Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Risk of Bias in Chest Radiography Deep Learning Foundation Models

Ben Glocker, PhD • *Charles Jones, MEng* • *Mélanie Roschewitz, MSc* • *Stefan Winzeck, PhD*

From the Department of Computing, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom. Received February 28, 2023; revision requested March 29; revision received August 7; accepted August 24. **Address correspondence to** B.G. (email: *b.glocker@imperial.ac.uk*).

B.G. received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 757173, Project MIRA). S.W. is supported by the UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare. C.J. is supported by Microsoft Research and Engineering and Physical Sciences Research Council. M.R. is funded through Imperial College London President's PhD Scholarship.

Conflicts of interest are listed at the end of this article.

See also commentary by Czum and Parr in this issue.

Radiology: Artificial Intelligence 2023; 5(6):e230060 • https://doi.org/10.1148/ryai.230060 • Content codes: AI CH

Purpose: To analyze a recently published chest radiography foundation model for the presence of biases that could lead to subgroup performance disparities across biologic sex and race.

Materials and Methods: This Health Insurance Portability and Accountability Act–compliant retrospective study used 127 118 chest radiographs from 42 884 patients (mean age, 63 years ± 17 [SD]; 23 623 male, 19 261 female) from the CheXpert dataset that were collected between October 2002 and July 2017. To determine the presence of bias in features generated by a chest radiography foundation model and baseline deep learning model, dimensionality reduction methods together with two-sample Kolmogorov–Smirnov tests were used to detect distribution shifts across sex and race. A comprehensive disease detection performance analysis was then performed to associate any biases in the features to specific disparities in classification performance across patient subgroups.

Results: Ten of 12 pairwise comparisons across biologic sex and race showed statistically significant differences in the studied foundation model, compared with four significant tests in the baseline model. Significant differences were found between male and female ($P < .001$) and Asian and Black ($P < .001$) patients in the feature projections that primarily capture disease. Compared with average model performance across all subgroups, classification performance on the "no finding" label decreased between 6.8% and 7.8% for female patients, and performance in detecting "pleural effusion" decreased between 10.7% and 11.6% for Black patients.

Conclusion: The studied chest radiography foundation model demonstrated racial and sex-related bias, which led to disparate performance across patient subgroups; thus, this model may be unsafe for clinical applications.

Supplemental material is available for this article.

Published under a CC BY 4.0 license.

Deep learning–based predictive models have found great success in medical imaging applications, such as disease detection in chest radiography (1). However, training of these models requires access to large amounts of representative data. Generalization across different clinical sites remains a major challenge for wider clinical adoption (2). Training on limited data makes models susceptible to failure whenever the data characteristics change, often caused by differences in the patient demographic characteristic (ie, population shift) and/or imaging technique (ie, acquisition shift) (3,4).

Foundation models have emerged as a promising solution to mitigate these issues (5,6). These models are pretrained on large-scale, heterogeneous, and diverse datasets, often by self-supervised or semisupervised learning strategies that do not require ground truth annotations, with the hope to provide robust backbones for task-agnostic feature extraction. These backbone features then serve as inputs for the subsequent, data-efficient training of task-specific prediction models. The term *foundation model* is now widely used to describe pretrained, versatile deep learning models that can be adapted to a wide range of downstream prediction tasks (7).

In medical imaging, pretraining is particularly attractive because of the difficulty of collecting large amounts of high-quality training data. Recent work includes self-supervised pretraining on large unlabeled medical imaging datasets, which appears to improve performance on data not only from similar sources (in-distribution) but also across various downstream tasks on new, out-of-distribution data (8). These findings were corroborated by Ghesu et al (9), who proposed a foundation model trained on more than 1 million diverse medical images. Similarly, Sellergren et al (10) recently developed a chest radiography foundation model, demonstrating that it can improve performance in downstream tasks as well as drastically reduce the amount of labeled training data required for task-specific fine-tuning. The pretrained model yielded an area under the receiver operating characteristic curve (AUC) of 0.95 for detecting tuberculosis when using only 45 chest radiographs for task-specific training, which was noninferior to radiologist performance. Outcome prediction after COVID-19 was better with

This copy is for personal use only. To order printed copies, contact reprints@rsna.org

Abbreviations

AUC = area under the receiver operating characteristic curve, FPR = false-positive rate, MLP = multilayer perceptrons, PCA = principal component analysis, TPR = true-positive rate, *t*-SNE = *t*-distributed stochastic neighbor embedding

Summary

Biases related to biologic sex and race in a recently published chest radiography foundation model for disease detection resulted in substantial performance disparities across protected subgroups.

Key Points

- Bias analysis of a recently published chest radiography foundation model showed significant differences between features related to disease detection across biologic sex ($P < .001$) and race ($P < .001$).
- Compared with average model performance, classification performance on the "no finding" label decreased between 6.8% and 7.8% for female patients, and performance in detecting "pleural effusion" decreased between 10.7% and 11.6% for Black patients.
- The studied foundation model may be unsafe for clinical applications because its use could amplify health disparities.

Keywords

Conventional Radiography, Computer Application-Detection/Diagnosis, Chest Radiography, Bias, Foundation Models

freezing the backbone versus fine-tuning the entire model on the complete dataset, when using only 528 chest radiographs for training.

Despite their increasing popularity, little is known about potential biases encoded and reinforced in these foundation models, as well as their effect on embedding biases in downstream models. Previous studies on foundation models in medical imaging largely lack a comprehensive bias analysis. This deserves a closer investigation in light of ethical and regulatory concerns regarding use of foundation models in health care applications (11) and, specifically, in radiology (12). Use of foundation models in medical imaging may be of particular concern given the recently demonstrated ability of deep learning models to accurately recognize protected characteristics, such as racial identity, and other demographic information (13,14).

In this study, we analyzed a recently published chest radiography foundation model proposed in the work by Sellergren et al (10). We inspected the generated features of this proprietary model for the presence of biases that could potentially lead to disparate performance across patient subgroups (15,16). We conducted a comprehensive subgroup performance analysis when using the foundation model for the downstream application of disease detection. Our performance analysis associates biases found in feature representation to specific performance disparities in protected subgroups.

