparity, and how is specificity affected?

In response to (1), the authors have added an extra experiment which controls for BMI and found similar results to their main analysis. This is a satisfactory response in my opinion.

In response to (2), the authors highlighted several parts of the paper in which such discussion was included, and slightly expanded this discussion. I agree with the authors that they have now sufficiently discussed this issue.

In response to (3), the authors pointed out that the calibration of models to other domains is a common issue in AI and not specific to their work. I agree with the authors that this is the case and that it does not significantly impact their findings. Other work has shown that fairness metrics do not always generalize well in the presence of other forms of domain shift and this is an open research question, but beyond the scope of this paper. If they wanted to, the authors could cite https://doi.org/10.48550/arXiv.2202.01034 as an example of such work.

Response: Thank you for pointing us to this interesting reference, which we agree is useful to include and we have now done so (it appears as reference 49).

In response to (4), the authors have now reported results which show that disparities in specificity were not significant. I am satisfied with their response to this point.

Overall, I believe that the authors have responded well to all concerns and I would be happy for the paper to be published.

I just found one minor typo in the caption of Figure 2 – "racial identify" should be "racial identity".

Response: Thank you very much for your positive feedback of the manuscript and our response to Reviewer 2. Thank you also for catching the typo in Figure 2, which we have now fixed.

***** All responses here are satisfactory.

Response to Reviewers

We thank the Reviewers for their comments and careful consideration of our manuscript. We provide a point-by-point reply to these comments below in **bold**. New text added in response to the comments is highlighted in **blue**.

Reviewer #2

I thank the authors for their responses to my earlier comments. Unfortunately, I still think the paper's claims are not adequately justified by the experiments.

1. It is not clear to me that zoom and contrast are the main factors driving the observed biases -- if they were, the data augmentation strategy (which should, in theory, make the zoom and contrast independent of race) should have reduced bias -- but it did not. This leads me to believe that other confounding factors are likely responsible.

Response: We have expanded upon this point in the discussion as follows (lines 332-341):

"In contrast to the score threshold strategy, we do 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 changes specifically related to x-ray image acquisition and processing. 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."

2. It is not clear how well the cropping and windowing simulate the actual field of view and exposure -- for example, higher radiation dose results in higher SNR, which is not possible to get with windowing

Response: We have expanded upon this point in the discussion as follows (lines 346-350):

"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."

3. Some of the figures are really not convincing -- for example, Extended Data Figure 4 -- on the CXP dataset, the score distribution is heavily tilted towards the "findings present" class. This likely suggests the scores are heavily miscalibrated (even compared to MXR). As such I am not sure how to interpret the claims about the scores increasing or decreasing when a particular parameter such as zoom or window is changed.

Response: We apologize for any lack of clarity but the increasing/decreasing score analysis and claims (e.g. Figure 2) pertain to the models trained to predict patient race (race prediction models), whereas Extended Data Figure 4 pertains to the models trained to predict the presence of pathologies (diagnostic models). We have modified the caption of Extended Data Figure 4 to increase clarity as follows:

"A kernel density estimate of the model score is shown for baseline diagnostic models trained and tested on each dataset. The score corresponds to the binary task of "No Findings" vs. "Findings Present" for the diagnostic models."

We agree that calibration is an important unsolved challenge in AI generally as discussed previously and addressed in Extended Data Figure 5 and lines 326-330 in the discussion. We also note an asymmetric score distribution does not necessarily mean a model is miscalibrated and the evaluation metrics used here (AUROC, sensitivity, specificity) depend on the relative ordering of predictions rather than their absolute value per se.

Reviewer #4

I have read the authors' response carefully as well as the other reviewers' comments. My previous opinion was that the paper was scientifically sound and a useful contribution to the literature in what I believe to be an important area. Based on what I have read in this resubmission I see no reason to change this opinion. I recommend acceptance of the paper without further revision.
I have included below the full response letter of the authors, which also includes the comments of the reviewers. I have added my opinions on the points made and the authors' responses into the below text, preceded by "*****".

