Fix and add link to git [MISSING]

The racial effect in radiographs interpretation   
Introduction  
The performance of deep learning models in the field of medical imaging has reached or even exceeded human-level performance, especially when it comes to diagnosing disease using chest X-rays. However, neural networks often learn to make predictions that overly rely on spurious correlation existing in the dataset, which causes the model to be biased. This kind of bias is often difficult to identify, due to the lake of explainability of such classifiers. As computer vision systems are deployed at scale in variety of settings, it becomes increasingly important to be aware of such drawbacks, especially in the medical domain.  
Previous studies in medical imaging have shown disparate abilities of deep learning models to detect a person’s race, yet there is no known correlation for race on medical imaging that would be obvious to human experts when interpreting the images.   
Our work aims to understand the impact race has on X-ray medical imaging diagnostics in deep learning models. To this extent, we’re utilizing two popular large-scale X-ray datasets: CheXpert and MIMIC-CXR.

More on problem [MISSING]  
More on experiments and methods [MISSING]

(1) to assess the ability of deep learning AI models to recognise race from medical images, including the ability of these models to generalise to new environments and across multiple imaging modalities; (2) to examine possible confounding anatomic and phenotype population features as explanations for these performance scores, and (3) to investigate the underlying mechanisms by which AI models can recognise race.

The results from our study emphasise that the ability of AI deep learning models to predict self-reported race is itself not the issue of importance. However, our finding that AI can accurately predict self-reported race, even from corrupted, cropped, and noised medical images, often when clinical experts cannot, creates an enormous risk for all model deployments in medical imaging [MISSING]

# Background and related work

**Bias.** Biased machine learning models is a topic of increasing attention, classifier biases have been discovered in various ranging domains, such as racial bias in criminal defendant risk assessments [17 MISSING] and gender bias in online recruiting [16 MISSING]. Studies in the medical domain have shown similar results in many health-care applications, such as mortality prediction [7 MISSING], and melanoma detection [5,6 MISSING].  
Sources of bias may originate in different points along the classical machine learning pipeline. Considering minority groups, collected features may not be as indicative compared to the rest of the population. This phenomenon of deep learning models to learn spurious correlations existing in the data is often a major obstacle in promising generalization. This is evident in the struggle of deep learning models to generalize on out-of-distribution data [MISSING].

The focus in our work [MISSING]

**Fairness.** Fairness has been conceptualized mathematically and philosophically in a variety of ways, such as error rate balance [MISSING], worst-case group accuracy [MISSING], and fairness through unawareness. There are several conflicting definitions of fairness, many of which are not simultaneously achievable, the appropriate choice of a disparity metric is generally task dependent. With the advancement of deep learning technology, artificial intelligence and decision support systems become more popular, the idea of unintentionally relying on protected attributes (such as race, gender, age, etc.) could be alarming, especially, in sensitive environments such as hospitals.

In our contest [MISSING]

**Protected Attributes Detection.** It’s been shown that deep learning algorithms can identify various patient demographic attributes, even when these do not form part of the input. For example, DK et al. [MISSING] utilized such abilities to improve a radiologist performance in skeletal age assessment. PH et al. [MISSING] Illustrated how such classifiers able to determine the age and sex of patients from chest radiographs. Furthermore, a study by Gichoya at el. demonstrates how race could be accurately identified from X-ray images alone, an ability that is unexplainable by physicians. Moreover, Gichoya at el. study shows that even with severe data augmentation the race identification abilities almost didn’t decrease.

Findings regarding the possibility of confounding of racial identity in deep learning models suggest a possible mechanism for racial disparities resulting from AI models: that AI models can directly recognise the race of a patient from medical images. However, this hypothesis is largely unexplored18 and, in contrast to other demographic factors (eg, age and sex), there is a widely held, but tacit, belief among radiologists that the identification of a patient’s race from medical images is almost impossible, and that most medical imaging tasks are essentially race agnostic (ie, the task is not affected by the patient’s race).

