The racial effect in radiographs interpretation

Sharon Peled

Abstract

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 diseases using chest X-rays. However, neural networks often learn to make predictions that overly rely on spurious correlations 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 X-ray images. Our work aims to understand the impact race has on X-ray medical imaging in deep learning models. To this extent, we're utilizing two popular large-scale chest X-ray datasets: CheXpert and MIMIC-CXR.
Git: https://github.com/SharonPeled/racial-effect-in-radiographs-interpretation

1 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 [3] and gender bias in online recruiting [5]. Studies in the medical domain have shown similar results in many healthcare applications, such as mortality prediction [14], and melanoma detection [4], [12]. Sources of bias may originate in different points along the classical machine learning pipeline. For example, considering minority groups, collected features may not be as indicative compared to the rest of the population [2]. Bias may also arise because of differences in how features interact with each other and with the target within subpopulations in the dataset, resulting in an underfit model for these subpopulations [2]. 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 (OOD) data [1]. In our work, we analyze how race affects deep learning models' interpretation of X-ray medical images, and validate our conclusions using OOD dataset.

Fairness. Fairness has been conceptualized mathematically and philosophically in a variety of ways, such as error rate balance [9], worst-case group accuracy [13], and fairness through unawareness [6]. 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. In out context, using X-ray medical images, we measure fairness through performance gaps among protected groups in terms of AUC scores for detecting diseases. Furthermore, in the second part of our work, we test for fairness through unawareness by examining whether the model encodes racial information during training.

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, utilizing such abilities to improve a radiologist performance in skeletal age assessment [7], and determine the age and sex of patients from chest radiographs [15]. Furthermore, it was demonstrated how race could be accurately identified from X-ray images alone [8], an ability that is unexplainable by physicians. Moreover, it was shown that even with severe data augmentation the race identification abilities almost

didn't decrease [8]. This ability of deep models to identify race is not necessary an issue of concern. In our experiments, we test weather the classifier uses this ability when interpreting X-rays images.
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.

2 The Data

CheXpert. A chest radiograph dataset from Stanford Hospital that contains 224,316 frontal and lateral chest radiographs of 65,240 patients [10]. The dataset includes a validation set which contains an addition 200 studies that has been verified by a board of three certificate radiologists. In our experiments, we denote it 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 next sections.
MIMIC-CXR. A chest radiograph dataset sourced from the Beth Israel Deaconess Medical Center between 2011 – 2016 [11]. The dataset consists of 371,920 chest X-rays associated with 227,943 imaging studies from 65,079 patients. From this dataset we constructed Validation Set 3, see next sections for more details.
Demographics. Both datasets include demographic data about most patients, such as their gender, age, and self-reported race. Statistical aggregates, comparisons, and other statistics can be found in Table 1 and Appendix A.
Automatic Labeler. For both datasets a rule-based labeler was used to extract observations from a free text radiology report to create structured labels for the images [10]. 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 was 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.
Validation Sets. The verified Validation Set 1 is relatively small and unbalanced between protected groups. For example, it contains only 4 studies of Hispanic patients and 8 studies of Black patients, see Appendix A. Therefore, we constructed Validation Set 2 from the CheXpert training dataset, such that there isn't overlap between patients in Validation Set 1, Validation Set 2, and patients considered during training. Validation Set 2 has been constructed in such a way that it is balanced across the 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, we constructed Validation Set 3. Similarly to Validation Set 2, Validation Set 3 is also balanced across protected groups, it contains 400 studies per race, gender, and age group, which totals 9,600 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 Table 1 and Appendix A.

3 Experiments

CheXpert Challenge. A competition for automated chest X-ray interpretation, targeting 5 observations: Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion. In our experiments we focus on these observations as well.
Uncertainty Approach. The training labels are either 0 (negative), 1 (positive), blank (not mentioned), or u (uncertain). Many approaches was shown in [10] for handling these uncertainties, such as U-Ignore, U-Zeros, and U-Ones. However, there wasn't a single approach that outperformed the rest in all categories. Our 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 [10]. 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).