Reviewer #1
The authors have attempted to provide an explanation for an ongoing question of significant importance where we don't understand the underlying mechanism. They have explored an area where previous work has not demonstrated any explanation for the mechanism behind why deep learning models can identify demographics and why there is persistent underdiagnosis for minority groups.
However, there are several concerns with their methodology and presentation of results. Firstly, the data set used to evaluate their hypothesis is flawed. The use of MIMIC chest X-ray and CheXpert datasets, which are released in JPEG format rather than the original DICOM format, makes it difficult to assess their results and conclusions. This is because preprocessing has already been performed on these images – and there is known variation of minor differences of even 8 bit versus 16 bit.
Response: We thank the Reviewer for expressing this point. While our primary goal is to better understand and help mitigate bias of standard AI approaches and datasets, which rely on preprocessed images in an "AI-ready" format, we agree that it is useful to understand if this preprocessing itself causes our observed results. Thus, we have also included evaluation on the original DICOM images for the MIMIC dataset. While the JPEG version is typically used in AI

work, these DICOM images are publicly available for MIMIC, though not for CheXpert. We have specifically tested our original models, which were trained on the JPEG images, directly on the MIMIC DICOM images to more robustly test the generalization and sensitivity of our results. Following the DICOM Standard (National Electrical Manufacturers Association, VA, USA), we extract and process the MIMIC images directly from the DICOM files using the default parameters contained in the DICOM headers before AI evaluation. This analysis is contained in a section in the manuscript titled “Analysis of potential confounding factors” with accompanying Extended Data Figures 1-3. Despite the models having been trained on images with different preprocessing, we find that the overall patterns regarding the technical parameter analysis and underdiagnosis bias remain when testing on the DICOM images. For the racial identity prediction task, the AI predictions are still influenced by the technical parameters, and on the disease classification task, the baseline model still shows underdiagnosis bias which is still reduced by our view-specific threshold approach. These analyses suggest that these results are not simply caused by the preprocessing used to create the AI-ready datasets. Nonetheless, we agree that this preprocessing is an important consideration and we have expanded on this consideration in the Discussion (lines 367-375).

***** Reviewer #1 expressed concern about the impact of the data format of the images. The authors have satisfactorily addressed this point by adding an extra experiment on the original DICOM images of the MIMIC dataset, showing results consistent with their original findings.

Moreover, the processing approaches suggested by the authors include zooming and windowing. These would not typically be considered image preprocessing but rather factors that are usually changed on the view/display by the end user – usually a radiologist . Therefore, their selection of preprocessing tasks is misleading.

Response: Thank you for raising this important point. We agree that it is challenging to define a precise delineation of what is considered image “preprocessing”. Related to the previous point, there are a number of processing stages that take place from initial x-ray exposure through to when the image is actually viewed by the end user (i.e., radiologist) or even AI model. For instance, most DICOM viewers implement image processing steps according to the DICOM Standard by default, including windowing according to the default values in the DICOM header before display to the end user, which can then be adjusted as desired as rightfully stated in the Reviewer’s comment above. Similarly, preprocessing stages such as windowing, normalization, resizing, and even bit depth conversion are commonly used in AI approaches. Ultimately, our goal was to simulate technical variations that are important in chest x-ray image acquisition and processing, including overall contrast/exposure and the relative size of the field of view, which can be changed through collimation. To study the effect of contrast, we perform windowing with different window widths to produce different levels of overall contrast. To study variations in collimation, we effectively perform ‘electronic collimation’ (Tsalafoutas et al., reference 37; Bomer et al., reference 38) to modify the relative field of view for the image. We had used the term “zoom” as an intuitive, non-technical way to explain this phenomenon in the initial submission, but we recognize that this term can have different meanings in different contexts and can cause confusion. Thus, we now refer to this aspect as the “field of view” parameter instead of the “zoom” parameter and have updated its use throughout the manuscript. Altogether, we believe that the window width and field of view parameters represent highly relevant transformations for chest x-ray acquisition and preprocessing, especially in the context of AI development. In addition to updating the field of view terminology since the initial submission, we have expanded upon these points in lines 70-71, 87-88, 367-375.

***** This point relates to the description of the zooming and windowing operations as "preprocessing". The authors have adjusted their terminology so I think this point has been satisfactorily addressed.

Secondly, there is concern for sensitivity to the labels that were provided in their data. For example, if you look at extended data table two in MIMIC chest X-ray dataset, most images are acquired using portable method as expected for ICU images. However, when you look at distribution between AP and lateral views quite a large number of lateral views in ICU which would not make sense. This may indicate a need for further cleaning of datasets by authors.

Response: Thank you for the opportunity to clarify. According to the initial paper describing the MIMIC dataset (Johnson et al., 2019), the authors state the following: "We queried the BIDMC EHR for chest radiograph studies acquired in the emergency department between 2011–2016, and extracted only the set of patient identifiers associated with these studies. A collection of images associated with a single report is referred to as a study, identified by a unique identifier, the study ID. We then extracted all chest radiographs and radiology reports available in the RIS for this set of patients between 2011–2016."