The results from our study emphasise that the ability of AI deep learning models to predict self-reported race is itself not the issue of importance. However, our finding that AI can accurately predict self-reported race, even from corrupted, cropped, and noised medical images, often when clinical experts cannot, creates an enormous risk for all model deployments in medical imaging [MISSING]

Our work takes one step further, beside identify race, we wanted to see if diagnosis prediction models are encoding racial features when making predictions. [MISSING]

**Self-Reported Race.** Race identity often conflated with biological constructs, in our study we define race as a social construct that pertains to how we interact with each other and how others perceive us. To this end, we use self-reported race as the racial identity of patients throughout the study.

# The Data

**CheXpert.** (Irvin et al., 2019) [MISSING] An X-ray dataset from Stanford Hospital that contains 224,316 frontal and lateral chest radiographs of 65,240 patients. The dataset includes a validation set, containing an addition 200 studies verified by a board three certificate radiologists, denoted as *Validation Set 1*. We also portion a second validation set from the training data, which we refer to as *Validation Set 2*. We elaborate more regarding the validation sets in the next sections.  
**MIMIC-CXR.** (Johnson et al., 2019) [MISSING] An X-ray dataset sourced from the Beth Israel Deaconess Medical Center between 2011 – 2016. The dataset consists of 371,920 chest X-rays associated with 227,943 imaging studies from 65,079 patients. A *Validation Set 3* was constructed, see next sections for more details.  
**Demographics.** Both datasets include demographic data about patients, such as their gender, age, and self-reported race. Statistical aggregates, comparisons, and other statistics can be found in *Appendix A* [MISSING]. And Table [MISSING]  
**Labeled.** For both datasets a rule-based labeler (Irvin et al., 2019) [MISSING] was used to extract observations from a free text radiology report to create structured labels for the images. The labeler was design to automatically detect the presence of 14 observations in radiology reports, capturing uncertainties inherent in radiograph interpretation. Each of the defined 14 observations were categorized by the labeler into 4 classes: confidently present (1), confidently absent (0), uncertainly present (u), or not mentioned (blank). The labeler was evaluated on 1000 distinct randomly sampled patient studies that were annotated by two board-certified radiologists. Disagreements were resolved by consensus discussion. For details regarding the labels, see Appendix A.  
**Validation Sets.** The verified *Validation Set 1* is relatively small and unbalanced between protected groups. For example, it contains only 4 Hispanic patients and [MISSING]. Therefore, we constructed *Validation Set 2* from the CheXpert training dataset, such that there isn’t overlap between patients in *Validation Set 2* and patients considered during training. *Validation Set 2* is balanced across protected groups, specifically, it contains 40 studies per group of race, gender, and age, which sums up to 960 studies overall. Using the MIMIC-CXR dataset, *Validation Set 3* was constructed, which contains 400 studies per gender, race, and age group, which totals 12,000 studies*.* Since *Validation Set 2* and *Validation Set 3* were labeled using an automatic labeler*,* only studies with no uncertainty observation were chosen. For more details and comparison between the three validation sets see *Appendix 1*.

# Experiments

**CheXpert Challenge.** A competition for automated chest x-ray interpretation, targeting 5 observations: Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion. Our experiments will focus on these observations as well.

**Uncertainty Approach.** The training labels in the dataset for each observation are either 0 (negative), 1 (positive), or u (uncertain). Irvin & Rajpurkar et al., 2019 [MISSING] shows many approaches for handling these anomalies, such as U-Ignore, U-MultiClass, U-Zeros, and U-Ones, however, there wasn’t a single approach that outperformed the rest. Out strategy was to combine U-Zeros and U-Ones, such that each ‘u’ observation was replaced by the better performing approach, according to the table presented in Irvin & Rajpurkar et al., 2019. Particularly, ‘u’ observation in Edema, Pleural Effusion, and Atelectasis pathologists was replaced by 1 (U-Ones), and in Cardiomegaly and Consolidation it was replaced by 0 (U-zeros).

[MISSING] referring CXP and MXR, left and right, unless specify otherwise validation set 1 2 3.