Pathology	Validation Set 1		Validation Set 2		Validation Set 3	
	Positive	Negative	Positive	Negative	Positive	Negative
Atelectasis	75 (37%)	125 (62%)	133 (13%)	827 (86%)	927 (9%)	8673 (90%)
Cardiomegaly	66 (33%)	134 (67%)	94 (9%)	866 (90%)	1025 (10%)	8575 (89%)
Consolidation	32 (16%)	168 (84%)	53 (5%)	907 (94%)	171 (1%)	9429 (98%)
Edema	42 (21%)	158 (79%)	138 (14%)	822 (85%)	359 (3%)	9241 (96%)
Pleural Effusion	64 (32%)	136 (68%)	236 (24%)	724 (75%)	773 (8%)	8827 (91%)

Table 1: The number (%) of studies which contain each of the targeted observations, per validation set. Note that there are no uncertainties in any of the validation sets.

3.1 Chest X-ray Classification

Reproducing Baseline. Training a DenseNet121 model with pretrained weights using the CheXpert dataset, as in [10]. 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 β parameters was used and decreasing learning rate with initial value of 1×10^{-4}. Training was executed for 5 epochs, see Table 2 for results on Validation Set 1, 2, and 3.

	Validation Set	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion	Mean
U-Zeros [10]	1 (CXP)	0.81	0.84	0.93	0.93	0.93	0.89
U-Ones [10]	1 (CXP)	0.86	0.83	0.90	0.94	0.93	0.89
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 2: 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, which 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 3.4 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.

In next sections, we refer to Validation Set 2 as CXP (since it was produced from CheXpert dataset), and Validation Set 3 as MXR (MIMIC-CXR). The following experiments focused on performance per protected group, therefore, we avoided using Validation Set 1 due to its size.

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 3, 4, and 5 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, 0.80	0.91, 0.79	0.89, 0.92	0.93, 0.91	0.89, 0.85
Black	0.74, 0.82	0.88, 0.80	0.81, 0.89	0.90, 0.92	0.92, 0.92	0.85, 0.87
Hispanic	0.74, 0.82	0.86, 0.83	0.83, 0.81	0.85, 0.93	0.94, 0.93	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 3: 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, 0.82	0.82, 0.82	0.86, 0.84	0.89, 0.93	0.93, 0.94	0.85, 0.87
Male	0.78, 0.83	0.94, 0.80	0.77, 0.83	0.90, 0.91	0.94, 0.92	0.87, 0.86

Table 4: In both validation sets (CXP, MXR) there aren't significant differences in terms of AUC scores across genders. In Cardiomegaly and Consolidation, there is some performance gap between genders in the CXP dataset, while in the MXR dataset there's no such gap. Furthermore, when considering these observations in males, there is also a difference between datasets. These differences might be due to cohort bias, as Validation Set 2 is relatively small and Validation Set 3 is x10 bigger.

	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion	Mean
20-40	0.79, 0.86	0.87, 0.80	0.93, 0.84	0.92, 0.95	0.93, 0.94	0.89, 0.88
40-70	0.78, 0.83	0.85, 0.79	0.81, 0.87	0.90, 0.93	0.94, 0.94	0.86, 0.87
70-90	0.69, 0.75	0.86, 0.77	0.72, 0.77	0.86, 0.88	0.91, 0.89	0.81, 0.81

Table 5: 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.

3.2 Race Prediction

Previous studies have shown that neural networks are able to detect race from X-ray images [8]. Particularly, [8] 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 5 epochs, with decreasing learning rate, Adam optimizer, and 16 size batches. Results are in Table 6.

There is a 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 A.

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

3.3 Race Encoding Through the Network

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 model considers race when making decisions. In an attempt to clarify this, we conducted a series of experiments, wherein each experiment we used our model from Table 2 and applied transfer learning on different parts of the network to learn race instead of chest X-ray pathologies. Particularly, we used our trained DenseNet121 model and trained it to predict race, such that each time a larger prefix 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 learn to predict race based on this representation. See Figure 1 for an illustration and Figure 2 for results.

According to our 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 architecture, 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, when the four denseblocks are frozen, it is likely that the steep drop in performance occurs due to the fact that only fully-connected layers are being learned (as all CNN layers are frozen), whereas a deep CNN-based model might still be able to predict race accurately.