Thus, while the set of patients was identified based on studies acquired in the emergency department, it is our interpretation that all studies for these patients were then extracted regardless of whether they originated from the emergency department or not. We have since expanded on the description of the MIMIC dataset in the Methods section (lines 556-560) to more fully reflect the original description of the dataset. Additionally, we realize that our original presentation of Extended Data Table 2 may have been confusing in that we displayed the breakdown of Standard vs Portable views for the AP view position, since this position is used for the vast majority of Portable views (as stated in the comment above). For improved clarity, we have added an additional row in this table since the initial submission showing the total proportion of Portable views amongst all images (33.8%) and updated the table legend with further description.

We have additionally pursued further data cleaning using three strategies to ensure that the AP and Lateral views are properly labeled as the Reviewer suggests. First, we plotted 100 random AP views and 100 random Lateral views according to the MIMIC metadata. We manually reviewed these 200 images and determined that they are all correctly labeled (e.g., there were no Lateral views that were labeled as AP and vice versa). Second, we extracted the View Position directly from the DICOM files that were obtained for the prior comment. We compared these extracted views to the metadata and there was a 100% correspondence. Third, we reviewed the code used by the dataset creators to create this dataset, which is publicly available on Github. We did not find any apparent issues through this review. While some amount of noise is likely to be expected in any clinical dataset, we believe that these efforts support the validity and proper curation of the MIMIC dataset, which is further supported by its popularity and the initial publication describing its curation.

***** This point relates to the concerns about the data labelling. The authors have made changes to the descriptions of the data and performed some further verification of label fidelity. In my opinion, this is a satisfactory response.

I recommend that the authors try a different modality such as brain MRI to demonstrate this process of preprocessing and underdiagnosis. The reason for this recommendation is because there is more harmonization and standardization of brain imaging. If there is a desire to work on chest X-rays, there is an opportunity to access and curate a dataset to make sure that it's applicable for this task.

Response: The goal for the current work is to better understand and address biases identified in recent high-profile work using highly-popular chest x-ray datasets, rather than identifying new biases in other domains altogether. In particular, we are unaware of AI results showing that race can be predicted from brain MRI, and we believe that training models to do so in this work would be counterproductive and may be deemed especially controversial given the modality (i.e., discriminating between brain images of different races). With our focus thus on popular chest x-ray datasets, we have aimed to ensure sufficient handling of potential confounders as suggested by the Reviewer. These important comments raised pertain to the use of DICOM images and sufficient curation of view information. Beyond these factors, we have additionally now controlled for potential confounders of age, disease prevalence, sex, and BMI, all of which resulted in similar core findings. We also note that studying preprocessing in a highly popular yet less standardized modality than brain MRI may actually be more beneficial in that it points to the potential of improved harmonization and standardization that could potentially reduce these issues. Thus, while we certainly agree that identifying potential AI bias and sources thereof is important in all medical imaging domains, it is essential to study existing findings, especially those involving high-profile work and prominent AI datasets.

***** The reviewer is suggesting to add extra experiments on brain MRI to strengthen the conclusions of the paper. I agree with the authors that this would be out of the scope of the paper and would not benefit it in terms of the clarity of the findings.

There are also other minor comments including some papers referenced in their work have been published but an archive link is provided instead (e.g., reference 27). I recommend that the authors go through their references and make sure they're referencing the most recent publication.

Response: Thank you for pointing this out. We have reviewed all references and updated to the most recent publications we could find, including removing arXiv links when appropriate.

In terms of information presentation, I thought figures two and three were very dense and difficult to follow. I recommend better visualization of results.

Response: Thank you for this feedback. We have updated Figures 2 and 3 since the initial submission to help improve interpretability, including increasing spacing and font sizes and better harmonizing the presentation of results across CXP and MXR.

***** The above are minor points which have been addressed by the authors.

While I do want to commend the authors for mentioning that training-based data augmentation did not reduce underdiagnosis (and increased it), presenting an area that can be further researched if we can understand the robustness of underlying methods proposed by authors; overall my recommendation is that the paper is flawed in its methodology

Response: Thank you for your consideration of our manuscript and for mentioning the significance of the underlying problem and potential implications of our results. We believe that the additional experiments and discussion since the original submission confirm the validity of our methods and support the robustness of our findings.

Reviewer #2

I thank the authors for adding additional experiments to strengthen the paper. To summarize, the manuscript first posits that technical factors such as field of view, exposure and view (AP vs lateral, for example) of chest x-rays affect the detection of race in chest x-rays. For example, higher contrast in the scan makes it more likely for the race-detection model to predict "white"; on the other hand, PA views tend to increase prediction scores for black and Asian categories. Next, the authors argue that underdiagnosis bias presented in Seyyed-Kalantari et al. can be mitigated (at least partially) by controlling for these technical factors. The authors study two approaches: 1. data augmentation with contrast (by windowing) and field of view variations, and 2. setting per-view thresholds after training the model. The conclusion is that the first approach of data augmentation did not work well, while the second approach of setting decision thresholds based on view helped reduce the underdiagnosis bias to some extent.