## X-ray Pathologies Detection

Pretrained weights [MISSING]

Recent advancements not produce improvement [MISSING]

**Reproducing Baseline.** Training a DenseNet121 model using the CheXpert dataset, as in Irvin & Rajpurkar et al., 2019. Reshaping the X-ray images to size 320 320 pixels, which then fed into the network in a fixed 16-sized batches. An Adam optimizer with default parameters was used and decreasing learning rate with initial value of . Training was executed for 10 epochs, see Table 1 [MISSING] for results on *Validation Set 1*, *2,* and *3*.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Validation Set | Atelectasis | Cardiomegaly | Consolidation | Edema | Pleural Effusion | Mean |
| U-Zeros (Irvin & Rajpurkar et al., 2019) | 1 (CXP) | 0.81 | 0.84 | 0.93 | 0.93 | 0.93 |  |
| U-Ones (Irvin & Rajpurkar et al., 2019) | 1 (CXP) | 0.86 | 0.83 | 0.90 | 0.94 | 0.93 |  |
| Ours | 1 (CXP) | 0.84 | 0.83 | 0.93 | 0.94 | 0.94 | 0.90 |
| Ours | 2 (CXP) | 0.76 | 0.87 | 0.82 | 0.92 | 0.93 | 0.86 |
| Ours | 3 (MXR) | 0.83 | 0.81 | 0.83 | 0.92 | 0.93 | 0.87 |

Table 1: Comparison of AUC scores. The results on *Validation Set 1* is very similar, as expected. There was a decrease in performance on *Validation Set 2*, it might be because of differences in the predictive ability on some groups. For example, we show in next sections that elderly patients are harder to diagnose correctly. A surprisingly high performance was achieved on *Validation Set 3*, even better than on *Validation Set 2*, which was constructed directly from CheXpert. This may be explained by the differences in the clinical reports themselves, see section [MISSING] for more information. Furthermore, we find that *Atelectasis* is harder to detect in *Validation Set 2* than in the other two validation sets. In the next section, we show that this holds true across various patient demographics.

**Performance per protected group.** The difference in performance between protected groups is a common criterion for measuring bias.We analyzed three protected attributes – Race, Gender, and Age. See Table 2, 3, and 4 [MISSING] for results on *Validation Set 2* and *Validation Set 3*. As we didn't find additional insights from the cross-comparative analysis of all three attributes (such as Asian, Female, aged 20-40), we left it out of the analysis.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Atelectasis | Cardiomegaly | Consolidation | Edema | Pleural Effusion | Mean |
| Asian | (0.75, 0.84 | (0.95, | (0.93, | (0.89, | (0.93, | 0.89, 0.85 |
| Black | (0.74, 0.82 | (0.88, | (0.81, | (0.90, | (0.92, | 0.85, 0.87 |
| Hispanic | (0.74, 0.82 | (0.86, | (0.83, | (0.85, | (0.94, | 0.84, 0.86 |
| White | (0.76, 0.83 | (0.81, 0.83 | (0.81, 0.85 | (0.89, 0.91 | (0.90, 0.94 | 0.83, 0.87 |

Table 2: Each cell presents the AUC score on Validation Set 2 (CXP, left) and Validation Set 3 (MXR, right). There isn’t a clear bias across any racial group, even though the training data composed mostly from white patients.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Atelectasis | Cardiomegaly | Consolidation | Edema | Pleural Effusion | Mean |
| Female | (0.73, | 0.95 |  |  |  | 0.85, 0.87 |
| Male | 0.78 |  |  |  |  | 0.87, 0.86 |

Table 3: In both dataset (CXP, MXR) there aren’t significant differences in terms of AUC score across genders. In *Cardiomegaly* and *Consolidation* of Males, there is a difference between datasets, although it might be due to cohort bias.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Atelectasis | Cardiomegaly | Consolidation | Edema | Pleural Effusion | Mean |
| 20-40 | (0.79, | 0.95 |  |  |  | 0.89, 0.88 |
| 40-70 | (0.78, | (0.85, |  |  |  | 0.86, 0.87 |
| 70-90 | (0.70, | (0.86, |  |  |  | 0.81, 0.81 |

Table 4: We observe a similar pattern in both datasets (CXP, MXR), a decrease in performance in elderly patients in terms of AUC score. This makes sense as clinical picture of elderly patients often complex and includes a long history, making them more difficult to diagnose.