Materials and Methods

This retrospective study is exempt from ethical approval because the analysis is based on secondary data that are publicly available, and no permission is required to access the data. The study was compliant with the Health Insurance Portability and Accountability Act.

Study Sample

We used a sample from the publicly available CheXpert dataset (17), which is composed of data from a total of 42 884 patients with 127 118 chest radiographs. The radiographs were divided into three sets for training (76 205 radiographs), validation (12 673 radiographs), and testing (38 240 radiographs) and were collected between October 2002 and July 2017. The study sample and data splits used in our study are identical to the ones used in the recent study by Gichoya et al (13). We refer the reader to the study by Gichoya et al, and specifically to their extensive supplementary material, for an excellent discussion and further information about the definitions of the used racial groupings. The code repository *(https://github.com/biomedia-mira/cxr-foundation-bias)* released with our study contains detailed information on how to construct the study sample from the original CheXpert dataset.

Models

The primary model of our investigation is the recently proposed chest radiography foundation model (10). According to the description, this model was first pretrained on a large corpus of natural images followed by a second pretraining on more than 800 000 chest radiographs from India and the United States. This second pretraining step is based on supervised contrastive learning to specifically train for the classification of images with and without abnormality, leveraging disease labels extracted from radiology reports using natural language processing. The foundation model is intended to serve as a robust feature extractor used for subsequent training of downstream, task-specific prediction models. Access to the foundation model is made available through a programming interface, which allows only the processing of input images, with the output corresponding to the generated features. The network weights of the chest radiography foundation model itself, however, are not publicly available; thus, the parameters of the feature extractor cannot be updated during training of downstream tasks.

To compare the chest radiography foundation model with a "traditional" approach of model development, we adopted a widely used deep convolutional neural network, DenseNet-121 (18), trained and validated on the CheXpert training and validation sets. This network is pretrained on natural images and then fine-tuned for the task of disease detection in chest radiography, using the disease annotations available in the CheXpert dataset. We used the identical, publicly available model described in the work by Glocker et al (19). The already fully trained model was obtained from their code repository, and no further modifications were made. Hereafter, we refer to this baseline model as the CheXpert model. A similar model has been used in other works and is considered state of the art for chest radiography disease detection (16,20).

Model Inspection

To analyze whether biases may persist in the features generated by the foundation model, we used the CheXpert test set with 38 240 scans. We used test set resampling to correct for varia-

tions across subgroups, such as racial imbalance, differences in age, and varying prevalence of disease. We then used the feature exploration framework proposed by Glocker et al (19). First, we obtained the corresponding features for the entire test set by passing each scan through the model backbones of the chest radiography foundation model and the CheXpert model. The high-dimensional feature vectors were then projected down to lower-dimensional feature spaces using principal component analysis (PCA) and *t*-distributed stochastic neighbor embedding (*t*-SNE). For PCA, these new dimensions (also called modes) capture the direction of the largest variation in the high-dimensional feature space. This means that for a model trained for disease detection, we find the strongest separation of samples with and without the disease in the first few modes of PCA. We applied *t*-SNE on top of PCA using all modes to retain 99% of the variance, aiming to capture the overall similarity between samples in the original high-dimensional feature space. For the bias analysis, we randomly sampled a set of 3000 patients (1000 samples from each racial group) and inspected whether the PCA modes that separate samples by disease may additionally separate non–disease-related patient characteristics, such as biologic sex or racial identity. Similarly, we inspected *t*-SNE projections to determine whether any groupings or distributional differences appear across patient subgroups. Differences found across subgroups in PCA and/or *t*-SNE projections may indicate that the underlying features not only capture variation in disease status but also encode biases with respect to protected patient characteristics.

Model Performance

Although biases in the features may not necessarily be problematic, it is important to assess whether such biases may affect downstream performance for disease detection. To this end, we performed a comprehensive subgroup performance analysis comparing different disease detection models built with and without the use of the chest radiography foundation model. First, we used the chest radiography foundation model as a feature extractor to build three different disease detection models by training classification submodels with increasing complexity. The classification submodels take the features generated by the chest radiography foundation model as inputs and produce multilabel, probabilistic outputs for different disease labels. The three submodels correspond to a single, fully connected classification layer, denoted as CXR–linear, and two multilayer perceptrons (MLPs) with three and five hidden layers, denoted as CXR-MLP-3 and CXR-MLP-5, respectively. These disease detection models represent the intended use of the chest radiography foundation model, acting as a mechanism to facilitate effective transfer learning for task-specific training of prediction models. All three classification models were trained using the CheXpert training set with the corresponding validation set being used for model selection. We then compared the performance of these models to our baseline CheXpert model, a DenseNet-121, trained on the exact same data. All models were then evaluated on the CheXpert test set, using test set resampling to correct for demographic variations across subgroups. Here, we followed the test set resampling strategy for an unbiased estimation of subgroup performance as described by Glocker et al (19). We used resampling with replacement to construct balanced test sets, correcting for racial imbalance, differences in age, and varying prevalence of disease. In this study, we evaluated and compared disease detection performance on four different labels ("no finding," "pleural effusion," "cardiomegaly," and "pneumothorax") to provide a variety of results and insights across different target predictions.

Statistical Analysis

To determine whether the features generated by a model were biased, we used two-sample Kolmogorov–Smirnov tests to determine *P* values for the null hypothesis that the marginal distributions for a given pair of subgroups are identical in each of the first four modes of PCA. These statistical tests were performed for all relevant pairwise comparisons regarding the presence of disease, biologic sex, and race. The *P* values were adjusted for multiple testing using the Benjamini–Yekutieli procedure, and significance was determined at a 95% confidence level ($P < .05$).

To evaluate and compare the disease detection performance of different models, we computed the AUC, true-positive rate (TPR), and false-positive rate (FPR). TPR and FPR in subgroups were determined at a fixed decision threshold, which was optimized for each model to yield an FPR of 0.20 on the whole patient sample. The fixed target FPR allows for immediate identification of performance deviations across subgroups. To provide a single measure of classification performance, we report the Youden J statistic (measured at the target FPR), which is defined as J = TPR − FPR. We used bootstrapping with 2000 samples to calculate 95% CIs.