I believe the paper still has several weaknesses:

1. First, the observed effects of the technical factors on race detection are likely reflective of the biases inherent in the training data. Indeed, the authors seem to agree and they state: "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." (lines 149-156). This is expected behavior of an AI model. Now, the question is: is the underdiagnosis bias reported in Seyyed-Kalantari et al. due (at least partially) to these confounding technical factors? The authors perform an experiment where they perform data augmentation by randomly changing the contrast (as proxy for exposure) and zoom (as proxy for field of view) while training the model. Assuming that windowing and zooming faithfully mimics the effects of exposure and field of view, respectively, if exposure and field of view were responsible for the observed underdiagnosis bias, one would think that the underdiagnosis bias would be reduced by this data augmentation strategy since done correctly, the data augmentation would ensure that the distribution of zoom and contrast were independent of the race. (As a sanity check, the authors should check if this removes the effect of these technical parameters in the race detection AI algorithm). Instead, the underdiagnosis bias persists and even slightly increases, suggesting that the technical factors are likely not responsible for the underdiagnosis bias observed in Seyyed-Kalantari et al. and reproduced here. In my mind, this undercuts the main stated takeaway of the paper: the importance of considering technical factors as they relate to race-based bias in AI models. The authors do not consider the view (AP vs PA) in this experiment -- so we are not sure if view has a role in the underdiagnosis bias -- the authors could have resampled the training data so that all races were equally represented in each view to determine its effect, but the authors do not perform that experiment.

Response: We are encouraged that the Reviewer seems to agree that biases exist regarding the technical parameters and that this could contribute to the effects observed on race prediction, where previously it was suggested that other confounders such as age, sex, disease

labels, and BMI might instead explain the observed effects. It is certainly not obvious that this would be the case, and we believe it supports a core message of our work on the underappreciation of technical preprocessing and acquisition parameters. In terms of the potential contribution of these factors to the underdiagnosis bias observed by Seyyed-Kalantari et al., the Reviewer rightfully suggests that there could be many possible approaches to try to mitigate these biases. We certainly do not claim that our methods are the only ones, but we were motivated to choose practical approaches that encompass both training and inference-time strategies. Inference-time strategies such as our threshold approach are particularly attractive compared to e.g. training data resampling given that they don't require training a new model and can more easily be adapted to new sites. As the Reviewer mentions in a later comment, we do indeed show that this method reduces underdiagnosis bias reported in Seyyed-Kalantari et al.. Regarding the data augmentation strategy, we commented on possible explanations in the Discussion (lines 337-340) and note that even if this strategy did not reduce the underdiagnosis bias, the reduction in bias through our per-view threshold strategy still supports the importance of considering technical factors as they relate to race-based bias in AI models.
*****The reviewer makes a good point here about the impact of other factors such as view angle on the results. I agree that the results of this paper are not completely clear-cut and unambiguous. But the authors also acknowledge this and discuss possible reasons for the results in their paper. So personally I do not see this ambiguity as a flaw in the methodology, just a slightly surprising finding that has been satisfactorily discussed in the text.
-----
2. I do not understand the motivation of using a race-blind factor to tune thresholds. Why not just use different thresholds for the different races in order to combat underdiagnosis bias? In fact, looking at Extended data table 3, it seems that in most cases, the higher sensitivity for whites comes at the cost of lower specificity (irrespective of the approach), suggesting that the performance in terms of metrics such as AUROC are similar across every race. The simplest approach would be choose the operating point based on race and mitigate the underdiagnosis bias.
Response: Race-based medicine has a very controversial history with many recent efforts pointing to its flaws. Beyond the ethical concerns, such an approach would not be practical or likely effective. Indeed, we've shown for instance that there are biases in views by race in the studied clinical datasets, making this a confounder that could lead to changes in performance if thresholds were simply chosen based on race and these view distributions changed over time or across clinical sites. Race, as opposed to view, is also not readily available in the DICOM images which would be used for inference in clinical practice; furthermore, regulatory agencies would likely express strong concern over approving products with race-based thresholds.
***** The reviewer makes a sensible suggestion about the use of race-specific thresholds. The authors make two points in response: the ethical aspect and the practicality. I accept that there is at least a debate to be had about the ethics of such an approach. The authors' answer about the practicality is also reasonable so overall I am happy with the response.
-----
3. Before using the view as the basis of setting the decision thresholds, the authors should check if it indeed has any effect on the underdiagnosis bias (see comment 1). It seems to me

that if one stratifies the patients into multiple subgroups and then optimizes the thresholds in each subgroup to minimize race-based bias, one would observe less bias overall, even if the factor on which the subgroups were made was completely unrelated. The fact that using per-view thresholds seems to show less underdiagnosis bias

Response: It appears that this comment may have been cut off or incomplete, but the start of the last sentence seems to concur that the per-view thresholds show less underdiagnosis bias. For the suggestion of optimizing thresholds for each patient subgroup, we assume this is referring to the previous suggestion of using different thresholds for different race subgroups, which has a number of drawbacks as detailed above.