## Race Prediction

Gichoya et al. previously showed the ability of neural nets to detect race from X-ray images. [REPHRASE] Specifically, Gichoya et al. focused on disparate abilities of Blacks and Whites, in our experiments, we expand this to Asians and Hispanics as well.   
**Training.** A pretrained DenseNet121 model using the CheXpert dataset. The training was executed for 10 epochs, with decreasing learning rate, Adam optimizer, and 16 size batches. Results are in Table 5.

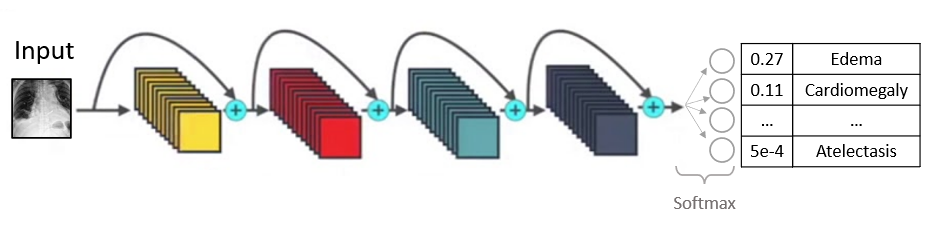
There’s a small performance gap between the 2 datasets, but overall trends look similar. We observed that in both datasets, Hispanics are significantly harder to detect than the rest of the races. This observation holds true across ages and genders as well, see Appendix [MISSING].

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Asian | Black | Hispanic | White | Mean |
| CXP | 0.95 | 0.95 | 0.77 | 0.90 | 0.90 |
| MXR | 0.90 | 0.87 | 0.66 | 0.87 | 0.83 |

Table 5: AUC scores for race detection, per race group. Performance on MXR validation declined compared to the CXP validation. The model struggles with detecting Hispanics.

## Race propagation [MISSING]

In previous studies, it was shown that X-ray images can be used to predict protected attributes, such as race, however, it is unclear whether the deep learning model takes race into account when making decisions. In an attempt to clarify this, we conducted a series of experiments, wherein each experiment we use our model from Table 1[MISSING], and apply a transfer learning technique on different parts of the network to learn race instead of chest X-ray pathologies. Particularly, we use our trained DenseNet121 model and train it to predict race, such that each time a larger part of the network is *frozen* (gradients are disabled). In essence, we are using the first part of the original network as an encoder, and then try to predict race based on this representation. See Figure 1 [MISSING] for an illustration and Figure 2 [MISSING] for results.   
According to the results, it seems that information regarding race does propagate through the network, as we were able to achieve relatively high AUC in most cases. However, when we use solely the 1024-sized feature vector produced by the four denseblocks (i.e. four denseblocks are frozen), the race predictive ability decreased significantly. Therefore, although race information is learned, its influence on the final decision might not be very significant. Furthermore, due to the concatenation of all feature maps in DenseNet, the information learned in early blocks is still passed through the network (with some information loss caused by pooling layers applied in transition between blocks). Consequently, the steep drop in performance may be due to a simple FC layer being used to predict race, whereas a deeper CNN-based model likely to still be able to predict race accurately.

  
Figure 1: DenseNet121 architecture. It consists of 4 denseblocks, whereas each block composed of multiple feature maps. In oppose to ResNets, DenseNets do not sum the output feature maps, but concatenate them. Each concatenation follows by a pooling layer to lower on computations. If a denseblock is *frozen*, the gradients won’t propagate through its layers, leaving it unchanged. Following the four denseblocks, a fully-connected layer named ‘classifier’ is applied.