All information to recreate the exact study sample used in this article, including splits of training, validation, and test sets, and all code that is required for reproducing the results are available under an open-source Apache 2.0 license in a dedicated GitHub repository *(https://github.com/biomedia-mira/cxr-foundation-bias)*. All deep learning models were implemented in PyTorch. The model inspection via PCA and *t*-SNE was performed using scikit-learn. All statistical tests were performed using SciPy, version 1.10.0.

Results

Patient Characteristics

The study included 127 118 chest radiographs from 42 884 Asian, Black, and White patients (mean age, 63 years ± 17 [SD]; 23 623 male, 19 261 female). Table 1 provides a full breakdown of the study sample characteristics.

Model Inspection

Figure 1 presents the PCA-based feature space analysis of the two inspected backbone models, showing marginal distributions for different subgroups across the first four PCA modes. The corresponding scatterplots are given in Figure S1. Visually, we observed more and larger differences in the marginal distributions for the chest radiography foundation model across the protected

Table 1: Characteristics of the Study Sample in CheXpert Dataset

Attribute	All Patients	White	Asian	Black	Male	Female
All data						
Patients	42 884	33 338	6642	2904	23 623	19 261
Scans	127 118	99 027 (78)	18 830 (15)	9261 (7)	74 682 (59)	52 436 (41)
Age (y)	63 ± 17	64 ± 17	61 ± 17	56 ± 17	62 ± 17	64 ± 18
Female	52 436 (41)	39 735 (40)	8132 (43)	4569 (49)	…	…
No finding	10 916 (9)	8236 (8)	1716 (9)	964 (10)	6280 (8)	4636 (9)
Pleural effusion	51 574 (41)	40 545 (41)	7953 (42)	3076 (33)	30 171 (40)	21 403 (41)
Cardiomegaly	15 790 (12)	11 561 (12)	2375 (13)	1854 (20)	9793 (13)	5997 (11)
Pneumothorax	12 237 (10)	9653 (10)	2013 (11)	571 (6)	7280 (10)	4957 (9)
Training data						
Patients	25 730	20 034	3945	1751	14 164	11 566
Scans	76 205	59 238 (78)	11 371 (15)	5596 (7)	44 773 (59)	31 432 (41)
Age (y)	63 ± 17	64 ± 17	62 ± 17	56 ± 17	62 ± 17	64 ± 18
Female	31 432 (41)	23 715 (40)	4976 (44)	2741 (49)	…	…
No finding	6514 (9)	4910 (8)	1046 (9)	558 (10)	3756 (8)	2758 (9)
Pleural effusion	31 015 (41)	24 405 (41)	4754 (42)	1856 (33)	18 174 (41)	12 841 (41)
Cardiomegaly	9353 (12)	6835 (12)	1397 (12)	1121 (20)	5864 (13)	3489 (11)
Pneumothorax	7435 (10)	5871 (10)	1235 (11)	329 (6)	4461 (10)	2974 (9)
Validation data						
Patients	4288	3348	666	274	2367	1921
Scans	12 673	9945 (79)	1809 (14)	919 (7)	7643 (60)	5030 (40)
Age (y)	62 ± 17	63 ± 17	62 ± 17	55 ± 16	62 ± 17	63 ± 18
Female	5030 (40)	3933 (40)	667 (37)	430 (47)	…	…
No finding	1086 (9)	817 (8)	175 (10)	94 (10)	602 (8)	484 (10)
Pleural effusion	5049 (40)	3988 (40)	738 (41)	323 (35)	3086 (40)	1963 (39)
Cardiomegaly	1548 (12)	1156 (12)	222 (12)	170 (18)	987 (13)	561 (11)
Pneumothorax	1220 (10)	956 (10)	168 (9)	96 (10)	744 10)	476 (9)
Test data						
Patients	12 866	9956	2031	879	7092	5774
Scans	38 240	29 844 (78)	5650 (15)	2746 (7)	22 266 (58)	15 974 (42)
Age (y)	63 ± 17	64 ± 17	61 ± 17	57 ± 16	63 ± 16	64 ± 18
Female	15 974 (42)	12 087 (41)	2489 (44)	1348 (49)	…	…
No finding	3316 (9)	2509 (8)	495 (9)	312 (11)	1922 (9)	1394 (9)
Pleural effusion	15 510 (41)	12 152 (41)	2461 (44)	897 (33)	8911 (40)	6599 (41)
Cardiomegaly	4889 (13)	3570 (12)	756 (13)	563 (21)	2942 (13)	1947 (12)
Pneumothorax	3582 (9)	2826 (9)	610 (11)	146 (5)	2075 (9)	1507 (9)

Note.—Breakdown of demographic characteristics over the set of patient scans by racial groups and training, validation, and test splits. Percentages in parentheses are with respect to the number of scans. Age is reported as mean ± SD. The number of unique patients is also reported.

characteristics of biologic sex and race. This is particularly visible in the subgroup distributions for biologic sex (second column in Fig 1) where clear shifts between male and female patients were observed in all four PCA modes, whereas no obvious separation is visible for the CheXpert model. Similarly, we observed larger differences in the distributions of racial groups in the chest radiography foundation model compared with the model trained on CheXpert (third column in Fig 1). Figure 2 presents the marginal distributions for the *t*-SNE projections, with the corresponding scatterplots shown in Figure S2. Because the orientation of *t*-SNE dimensions is somewhat arbitrary, it was generally more difficult to visually observe any potential relationship between disease information and protected characteristics. However, we still observed larger differences between subgroup distributions for both biologic sex and race for the chest radiography foundation model compared with the CheXpert model, which was visible when we focused on the marginal distributions in the second and third columns of Figure 2.