***** Here, the original comment is not clear to me and does appear to be incomplete.

-----

4. The authors claim that their method reduces bias by 50%. This is technically true, but we should remember that the bias in the baseline was only a few percent in sensitivity (at the cost of lower specificity), and also that most of these disappear when controlled for age, sex, BMI, etc (Extended data table 4). As I mentioned in comment 1, the results suggest that the technical factors have little influence/effect on the underdiagnosis bias (if there is one in the first place.)

Response: We apologize if it was unclear, but the values in Extended Data Table 4 reflect the distribution of the original and resampled test sets and not the AI model's predictions. When controlling for age, sex, and BMI (such as in the resampled test sets), we do indeed still observe sensitivity disparities that are reduced by the per-view threshold approach (e.g., Extended Data Figure 3).

***** The authors' response to this minor point is satisfactory.

Reviewer #3

CONTEXT:

I have been brought in as an additional reviewer to replace Reviewer 1, that was no longer able to comment on the paper. Thus, my first set of comments aim to assess whether the previous reviewer's comments have been appropriately taken into account. The authors' "Response to Reviewers" didn't actually include any responses to Reviewer 1, which seems odd -- if these were indeed supposed to be there, I was not able to find them. I also was not able to find the review from Reviewer 2, but I have seen those parts of it that were addressed in the Response to Reviewers.

Response: Thank you for reviewing our manuscript. It is our understanding that a system glitch caused our prior response to Reviewer 1 to not be available to you. Please find this response above. We welcome your comments to this response.

I will, moreover, add two major concerns regarding the authors' modelling objective and the used datasets.

SUMMARY:

This paper studies the extent to which acquisition choices are the source of racial biases observed in state-of-the-art chest X-ray diagnostic AI. While this is a very important question, I find the authors' analysis too superficial and making conclusions that the analysis does not sufficiently justify. In particular, as partially also pointed out by Reviewer 1, the used datasets are notorious for their hidden biases, and I don't think confounders are taken into account to the degree necessary to support conclusions that could actually affect how X-ray acquisition is done in the clinic.

Response: Thank you for the opportunity to clarify the goals and specific conclusions of our work. We agree that potential hidden biases in highly-popular AI datasets is of critical importance. In fact, this consideration was a core motivation for our work in studying if underappreciated "technical" biases exist in these popular datasets and if these factors may help partially explain previous high-profile AI work using these datasets, as we describe in lines 294-307). Thus, our focus is driven from an AI perspective and we further comment on hidden biases below. Separately, we agree that there are important considerations for how x-ray acquisition is done in the clinic as the Reviewer specifically mentions, but we do not intend for our conclusions to directly inform clinical x-ray acquisition and have not made such claims. We have added the following text in lines 389-393 to make this more explicit (updates highlighted in blue):

"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."

Instead our conclusions are summarized in the first paragraph of the Discussion as follows, with updates highlighted in blue which were made to further emphasize the focus on AI and popular datasets:

"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."

Thus, explicitly we conclude from our results that a) "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" and b) "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". We believe that these conclusions are directly supported by our analysis showing that the race prediction models show significant changes in predictions when varying the studied technical factors (e.g., Figures 2 and 3) and that our per-view threshold strategy reduces the underdiagnosis bias (e.g., Figure 4) reported by Seyyed-Kalantari et al. (Nature Medicine, 2021). Importantly, as detailed in the "Analysis of potential confounder factors" results section, these core findings persist even when controlling for numerous potential confounders/hidden biases of age, sex, disease prevalence, BMI, and DICOM-based processing using multiple strategies.

*****The reviewer’s point here seems to be related to the points made by Reviewer #1, which I have already commented on above. In the above text, I do not think Reviewer #3 adds any further specific criticisms supported by detailed evidence, but rather expresses an opinion on the points of Reviewer #1. Personally, I find the authors’ responses to Reviewer #1 and to this point by Reviewer #3 satisfactory.

Thus: While I find the studied problem extremely important, I don't think the concerns of Reviewer 1 are taken sufficiently into account, and I unfortunately have to add some concerns of my own. As a result of these concerns, I don't think the drawn conclusions -- which are what brings the paper to the level of interest of Nature Communications – are sufficiently supported to be published at present.