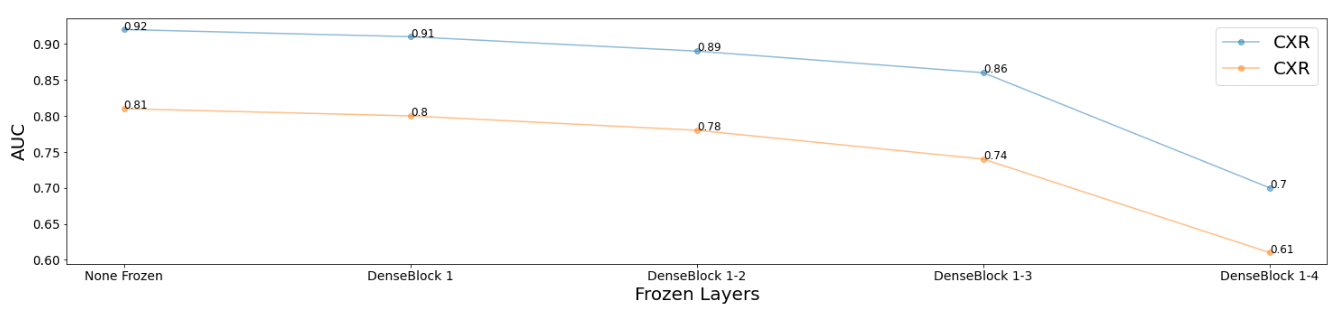


Figure 2: AUC score for predicting race, such that an increasing part of the network is being frozen. Both datasets show a similar pattern - performance slightly declines when more denseblocks are frozen, except when all four denseblocks are frozen, which results in a significant decline in AUC.

## Labeler Analysis

**Labeler Performance.** As part of our experiments, we used data that has been automatically labeled, both for training and for validation. The two datasets were labeled using the same labeler, which was assessed on 1000 randomly selected reports from CheXpert and 687 randomly selected reports from MIMIC-CXR. Each report was verified by a board of certificate radiologists. Generally, the labeler performed similar across datasets, with some differences favoring the CheXpert dataset. For example, detection of 'Edema' in radiology reports from CXP dataset achieved an F1-score of 0.993 (Table 2, Ivrin [MISSING]), versus 0.888 for radiology reports from MXR dataset (Table 5, Johnson [MISSING]).

**Uncertainty.** We avoided using observations with one or more uncertainties in our validations. This section analyses the differences in uncertainty between datasets, based on quantity and distribution across patient demographics. See Table 8 for results.  
We found a significant difference across datasets in terms of number of uncertain observations. The CXP dataset contained an average of 0.421 uncertainty tags per study, where an average of 0.148 uncertain observations was measure in MXR studies. This could be due to the nature of how radiology reports produced in each dataset, different hospitals might have different policies to what include and exclude from radiology reports, which could be reflected here.

Performance of MXR [MISSING]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Atelectasis | Cardiomegaly | Consolidation | Edema | Pleural Effusion | Mean per Entire Study |
| CXP | 0.151 (36%) | 0.036 (9%) | 0.124 (29%) | 0.058 (14%) | 0.052 (12%) | 0.421 |
| MXR | 0.038 (25%) | 0.021 (15%) | 0.014 (10%) | 0.052 (35%) | 0.022 (15%) | 0.148 |

Table 8: Mean number of uncertainties per study, and across label categories. The mean number of uncertainties is significantly bigger in the CXP dataset.

# Summary

In our work, we analyzed potential pitfalls of deploying deep neural networks. The first part of our work involved around finding model bias, specifically in terms of performance gaps between protected groups. We haven’t detected any unreasonable bias across any demographic group. In the second part of our work, we focused on the ability of deep models to detect race from X-ray medical images. Although, the ability itself is not necessarily an issue of concern, our finding that racial information propagates through the entire network, a network which was trained to find pathologies, a task that should have little to no correlation to race, creates a risk for deep models’ deployment in real life settings.   
Furthermore, through our experiments we validate our results on an out of distribution validation set, to reassure our conclusions.  
There were several limitations to this work. First, we relied on self-reported race as the ground truth when making predictions. Race identity often conflated with biological constructs, in our study we define race as a social construct that pertains to how we interact with each other and how others perceive us. Secondly, as self-reported race considered a strong proxy to racial identity (genetic ancestry), it could be a potential confounder for the detection of some diseases. Future work could deepen the analysis and stratified the validation sets according to the presence of diseases. This line of research allows to better understand the correct way of deploying a model in medical imaging and reassure the model’s behavior in new environments.