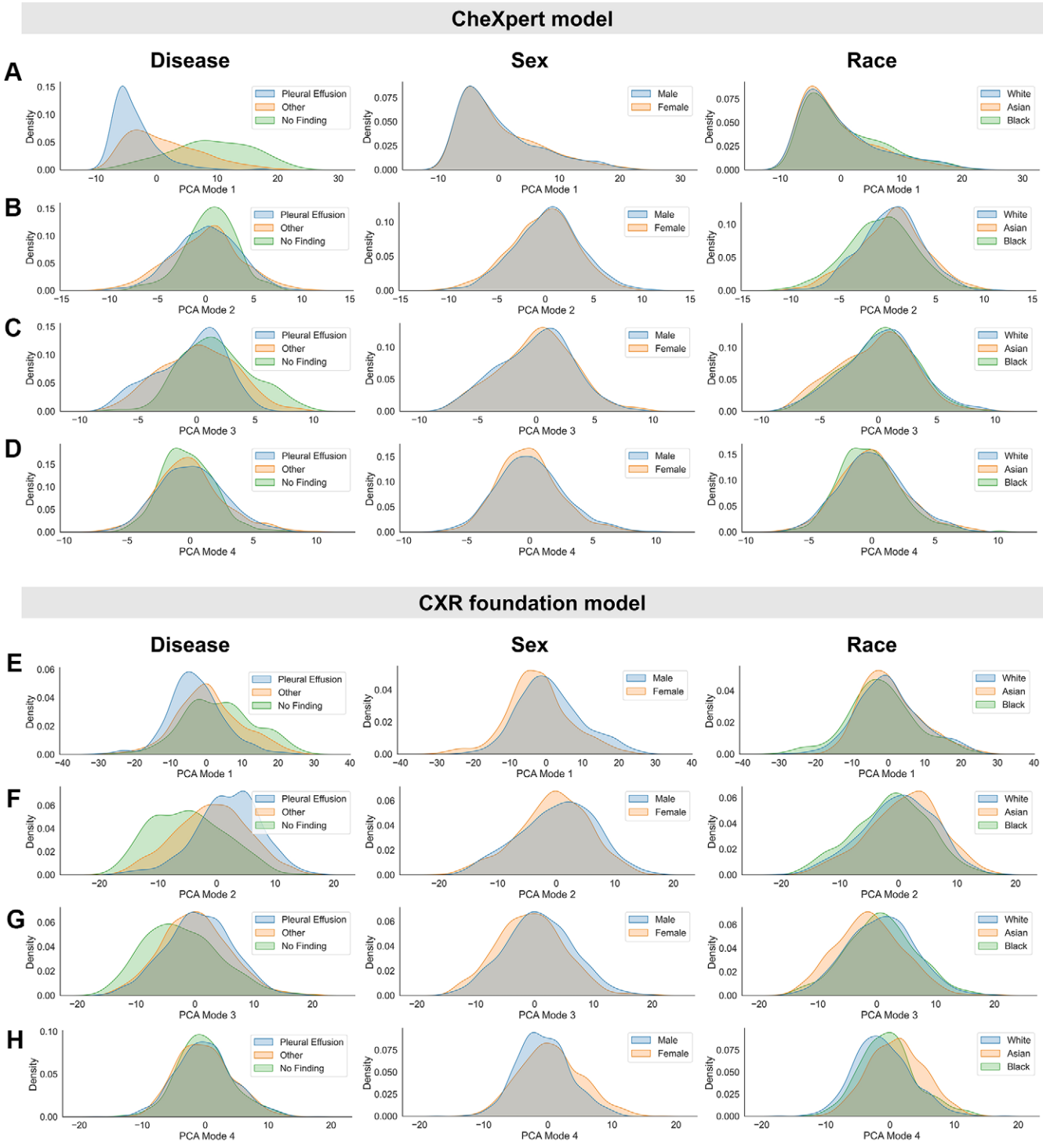


Figure 1: Inspection of subgroup distribution shifts in the principal component analysis (PCA) feature space projections. Marginal distributions are plotted across subgroups for the first four modes of PCA applied to the extracted feature vectors of the CheXpert test data for **(A–D)** the CheXpert model and **(E–H)** the chest radiography (CXR) foundation model. The plots were generated using a random set of 3000 patients (1000 samples from each racial group). Marginal distributions were normalized independently to remove differences in subgroup base rates and are shown for different characteristics (from left to right): presence of disease, biologic sex, and racial identity. Larger distribution shifts across sex and race are observed for the chest radiography foundation model.

The statistical analysis confirmed these qualitative observations (Table 2). For biologic sex, we found significant differences between the marginal distributions for male and female patients in all four PCA modes ($P < .001$, $P = .0013$, $P < .001$, $P < .001$), compared with no evidence of differences found in the CheXpert model ($P > .99$, $P = .26$, $P > .99$, $P = .15$). Significant differences are also found between the groups of Asian and Black patients in all four PCA modes in the chest radiography foundation model (all $P < .001$) versus two significant differences in the first and second mode of PCA for the CheXpert model ($P = .021$, $P < .001$, $P = .29$, $P = .40$). More differences were also observed between White and Asian and White and Black