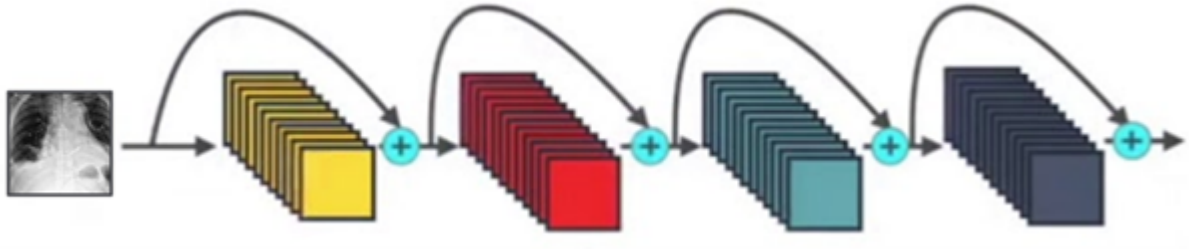

Figure 1: DenseNet121 architecture. It consists of 4 denseblocks, whereas each block composed of multiple feature maps. In oppose to ResNets, DenseNets do not sum the output feature maps, but concatenate them. Each concatenation follows by a pooling layer to lower on computations. If a denseblock is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

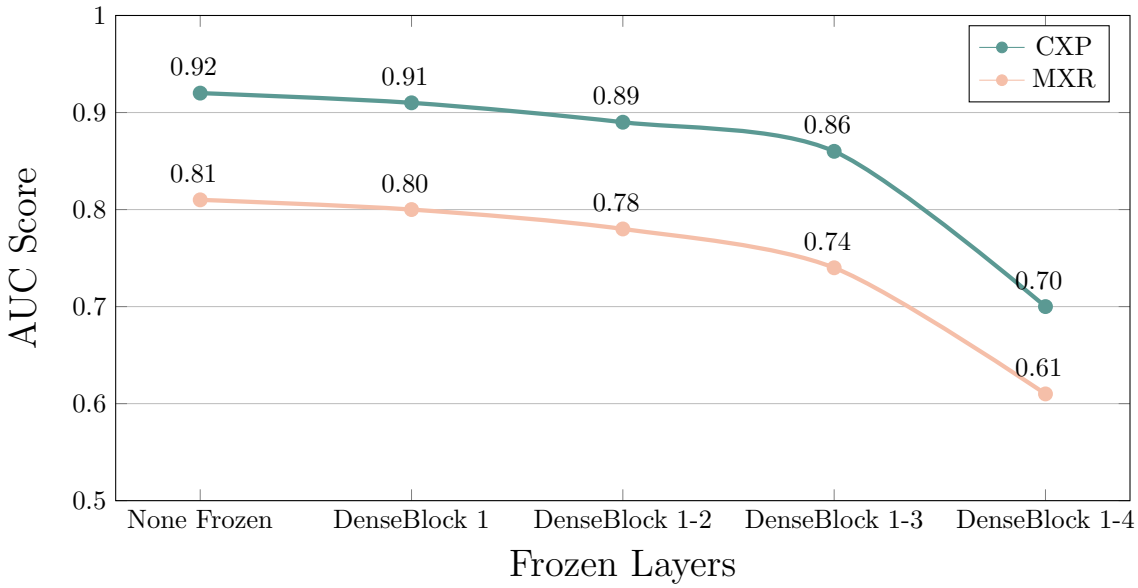

Figure 2: AUC score for predicting race, such that an increasing prefix of the network is being frozen. Both datasets show a similar pattern - performance slightly declines when more denseblocks are frozen, except when all four denseblocks are frozen, which results in a significant decline in AUC.

3.4 Automatic 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 [10], versus 0.888 for radiology reports from MXR dataset [11].

Uncertainty. We avoided using observations with one or more uncertainties in our validations. In this section, we analyze the differences in uncertainty between the two datasets. See Table 7 for results. We found a significant difference across datasets in terms of number of uncertain observations per study. 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 are produced in each dataset, different hospitals might have different policies to what include or exclude from the radiology report, which could be reflected here.

	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion	Mean per Study
CXP	0.151	0.036	0.124	0.058	0.052	0.421
MXR	0.038	0.021	0.014	0.052	0.022	0.148

Table 7: Mean number of uncertainties per study in each dataset. The mean number of uncertainties is significantly bigger in the CXP dataset.