is very important to further understand the correct way to deploy such medical tools and reassure the behaviour of such classifier to generalize to new environments.

deployments in medical imaging

OOD [MISSING]

Self-reported attribute is

Race identity often conflated with biological constructs, in our study we define race as a social construct that pertains to how we interact with each other and how others perceive us. To this end, we use self-reported race as the racial identity of patients throughout the study.

The connection between self-reported race and genetic ancestry [MISSING], that concludes that race is more of social construct than a biological construct, which makes self-reported race a strong proxy for racial identity. Furthermore,

How others see me

Furthermore, an additional analyses could be done on the distribution of diseases across races, as in our experiments we didn’t stratified the validation set according to disease finding.

There were several limitations to this work. Most importantly, we relied on self-reported race as the ground Articles e413 www.thelancet.com/digital-health Vol 4 June 2022 truth for our predictions. There has been extensive research into the association between self-reported race and genetic ancestry, which has shown that there is more genetic variation within races than between races, and that race is more a social construct than a biological construct.24 We note that in the context of racial discrimination and bias, the vector of harm is not genetic ancestry but the social and cultural construct that of racial identity, which we have defined as the combination of external perceptions and self-identification of race. Indeed, biased decisions are not informed by genetic ancestry information, which is not directly available to medical decision makers in almost any plausible scenario. As such, self-reported race should be considered a strong proxy for racial identity

regardless of race,

(1) to assess the ability of deep learning AI models to recognise race from medical images, including the ability of these models to generalise to new environments and across multiple imaging modalities; (2) to examine possible confounding anatomic and phenotype population features as explanations for these performance scores, and (3) to investigate the underlying mechanisms by which AI models can recognise race.

The results from our study emphasise that the ability of AI deep learning models to predict self-reported race is itself not the issue of importance. However, our finding that AI can accurately predict self-reported race, even from corrupted, cropped, and noised medical images, often when clinical experts cannot, creates an enormous risk for all model deployments in medical imaging [MISSING]

In our work, we focused on analyzing potential pitfalls when using a blackbox model such as NN. We didn’t find unreasonable bias across any demographic group, in terms of pathologies detection.

The second part of our work evolved around NN ability to detect race directly from X-ray images. Although it’s not necessary issue of importance, it’s

We showed that Hispanics are harder to detect, and that features regarding race propagates to almost every part of the network.

to generalise to new environments

Discussion

Results

Motivation

(1) to assess the ability of deep learning AI models to recognise race from medical images, including the ability of these models to generalise to new environments and across multiple imaging modalities; (2) to examine possible confounding anatomic and phenotype population features as explanations for these performance scores, and (3) to investigate the underlying mechanisms by which AI models can recognise race.

The results from our study emphasise that the ability of AI deep learning models to predict self-reported race is itself not the issue of importance. However, our finding that AI can accurately predict self-reported race, even from corrupted, cropped, and noised medical images, often when clinical experts cannot, creates an enormous risk for all model deployments in medical imaging [MISSING]

Limitation to the study - In our experiments we’re using data labeled by an automatic labeler, both for training and validation. Unreached clinical notes.

Disease distribution as future

Future work – understanding how race being predicted, using feature maps for example.

# Appendix

Race prediction

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Asian | Black | Hispanic | White | Mean |
| Female |  |  | 0.80, 0.66 |  | 0.91, 0.82 |
| Male |  |  | 0.78, 0.66 |  | 0.90, 0,83 |

Table 6: little to no difference between ages.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Asian | Black | Hispanic | White | Mean |
| 20-40 |  |  | 0.80, 0.66 |  | 0.91, 0.81 |
| 40-70 |  |  | 0.78, 0.66 |  | 0.90, 0,85 |
| 70-90 |  |  |  |  | 0.92, 0.83 |

Table 7: small changes between ages.