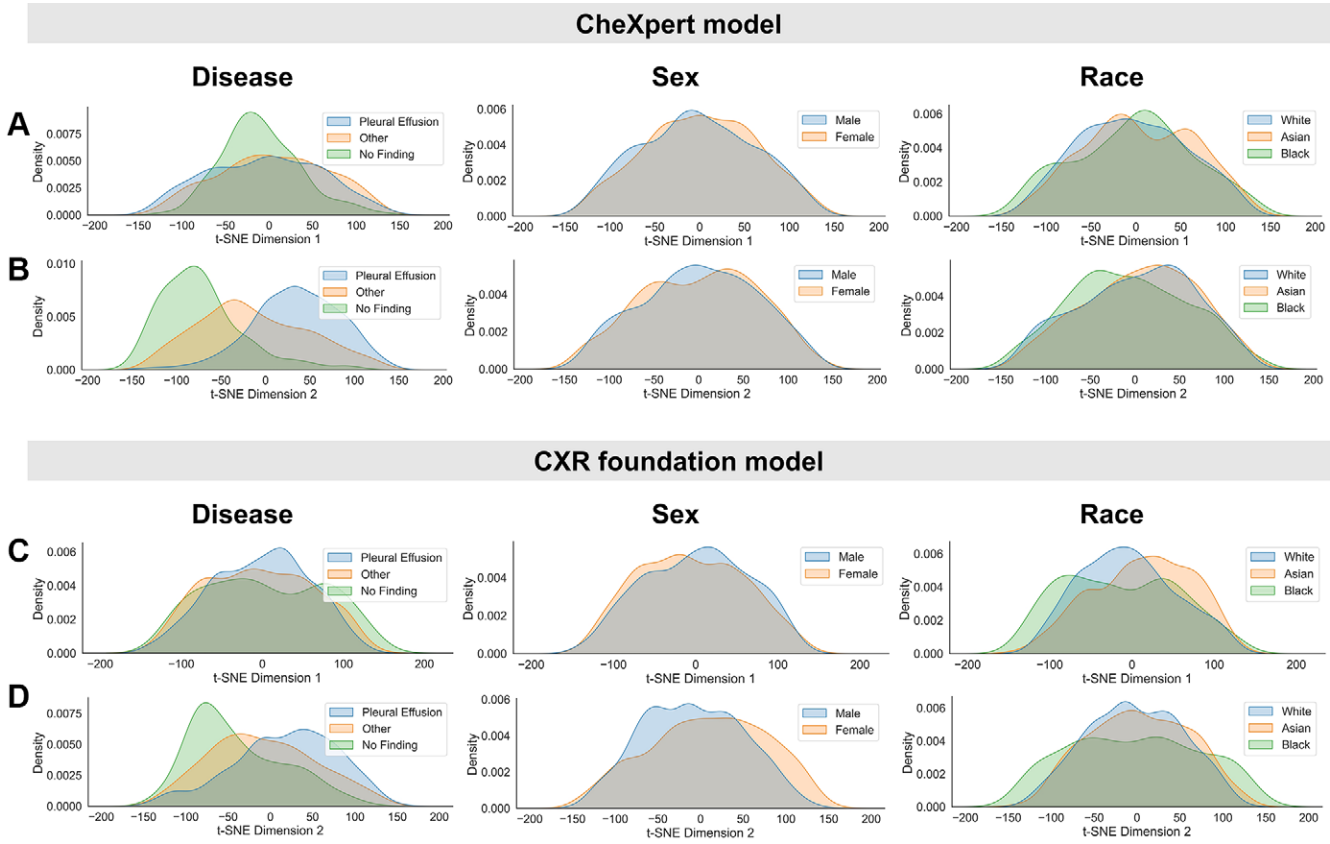


Figure 2: Inspection of subgroup distribution shifts in the *t*-distributed stochastic neighbor embedding (*t*-SNE) feature space projections. Marginal distributions are plotted across subgroups for the two dimensions of *t*-SNE applied to the extracted feature vectors of the CheXpert test data for **(A, B)** the CheXpert model and **(C, D)** the chest radiography (CXR) foundation model. The plots were generated using a random set of 3000 patients (1000 samples from each racial group). Marginal distributions were normalized independently to remove differences in subgroup base rates and are shown for different characteristics (from left to right): presence of disease, biologic sex, and racial identity. Larger distribution shifts across sex and race are observed for the chest radiography foundation model.

Table 2: Kolmogorov–Smirnov Tests for Comparing Marginal Distributions across Principal Component Analysis Modes

		P Values				
Mode	Exp. Var. (%)	No Finding vs Pleural Effusion	White vs Asian Patients	Asian vs Black Patients	Black vs White Patients	Male vs Female Patients
CheXpert model						
PCA mode 1	31.2	<.0001*	>.99	.021†	.40	>.99
PCA mode 2	9.0	.012†	>.99	<.0001*	<.0001*	.26
PCA mode 3	7.9	<.0001*	.025†	.29	>.99	>.99
PCA mode 4	5.1	.12	>.99	.40	.16	.15
Chest radiography foundation model						
PCA mode 1	16.8	<.0001*	.96	<.0001*	.00031*	<.0001*
PCA mode 2	8.6	<.0001*	.0013†	<.0001*	.0044†	.0013†
PCA mode 3	7.3	<.0001*	<.0001*	<.0001*	>.99	<.0001*
PCA mode 4	4.2	>.99	<.0001*	<.0001*	.00018*	<.0001*

Note.—Two-sample Kolmogorov–Smirnov tests were performed between the pairs of subgroups indicated in each column. The *P* values are adjusted for multiple testing using the Benjamini–Yekutieli procedure, and significance is determined at a 95% confidence level. Exp. Var. = explained variance, PCA = principal component analysis.
*$P < .001$.
†$P < .05$.