*** FOLLOW-UP ON COMMENTS BY THE PREVIOUS REVIEWER ***

The previous reviewer had 4 main concerns regarding the appropriateness of the study. I will list these 4 main concerns of this reviewer along with my assessment of the authors' adaptations to the concerns.

Response: Please find our complete response to Reviewer 1 above. We additionally respond to the specific comments regarding datasets in the section below. We also note that Reviewer 1’s comments were made prior to much of the existing confounder analysis, as detailed in our response to their comments above.

***** As stated above, I find the authors’ responses to Reviewer #1 satisfactory.

-----

1. The use of JPEG image format in place of the original DICOM, which is particularly important when you want to address image acquisition. The authors repeated the experiments using DICOM images and saw similar results.

2. The branding of zooming and windowing as preprocessing I don't see any changes in this regard, but to me this also isn't an important concern.

3. Whether there are hidden biases in the dataset selection, as there is an unexpectedly large proportion of lateral images in ICU. I don't see any comment on this, and I do think hidden biases are an important concern -- please see my further concerns below.

4. The reviewer recommended including experiments on brain MRI to validate the method in a modality where the entire process is more controlled

The authors have not included such an analysis, and also do not seem to comment on it. To me, if the flaws of the chest X-ray analysis could be brought down, that would make the need for another dataset less prominent. But this is difficult.

*** FURTHER CONCERNS ***

I have two important further concerns regarding the potential hidden biases in the used datasets, as well as with the paper's motivation.

1. Potential hidden biases

Chest X-ray datasets are notorious for their potential built-in hidden biases, which include but are not limited to: a) Known errors in diagnostic labels, which are inferred using NLP tools [1]. These errors might have a racial bias -- which could happen e.g. if one group has more follow-up scans (associated with higher error) than another group. Such biases could automatically recalibrate the algorithm towards over- or underdiagnosis. Such biases would also make a proper assessment of group-wise performance -- as carried out in this paper -- impossible.

Response: We agree that there may be several sources of bias contributing to the group-wise performance differences observed by Seyyed-Kalantari et al., who reported an underdiagnosis bias when using standard AI approaches and standard AI datasets. To reiterate, our goal was to assess if potential hidden biases related to technical parameters, which are underappreciated by the AI community and offer a means for reducing AI diagnostic performance bias, partially contribute to these results. We indeed find that our strategy to address a previously-unreported technical bias significantly reduces the underdiagnosis bias reported by Seyyed-Kalantari et al., even when controlling for a number of possible confounders. However, as we acknowledge in lines 331-332, this strategy did not completely eliminate the performance bias, leaving room for improvement for other strategies and for identifying other sources of bias that may contribute, including potentially label noise/bias. We have now expanded upon this consideration regarding NLP-based labeling by including the following sentence in line 375: "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." We note also that the diagnostic labels are not used for the racial identity prediction analysis (e.g., Figures 2 and 3) and thus only apply to the diagnostic performance analysis (e.g., Figure 4).

We note also that the original CheXpert paper (Irvin et al., 2019) included a validation of their NLP tool compared to radiologists, which demonstrated high performance (average F1 scores > 0.90) that exceeded prior NLP tools, including that of the NIH dataset mentioned in the referenced blog post [1] by Dr. Oakden-Rayner. In this blog post, Dr. Oakden-Rayner states "Unlike the earlier NIH paper, where they did not initially test the performance of their labeller against human labelled cases in their dataset, the Stanford team produced a dataset of 1000 hand labelled reports, and the MIT team produced a dataset of 687 hand labelled reports. These results are believable, and show very good performance across the board." While they certainly acknowledge limitations, they also state "This is a first impressions post, so I don't want to make a strong judgement too early, but I think it is pretty safe to say that this dataset is the best quality chest x-ray dataset we currently have." Thus, while no AI dataset is perfect, this dataset is highly used by the AI community including in the referenced high-profile work, supporting the importance of analysis on these datasets.

***** I think that the authors' response to the reviewer's point about possible errors in diagnostic labels is satisfactory.

-----

b) Potential group-wise differences in disease prevalence or -severity.

Response: Potential group-wise differences in disease prevalence is certainly an important consideration. We have accounted for this in our confounder analyses, where we find that the observed effects remain even when performing training and test set resampling to control for group-wise differences in this factor (please refer to lines 221-266 and Extended Figures 1 and 2).

***** Likewise, the response here is satisfactory.

-----

c) Potential group-wise differences in the use of support devices, which are known for their ability to act as shortcuts for algorithms [2].