The trend we observed in the previous task, that performance on older patients is lower, does not exists in this task.

[MISSING]

[OPTIONAL] label analysis: The difference between dataset holds true across other demographics (race, gender, and age). Across races and genders distribution were the same, ages 40% than the rest of the population.

# References

[1]

16. M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik,K. Kenthapadi and A. T. Kalai, Bias in bios: a case study of semantic representation bias in a Pacific Symposium on Biocomputing242high-stakes setting (2019), Atlanta, GA.

17. A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism predictioninstruments,Big data5, 153 (2016).

18. J. Buolamwini and T. Gebru, Gender Shades: Intersectional Accuracy Disparities in CommercialGender Classification,81, p. 15 (2018).

19. I. Chen, F. D. Johansson and D. Sontag, Why Is My Classifier Discriminatory?, inNIPS’31,2018 pp. 3539–3550.20. J. Kleinberg, S. Mullainathan and M. Raghavan, Inherent Trade-Offs in the Fair Determinationof Risk Scores,arXiv:1609.05807(2016).21. M. Srivastava, H. Heidari and A. Krause, Mathematical notions vs. human perception of fairness:a descriptive approach to fairness for machine learning,arXiv preprint arXiv:1902.04783(2019).22. D. S. Char, N. H. Shah and D. Magnus, Implementing machine learning in health care —addressing ethical challenges,New England Journal of Medicine378, 981 (2018).23. Z. Obermeyer and S. Mullainathan, Dissecting racial bias in an algorithm that guides healthdecisions for 70 million peoples, p. 89 (2019), Atlanta, GA.24. M. A. Gianfrancesco, S. Tamang, J. Yazdany and G. Schmajuk, Potential biases in machinelearning algorithms using electronic health record data.,JAMA internal medicine178, 1544(2018).25. S. Akbarian, L. Seyyed-Kalantari, F. Khalvati and E. Dolatabadi, Evaluating knowledge transferin neural network for medical images,arXiv preprint arXiv:2008.13574(2020).26. G. Huang, Z. Liu, L. v. d. Maaten and K. Q. Weinberger, Densely Connected ConvolutionalNetworks, 2261 (2017).27. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale HierarchicalImage Database (2009).28. D. P. Kingma and J. Ba, Adam: A method for stochastic optimization,arXiv:1412.6980v9(2017).29. A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone and E. Ferrante, Gender imbalance inmedical imaging datasets produces biased classifiers for computer-aided diagnosis,Proceedingsof the National Academy of Sciences117, 12592 (2020).30. T. B. Hashimoto, M. Srivastava, H. Namkoong and P. Liang, Fairness Without Demographicsin Repeated Loss Minimization,arXiv:1806.08010(2018).31. R. G. J. Miller,Simultaneous Statistical Inference(Springer-Verlag New York, 1981).32. J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano and E. K. Oermann, Confoundingvariables can degrade generalization performance of radiological deep learning models,PLOSMedicine15, p. e1002683 (2018).33. N. Tomaˇsev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer,S. Ravuri, I. Protsyuket al., A clinically applicable approach to continuous prediction of futureacute kidney injury,Nature572, p. 116 (2019

Yi PH, Wei J, Kim TK, et al. Radiology “forensics”: determination of age and sex from chest radiographs using deep learning. Emerg Radiol 2021; 28: 949–54.

12

Eng DK, Khandwala NB, Long J, et al. Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: a prospective multicenter randomized controlled trial. Radiology 2021; 301: 692–99. 13 Rim TH, Lee G, Kim Y, et al. Prediction of systemic biomarkers from retinal photographs: development and validation of deeplearning algorithms. Lancet Digit Health 2020; 2: e526–36.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al., 2019. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. arXiv:1901.07031.