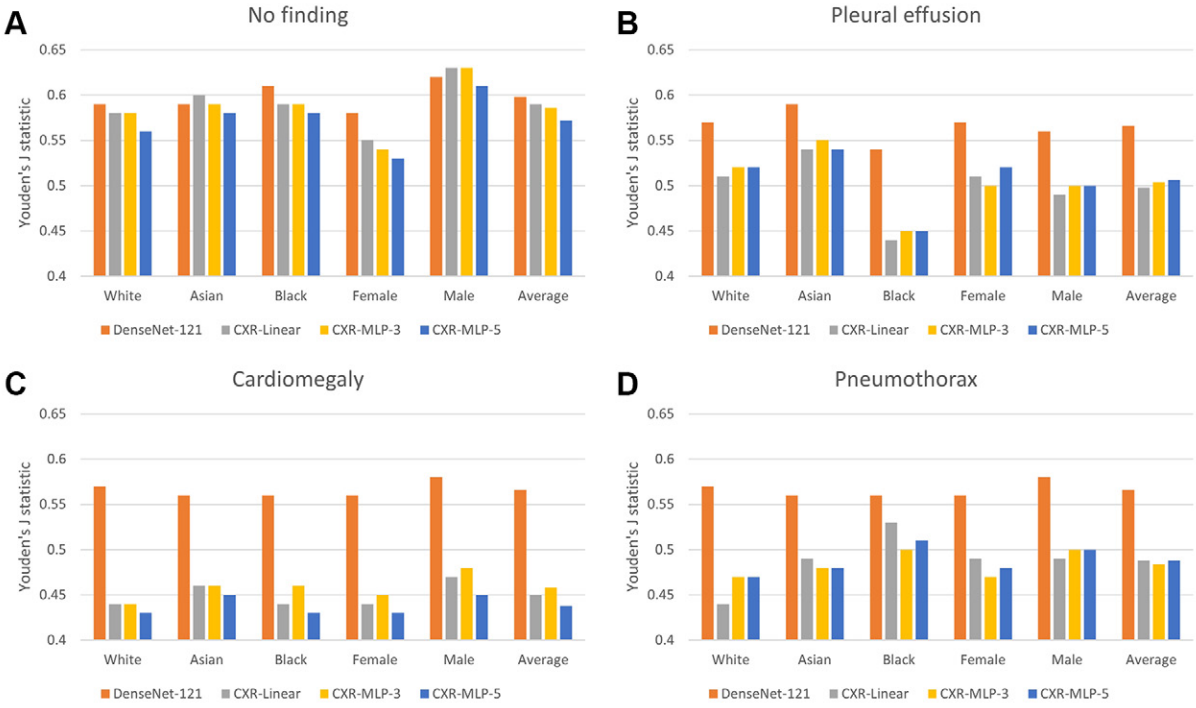


Figure 3: Comparison of disease detection performance across patient subgroups. Average classification performance across patient subgroups is shown in terms of Youden J statistic for the DenseNet-121 CheXpert model and three variants of the chest radiography foundation model. Classification performance is shown on four different labels of **(A)** "no finding," **(B)** "pleural effusion," **(C)** "cardiomegaly," and **(D)** "pneumothorax." The chest radiography foundation models consistently underperformed compared with the CheXpert model, with specific underperformance on the subgroup of female patients for "no finding" and the subgroup of Black patients on "pleural effusion." There was also a drastic decrease in overall performance across all subgroups for the chest radiography foundation models for "cardiomegaly." CXR-linear = submodel for chest radiography with single, fully connected classification layer, CXR-MLP-3 = submodel for chest radiography with three hidden layers, CXR-MLP-5 = submodel for chest radiography with five hidden layers, MLP = multilayer perceptrons.

patients in the chest radiography foundation model compared with the CheXpert model. Focusing on the first three PCA modes, which primarily capture differences in the features related to presence of disease (indicated by the significant differences between "no finding" and "pleural effusion"), we found that 10 of 12 pairwise comparisons on protected characteristics of biologic sex and race showed significant differences in the chest radiography foundation model, compared with four of 12 significant tests in the CheXpert model. Considering the explained variance for each PCA mode (see Table 2), we found that the first three PCA modes combined explained more than 53% of the variance in the CheXpert model compared with 37% in the chest radiography foundation model, indicating that the latter captures substantially more information in its feature representation that may be unrelated to disease prediction. To rule out within-patient cluster effects due to the presence of multiple scans per patient within the test set, we redid the analysis for a subsampled test set with only one scan per patient. The overall findings and conclusions remained unchanged, confirming the larger disparities for the chest radiography foundation model.

Model Performance

The differences in performance in terms of the Youden J statistic across models and patient subgroups are summarized in Figure 3. The models built on top of the chest radiography foundation model (CXR-linear, CXR-MLP-3, and CXR-MLP-5) consistently underperformed compared with the CheXpert model. Compared with average model performance across all subgroups, performance in detecting "no finding" decreased between 6.8% and 7.8% for female patients, and performance in detecting "pleural effusion" decreased between 10.7% and 11.6% for Black patients. We also observed a drastic decrease in overall performance in classifying "cardiomegaly" across all patient groups. In addition, we observed a clear difference in relative performance, leading to concerning subgroup disparities. Figure 4 presents the relative change in performance in terms of the Youden J statistic for each subgroup when compared with each model's average performance over all subgroups. We observed substantially larger disparities in relative performance across biologic sex and race for the chest radiography foundation models compared with the CheXpert model. The absolute and relative performances in terms of AUC are summarized in Figures S3 and S4, with similar findings of larger performance disparities across subgroups for the chest radiography foundation model. Detailed results of the subgroup performance analysis with various performance metrics are given in Tables S1–S4.

Discussion

This investigation aimed to highlight the potential risks of using foundation models in the development of medical imag-