Response: Thank you for this suggestion. We have now examined the frequency of support devices per subgroup and observe similar frequencies across subgroups for both datasets. For CXP, the frequencies are 94.0%, 94.3%, and 93.9% per image for White, Black, and Asian patients respectively. For MXR, the frequencies are 94.2%, 93.3%, and 91.9% for White, Black, and Asian patients respectively. We have now added these numbers in lines 592-595. While the support devices frequencies are similar across subgroups, the notion of shortcut connections is indeed an important consideration and was a motivating factor for the per-view threshold approach, as we describe in lines 314-325:

“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 correlations and even shortcut connections between the view type and the presence of pathological findings. Indeed, we do find that AI models trained to predict pathological findings exhibit different score distributions for different views (Extended Data Figure 4). This observation can help explain why choosing score thresholds per view can help mitigate the underdiagnosis bias.”

***** In response to this point about groupwise differences in the use of support devices, the authors have performed some further analysis so I think their response is satisfactory.

-----

d) Potential group-wise differences in the effective dataset size -- if there are generally more views included for one group than another, its effective size goes down, which could affect both training and testing.

Response: The relative amount of data available for each subgroup is certainly an important question, where public AI datasets, including those studied, are notoriously skewed towards White patients. Regarding effective dataset size, we observe similar subgroup proportions whether calculating by patient or view for both datasets. For MXR, the percentages are 67.2% White, 17.4% Black, 3.8% Asian by patient and 68.4% White, 17.4% Black, 3.5% Asian by view. For the CXP dataset, the subgroup percentages are 63.6% White, 12.2% Asian, and 5.4% Black by patient and 63.8% White, 11.8% Asian, 6.1% Black by view. Thus, the “effective” and “absolute” dataset sizes are similar. We have now added these numbers to lines 573-577 in the Methods to complement the per view numbers reported in Extended Data Table 2. Regarding absolute dataset sizes, we note that the observed effects of the technical parameters remained when performing resampling to equalize subgroup proportions, as described in the “Analysis of potential confounding factors” section.

***** The authors have also performed some further analysis in response to this point about dataset size, so I find their answer satisfactory.

The authors don't even provide group-wise numbers that allow us as readers to assess whether such hidden biases might be affecting the algorithm, which leaves me concerned. The label errors are particularly problematic -- I don't think this dataset is suitable for doing any analyses that inform actual real-world choices unless the disease labels are revisited and performed manually by a qualified clinician, at least on the test set.

Response: In terms of group-wise numbers, we note that such values are reported across views in Extended Data Table 2 and by age, sex, BMI, and disease prevalence in Extended Data Table 4. Based on the prior comment, we have now also added numbers for supported devices and relative data sizes which show similar patterns across subgroups. This adds to our prior analysis controlling for potential hidden biases in age, sex, BMI, disease prevalence, and DICOM processing using multiple approaches, where this level of confounder/hidden bias analysis exceeds that of the referenced related work in the field.

We additionally highlight that the studied datasets are actively being used to develop AI to inform real-world choices and are supported by thousands of citations. Thus, the concern of other potential hidden biases and diagnostic label errors is not specific to our work, but any work using these and related datasets which have become benchmarks in the field. To this end, our efforts align with this core message of carefully considering the construction and subsequent use of AI datasets, where we specifically focus on technical factors for the reasons described above. We certainly agree that manual labeling of these datasets would be useful for the entire AI community, but this would require clinician review of tens of thousands of images (even for the test sets) which we believe is not within the scope of the current work. Given these considerations, in addition to the changes previously described, we have made text changes in lines 48, 163, 286, 291, and 401 to reiterate the focus and importance of studying popular AI datasets and their potential hidden biases.

***** The response here is also satisfactory in my opinion. The authors pointed to tables where some of the requested is provided. Regarding the suitability of this (widely used) dataset for performing such research, I agree with the reviewer and authors that relabelling would be useful but this is certainly beyond the scope of this paper. There are clearly issues with some of the datasets used in the work but I don’t think this invalidates the work – the issues just need to be acknowledged and the conclusions drawn should be qualified by the limitations of the datasets. I think the authors have done this.

-----

2. The paper's motivation

The paper is motivated by AI algorithms' ability to recognize race from chest X-ray images. While I was, as the rest of the community, surprised to see this, I disagree with the narrative that paints this as a problem that you want to remove. Consider for a second that disease X has different prevalences between different groups. If this is the case, then the diagnosis label itself will be enough to predict race above chance. Which means that an algorithm that is *unable* to predict race, necessarily has to predict equal disease prevalance across races. If the true prevalence is different across races, the algorithm has no choice but to have a racial performance bias. In other words, reducing the algorithm's ability to predict race is not necessarily good for its ability to predict disease with equal performance across race. Please see [3] for further details. In their study, the authors do actually verify that their performance does not go down -- but their discussion does not reflect this potential challenge. If this paper is to be published in Nature Communications, I think it needs to make sure that this motivating factor is not misrepresented -- otherwise, they risk inspiring the development of more biased methods in our community.