4 Summary

In our work, we analyzed potential pitfalls of deploying deep neural networks. In the first part of our work we tested for model bias, particularly 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 X-ray 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. In our context, racial bias is not a genetic trait but rather formed by social and cultural factors. 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. In addition, since our experiments use data that had been automatically labelled, we are vulnerable to errors in our validation sets.

We believe that this line of research allows to better understand the correct way of deploying deep learning models in medical imaging, and reassuring the model's behavior in new environments.

References

[1] Ehab Albadawy, Ashirbani Saha, and Maciej Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical Physics*, 45, 01 2018.

[2] Irene Y. Chen, Fredrik D. Johansson, and David Sontag. Why is my classifier discriminatory? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3543–3554, Red Hook, NY, USA, 2018. Curran Associates Inc.

[3] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. 2016.

[4] Roxana Daneshjou, Kailas Vodrahalli, Weixin Liang, Roberto A Novoa, Melissa Jenkins, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, James Zou, and Albert Chiou. Disparities in dermatology ai: Assessments using diverse clinical images, 2021.

[5] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, jan 2019.

[6] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

[7] David Eng, Nishith Khandwala, Jin Long, Nancy Fefferman, Shailee Lala, Naomi Strubel, Sarah Milla, Ross Filice, Susan Sharp, Alexander Towbin, Michael Francavilla, Summer Kaplan, Kirsten Ecklund, Sanjay Prabhu, Brian Dillon, Brian Everist, Christopher Anton, Mark Bittman, Rebecca Dennis, and Safwan Halabi. Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: A prospective multicenter randomized controlled trial. *Radiology*, 301:204021, 09 2021.

[8] Judy Gichoya, Imon Banerjee, Ananth Bhimireddy, John Burns, Leo Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew Lungren, Lyle Palmer, Brandon Price, Saptarshi Purkayastha, Ayis Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, and Haoran Zhang. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4, 05 2022.

[9] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[10] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David Mong, Safwan Halabi, Jesse Sandberg, Ricky Jones, David Larson, Curtis Langlotz, Bhavik Patel, Matthew Lungren, and Andrew Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:590–597, 07 2019.

[11] Alistair Johnson, Tom Pollard, Seth Berkowitz, Nathaniel Greenbaum, Matthew Lungren, Chih-ying Deng, Roger Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs, 01 2019.

[12] Cristian Navarrete-Dechent, Stephen Dusza, Konstantinos Liopyris, Ashfaq Marghoob, Allan Halpern, and Michael Marchetti. Automated dermatological diagnosis: Hype or reality? *Journal of Investigative Dermatology*, 138, 06 2018.

[13] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[14] Rahuldeb Sarkar, Christopher Martin, Heather Mattie, Judy Wawira Gichoya, David J Stone, and Leo Anthony Celi. Performance of intensive care unit severity scoring systems across different ethnicities in the usa: a retrospective observational study. *The Lancet Digital Health*, 3(4):e241–e249, 2021.

[15] Paul Yi, Jinchi Wei, Tae Kyung Kim, Jiwon Shin, Haris Sair, Ferdinand Hui, Gregory Hager, and Cheng Lin. Radiology "forensics": determination of age and sex from chest radiographs using deep learning. *Emergency Radiology*, 28, 06 2021.

5 Appendix A

5.1 Cohort Statistics

	CheXpert	MIMIC-CXR
# Images	222,792	376,206
# Patients	64,427	65,152
# Frontal	190,498	242,754
# Lateral	32,294	133,452
Male	59.4%	52.2%
Female	40.7%	47.8%
White	55.8%	64.5%
Black	4.8%	15.3%
Asian	10.8%	4.7%
Hispanic	2.2%	5.7%
Other	26.5%	9.8%
20-40	15.2%	39.7%
40-70	50.0%	41.2%
70-90	34.7%	19.1%

Table 8: Summary statistics for MIMIC-CXR and CheXpert. Percentages shown correspond to the fraction of the population belonging to a particular group.