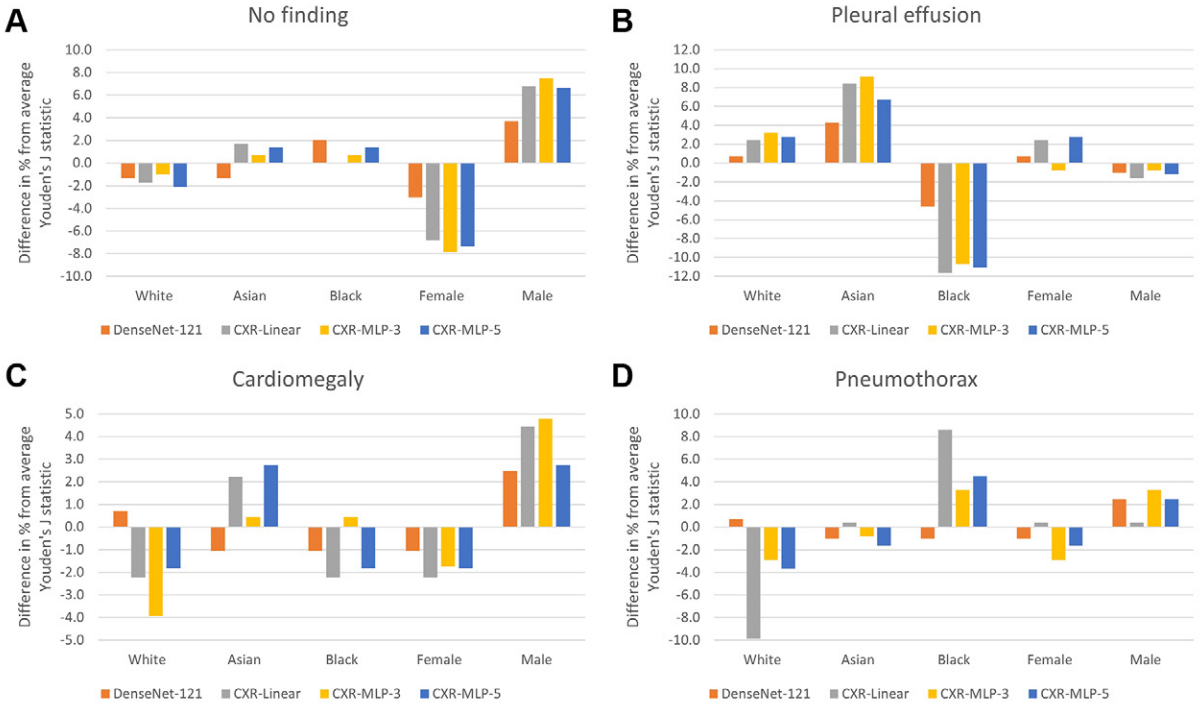


Figure 4: Relative change in disease detection performance across patient subgroups. The relative change in performance for each subgroup was measured by comparing the subgroup performance with each model's average performance over all subgroups. Performance is measured in terms of Youden J statistic on the labels of **(A)** "no finding," **(B)** "pleural effusion," **(C)** "cardiomegaly," and **(D)** "pneumothorax." There were substantially larger disparities in relative performance across biologic sex and race for the three chest radiography foundation models, CXR-linear, CXR-MLP-3, and CXR-MLP-5 when compared with the DenseNet-121 CheXpert model. CXR-linear = submodel for chest radiography with single, fully connected classification layer, CXR-MLP-3 = submodel for chest radiography with three hidden layers, CXR-MLP-5 = submodel for chest radiography with five hidden layers, MLP = multilayer perceptrons.

ing artificial intelligence. The fact that the investigated chest radiography foundation model encodes protected characteristics more strongly than does a task-specific backbone raises concerns because these biases could amplify already existing health disparities (21–24). Our bias analysis showed significant differences between features related to disease detection across biologic sex ($P < .001$) and race ($P < .001$). When using the foundation model in downstream disease detection, our subgroup performance analysis revealed a substantial degradation in classification performance, with specific disparities in protected subgroups. Classification performance on the "no finding" label decreased between 6.8% and 7.8% for female patients, and performance in detecting "pleural effusion" decreased between 10.7% and 11.6% for Black patients.

These findings are in line with the study by Seyyed-Kalantari et al (16), who found performance disparities in chest radiograph disease detection across underrepresented subgroups. Our results indicate a risk of bias for classification models built on top of features extracted with the chest radiography foundation model. Identifying these issues for the CheXpert dataset is noteworthy because this dataset was specifically used in the original study to evaluate the generalization ability of the foundation model (10). This highlights that even for artificial intelligence developers, it remains difficult to assess whether their models may be suitable for a specific target dataset. End users of third-party foundation models, who may have less knowledge of and insight into model pretraining, may find it even more difficult to assess risk of bias for their specific application and data. This is particularly concerning in light of recent studies demonstrating that medical images encode protected characteristics that can be recognized by deep learning models (13,14). The availability of diverse and representative datasets with detailed demographic information will be key for algorithmic auditing and comprehensive assessment of algorithmic bias (25–27).

We believe that our findings have implications beyond the studied model because the difficulty of scrutinizing foundation models applies in general, as pointed out in the work by Bommasani et al (5). This difficulty stems from the fact that detailed information about the data-generating processes and the exact training strategies is often missing. If biases remain undetected, they can cause serious harm, such as underdiagnosis of underserved populations (16). To mitigate these risks, it will be important to better understand how biases are encoded and how we may prevent the use of undesired information in prediction tasks (19,28). Of note, when a prediction model is trained via fine-tuning of a pretrained foundation model, we typically have two options: *(a)* unfreeze the backbone model, allowing the fine-tuning process to modify the mapping from input images to features, potentially overriding biases in the backbone, or *(b)* freeze the backbone model and learn only the parameters specific to the task prediction model, which usually requires substantially fewer training data and is therefore more appealing in practice. Arguably, the latter is more likely to carry forward any biases from the backbone because the method with which features are

generated remains unchanged. If the backbone features separate patients based on a protected characteristic, it is likely that the task-specific prediction model will learn separate mechanisms for the different subgroups. Fine-tuning is unlikely to be able to unlearn biases and may exploit shortcuts for making predictions because of the presence of undesirable correlations in the training data (29,30). In this context, the provided access to the chest radiography foundation model by Sellergren et al (10) is problematic because the developers offer option *b*. The original backbone model is not publicly shared; hence, one cannot update the mechanism for feature extraction when performing the task-specific fine-tuning, which limits the use of debiasing techniques (28,31). The observed differences in absolute and relative subgroup performance may be partly explained by the fact that the chest radiography foundation model was frozen during training of the classification submodels.

Our study had important limitations. We analyzed only one foundation model that was trained in a specific way, using supervised contrastive learning. Future work should explore whether biases also manifest in other chest radiography foundation models that are trained differently, such as via self-supervision without any annotations (8,9). Such models, however, are currently not publicly available. We may expect to find similar biases in models with fully self-supervised training because such training strategies encourage grouping of individuals in feature spaces that are visually similar. Therefore, we may expect to see clusters for biologic sex and potentially race; these characteristics are known to be separable with high predictive accuracy (13,14,32). We believe that our work may provide a methodologic basis for future bias analyses. Another limitation was that we could not shed light on the exact origin of the biases in the studied foundation model because of insufficient insights into the exact training data characteristics. Although the amount of training data for the foundation model was considerably large, with more than 800 000 chest radiographs, it was limited to data from two countries: India and the United States. Most images, more than 700 000, were reported to come from India, which may contribute to the observed bias across racial subgroups. It has been argued previously that mixing effects from dataset bias are notoriously difficult to analyze (33). Thus, a more systematic approach with controlled, simulated environments to specifically inject different types of bias may be required to isolate the effect of each bias on classification performance.

In conclusion, our study demonstrates that biases in the chest radiography foundation model related to race and biologic sex led to substantial performance disparities across protected subgroups. To minimize the risk of bias associated with use of foundation models in critical applications such as clinical decision-making, we argue that these models need to be fully accessible and transparent. This is important for allowing a more detailed analysis of potential biases and scrutiny of the resulting task-specific prediction models. Here, we advocate for comprehensive bias analysis and subgroup performance analysis to become integral parts in the development and auditing of future foundation models, which is essential for their safe and ethical use in health care applications.