Response: We agree that removing the ability to recognize race does not ensure fairness. We certainly would like to be clear that this was not the goal of our work, and we were careful not to make this claim at any point throughout the manuscript. To avoid any potential misinterpretation, we have made this more explicit in lines 307-310 in the Discussion, which we include below (updates are in blue). Thank you for the opportunity to clarify.

“As such, our goal was not to elucidate all of the features enabling AI-based race prediction, but instead focusing 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.”

*****The reviewer makes a good point that not all bias is bad. But I agree with the authors that they were not claiming this. The additional text added now makes this even clearer.

-----

3. Details

I think the authors are sometimes interpreting too much from their quantitative results. An example is the discussion of Extended Data Figure 6, where the authors write "For instance, the average image for White patients has relatively high contrast between lung and non-lung regions, which is qualitatively similar to the observed effect of an increased average White prediction score when the window width is decreased." I don't see any racial differences in the contrast of the average images. However, it seems very likely that the differences observed in Extended Data Figure 6 could be caused by Asian patients on average having a lower BMI than Black and White patients. Also, there is no colorbar, which makes it very hard to interpret the scale of the shown differences. I don't myself see any visual difference between the different average images.

Response: Thank you for this input. The referenced sentence was added after the last round of review based on a comment from another reviewer, but we agree that it is not necessary. We have now removed this line. We have also added a colorbar and a version of the plot based on the resampled MXR test set that controls for BMI shifts. The differences across the mean images are indeed subtle but reach relative magnitudes of ~10-20%, and similar overall patterns are observed when controlling for BMI.

***** This response is satisfactory.

References:

[1] https://laurenoakdenrayner.com/2019/02/25/half-a-million-x-rays-first-impressions-of-thestanford-and-mit-chest-x-ray-datasets/

[2] Oakden-Rayner, Luke, et al. "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. arXiv." (2019).

{3] Petersen, Eike, et al. "Are demographically invariant models and representations in medical imaging fair?." arXiv preprint arXiv:2305.01397 (2023).

Reviewer #4

I am reviewing this paper for this first time and have been asked to consider whether the authors have satisfactorily responded to the concerns of the last review.

Response: Thank you for reviewing our manuscript. It is our understanding that a system glitch may have caused only our response to Reviewer 2 to be available to you and not our response to Reviewer 1. Please also find this response above. We welcome your comments to this response.

I found the paper to be interesting, well-written and scientifically sound. The issues of AI bias and prediction of race from medical images are certainly topical and important, and I believe that the investigation performed in this paper makes a useful contribution to this area.

The previous reviewer's comments focus on the following points:

(1) Can an association between patient weight/BMI and both race and acquisition parameters be having a confounding effect on the results?

(2) Lack of discussion of possible underlying reasons for results found.

(3) Lack of generalization of per-view thresholding approach.

(4) Modest reduction in sensitivity disparity, and how is specificity affected?

In response to (1), the authors have added an extra experiment which controls for BMI and found similar results to their main analysis. This is a satisfactory response in my opinion.

In response to (2), the authors highlighted several parts of the paper in which such discussion was included, and slightly expanded this discussion. I agree with the authors that they have now sufficiently discussed this issue.

In response to (3), the authors pointed out that the calibration of models to other domains is a common issue in AI and not specific to their work. I agree with the authors that this is the case and that it does not significantly impact their findings. Other work has shown that fairness metrics do not always generalize well in the presence of other forms of domain shift and this is an open research question, but beyond the scope of this paper. If they wanted to, the authors could cite https://doi.org/10.48550/arXiv.2202.01034 as an example of such work.

Response: Thank you for pointing us to this interesting reference, which we agree is useful to include and we have now done so (it appears as reference 49).

In response to (4), the authors have now reported results which show that disparities in specificity were not significant. I am satisfied with their response to this point.

Overall, I believe that the authors have responded well to all concerns and I would be happy for the paper to be published.

I just found one minor typo in the caption of Figure 2 – "racial identify" should be "racial identity".

Response: Thank you very much for your positive feedback of the manuscript and our response to Reviewer 2. Thank you also for catching the typo in Figure 2, which we have now fixed.

***** All responses here are satisfactory.

Response: We sincerely thank the Reviewer for their time and effort in reviewing all of these comments.

REVIEWERS' COMMENTS

Reviewer #2 (Remarks to the Author):

I thank the reviewers for their responses and their revisions. They are satisfactory to me.