Race	Gender	Age	Validation Set 1	Validation Set 2	Validation Set 3	CXP Training
White	Female	20-40	9 (4.5%)	40 (4.2%)	400 (4.2%)	4345 (2.3%)
		40-70	21 (10.5%)	40 (4.2%)	400 (4.2%)	20290 (10.9%)
		70-90	15 (7.5%)	40 (4.2%)	400 (4.2%)	17772 (9.6%)
	Male	20-40	7 (3.5%)	40 (4.2%)	400 (4.2%)	6226 (3.4%)
		40-70	29 (14.5%)	40 (4.2%)	400 (4.2%)	32911 (17.7%)
		70-90	31 (15.5%)	40 (4.2%)	400 (4.2%)	22573 (12.2%)
Black	Female	20-40	- (0.0%)	40 (4.2%)	400 (4.2%)	785 (0.4%)
		40-70	3 (1.5%)	40 (4.2%)	400 (4.2%)	2785 (1.5%)
		70-90	2 (1.0%)	40 (4.2%)	400 (4.2%)	1213 (0.7%)
	Male	20-40	2 (1.0%)	40 (4.2%)	400 (4.2%)	858 (0.5%)
		40-70	1 (0.5%)	40 (4.2%)	400 (4.2%)	3219 (1.7%)
		70-90	- (0.0%)	40 (4.2%)	400 (4.2%)	744 (0.4%)
Asian	Female	20-40	3 (1.5%)	40 (4.2%)	400 (4.2%)	1105 (0.6%)
		40-70	5 (2.5%)	40 (4.2%)	400 (4.2%)	4123 (2.2%)
		70-90	4 (2.0%)	40 (4.2%)	400 (4.2%)	3072 (1.7%)
	Male	20-40	1 (0.5%)	40 (4.2%)	400 (4.2%)	1309 (0.7%)
		40-70	7 (3.5%)	40 (4.2%)	400 (4.2%)	5673 (3.1%)
		70-90	4 (2.0%)	40 (4.2%)	400 (4.2%)	3862 (2.1%)
Hispanic	Female	20-40	1 (0.5%)	40 (4.2%)	400 (4.2%)	225 (0.1%)
		40-70	- (0.0%)	40 (4.2%)	400 (4.2%)	781 (0.4%)
		70-90	- (0.0%)	40 (4.2%)	400 (4.2%)	437 (0.2%)
	Male	20-40	1 (0.5%)	40 (4.2%)	400 (4.2%)	805 (0.4%)
		40-70	2 (1.0%)	40 (4.2%)	400 (4.2%)	1187 (0.6%)
		70-90	- (0.0%)	40 (4.2%)	400 (4.2%)	388 (0.2%)
Other	Female	20-40	3 (1.5%)	- (0.0%)	- (0.0%)	3884 (2.1%)
		40-70	18 (9.0%)	- (0.0%)	- (0.0%)	10697 (5.8%)
		70-90	10 (5.0%)	- (0.0%)	- (0.0%)	5805 (3.1%)
	Male	20-40	2 (1.0%)	- (0.0%)	- (0.0%)	5498 (3.0%)
		40-70	14 (7.0%)	- (0.0%)	- (0.0%)	16447 (8.9%)
		70-90	5 (2.5%)	- (0.0%)	- (0.0%)	6708 (3.6%)

Table 9: Number of studies (%) per race, gender, and age in training and in each of the validation sets.

5.2 Race Prediction Performance per Protected Group

	Asian	Black	Hispanic	White	Mean
Female	0.96, 0.90	0.96, 0.86	0.80, 0.66	0.91, 0.86	0.91, 0.82
Male	0.95, 0.91	0.96, 0.88	0.78, 0.66	0.92, 0.87	0.90, 0.83

Table 10: AUC scores (CXP, MXR) for predicting race. Little to no difference between genders.

	Asian	Black	Hispanic	White	Mean
20-40	0.93, 0.87	0.98, 0.88	0.81, 0.65	0.91, 0.84	0.91, 0.81
40-70	0.97, 0.93	0.95, 0.89	0.78, 0.67	0.91, 0.89	0.90, 0.85
70-90	0.96, 0.92	0.97, 0.87	0.81, 0.66	0.95, 0.87	0.92, 0.83

Table 11: AUC scores (CXP, MXR) for predicting race. Contrary to the X-ray interpretation task, where we observed a lack of performance among elderly patients, we find balanced performance across all ages.