Author contributions: Guarantors of integrity of entire study, all authors; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, all authors; experimental studies, all authors; statistical analysis, all authors; and manuscript editing, all authors

Disclosures of conflicts of interest: B.G. EU Commission (grant agreement no. 757173, Project MIRA, ERC-2017-STG, funding); Innovate UK/UKRI (UKRI London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare); Royal Academy of Engineering (Research Chair in Safe Deployment of Medical Imaging AI); scientific advisor for Kheiron Medical Technologies (January 2018–September 2021); part-time employment with Kheiron Medical Technologies with stock options as part of the standard compensation package (since October 2021); part-time employment with HeartFlow with stock options as part of the standard compensation package (since September 2018). **C.J.** Support from Microsoft Research—EPSRC iCASE Microsoft PhD Scholarship. **M.R.** Imperial College London, President's PhD Scholarship. **S.W.** Innovate UK/UKRI (UKRI London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare).

References

1. Çallı E, Sogancioglu E, van Ginneken B, van Leeuwen KG, Murphy K. Deep learning for chest X-ray analysis: a survey. Med Image Anal 2021;72:102125.
2. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021;385(3):283–286.
3. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. Nat Commun 2020;11(1):3673.
4. Cohen JP, Hashir M, Brooks R, Bertrand H. On the limits of cross-domain generalization in automated X-ray prediction. In: Arbel T, Ben Ayed I, de Bruijne M, Descoteaux M, Lombaert H, Pal C, eds. Proceedings of the Third Conference on Medical Imaging with Deep Learning. PMLR; 06–08 Jul 2020; 136–155. https://proceedings.mlr.press/v121/cohen20a.html.
5. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. arXiv 2108.07258 [preprint] https://arxiv.org/abs/2108.07258. Posted August 16, 2021. Accessed January 25, 2023.
6. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature 2023;616(7956):259–265.
7. Willemink MJ, Roth HR, Sandfort V. Toward foundational deep learning models for medical imaging in the new era of transformer networks. Radiol Artif Intell 2022;4(6):e210284.
8. Azizi S, Culp L, Freyberg J, et al. Robust and efficient medical imaging with self-supervision. arXiv 2205.09723 [preprint] https://arxiv.org/abs/2205.09723. Posted May 19, 2022. Accessed January 25, 2023.
9. Ghesu FC, Georgescu B, Mansoor A, et al. Self-supervised learning from 100 million medical images. arXiv 2201.01283 [preprint] https://arxiv.org/abs/2201.01283. Posted January 4, 2022. Accessed January 25, 2023.
10. Sellergren AB, Chen C, Nabulsi Z, et al. Simplified transfer learning for chest radiography models using less data. Radiology 2022;305(2):454–465.
11. Wójcik MA. Foundation Models in Healthcare: Opportunities, Biases and Regulatory Prospects in Europe. Electronic Government and the Information Systems Perspective. Springer International Publishing, 2022; 32–46.
12. Wiggins WF, Tejani AS. On the opportunities and risks of foundation models for natural language processing in radiology. Radiol Artif Intell 2022;4(4):e220119.
13. Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health 2022;4(6):e406–e414.
14. Adleberg J, Wardeh A, Doo FX, et al. Predicting patient demographics from chest radiographs with deep learning. J Am Coll Radiol 2022;19(10):1151–1161.
15. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci USA 2020;117(23):12592–12594.
16. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27(12):2176–2182.
17. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proc AAAI Conf Artif Intell 2019;33(01):590–597.
18. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017; 4700–4708.

19. Glocker B, Jones C, Bernhardt M, Winzeck S. Algorithmic encoding of protected characteristics in chest X-ray disease detection models. EBioMedicine 2023;89:104467.
20. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. Pac Symp Biocomput 2021;26:232–243.
21. Pinkerton KE, Harbaugh M, Han MK, et al. Women and lung disease. Sex differences and global health disparities. Am J Respir Crit Care Med 2015;192(1):11–16.
22. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol 2018;154(11):1247–1248 .
23. Cho MK. Rising to the challenge of bias in health care AI. Nat Med 2021;27(12):2079–2081.
24. Cullen MR, Lemeshow AR, Russo LJ, Barnes DM, Ababio Y, Habtezion A. Disease-specific health disparities: a targeted review focusing on race and ethnicity. Healthcare (Basel) 2022;10(4):603.
25. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. Lancet Digit Health 2022;4(5):e384–e397. [Published correction appears in Lancet Digit Health 2022;4(6):e405.]
26. McCradden MD, Anderson JAA, A Stephenson E, et al. A research ethics framework for the clinical translation of healthcare machine learning. Am J Bioeth 2022;22(5):8–22.
27. Yi PH, Kim TK, Siegel E, Yahyavi-Firouz-Abadi N. Demographic reporting in publicly available chest radiograph data sets: opportunities for mitigating sex and racial disparities in deep learning models. J Am Coll Radiol 2022;19(1 Pt B):192–200.
28. Marcinkevics R, Ozkan E, Vogt JE. Debiasing deep chest x-ray classifiers using intra- and post-processing methods. In: Lipton Z, Ranganath R, Sendak M, Sjoding M, Yeung S, eds. Proceedings of the 7th Machine Learning for Healthcare Conference. PMLR; 5–6 Aug 2022; 504–536. https://proceedings.mlr.press/v182/marcinkevics22a.html.
29. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. medRxiv 2020.09.13.20193565 [preprint]. https://www.medrxiv.org/content/10.1101/2020.09.13.20193565v2. Posted October 8, 2020. Accessed January 25, 2023.
30. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: CHIL '20: Proceedings of the ACM Conference on Health, Inference, and Learning, 2020; 151–159.
31. Nazer LH, Zatarah R, Waldrip S, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digit Health 2023;2(6):e0000278.
32. 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(5):949–954.
33. Bernhardt M, Jones C, Glocker B. Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. Nat Med 2022;28(6):1157–1158.