On Mitigating Shortcut Learning for Fair Chest X-ray Classification under Distribution Shift

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

As machine learning models reach human level performance on many real-world medical imaging tasks, it is crucial to consider the mechanisms they may be using to make such predictions. Prior work has demonstrated the surprising ability of deep learning models to recover demographic information from chest X-rays. This suggests that disease classification models could potentially be utilizing these demographics as shortcuts, leading to prior observed performance gaps between demographic groups. In this work, we start by investigating whether chest X-ray models indeed use demographic information as shortcuts when classifying four different diseases. Next, we apply five existing methods for tackling spurious correlations, and examine *performance* and *fairness* both for the ***original*** dataset and five ***external*** hospitals. Our results indicate that shortcut learning can be corrected to remedy in-distribution fairness gaps, though this reduction often does not transfer under domain shift. We also find trade-offs between fairness and other important metrics, raising the question of whether it is beneficial to remove such shortcuts in the first place.

1 Introduction

Real-world data often contain *spurious correlations* [1, 2], which are features in the training data that are correlated with the label, but are not used in the true label function [3]. Models trained to minimize empirical risk often utilize these correlations as *shortcuts*, relying solely on these features to make predictions. Such models then exhibit poor worst-group accuracy (WGA), gaps in class-conditioned accuracy across different attributes, as well as catastrophic performance drops when deployed in an environment where attribute characteristics change [4].

In the field of medicine, machine learning models are being increasingly deployed in real-world clinical environments [5, 6]. In such settings, it is important to consider not only overall model performance, but also potential model biases across demographic groups [7, 8]. Though deep learning has reached human level performance in many tasks in the medical imaging domain [9, 10, 11], prior works have found that they often exhibit biases in the form of performance disparities across protected groups [12, 13, 14, 15]. For example, It has been shown that chest X-ray classifiers trained to predict the presence of any disease systematically underdiagnose Black patients [16], which could lead to delays in care. In order to ensure safe and equitable deployment of such models, it is crucial to understand the source of such biases, and, where possible, take actions to correct them [17, 18].

In a parallel line of work, researchers have found the surprising ability of deep models to predict demographic information from medical images, achieving performance far beyond that of radiologists. For example, self-reported patient race can be predicted with high accuracy from chest X-rays, chest CTs, and mammographs [19], and gender and age can also be predicted from X-rays with high accuracy [20]. This suggests that such demographic attributes may be used as a potential *shortcut* for disease prediction models.

Submitted to the Workshop on Distribution Shifts (DistShift@NeurIPS 2023). Do not distribute.

In this work, we connect these findings from shortcut learning and algorithmic fairness to ask the question: Do chest X-ray disease classification models use demographics as shortcuts, and what happens if we remove this shortcut when learning the model? We make the following empirical contributions:

1. We follow prior work [21] in showing that representations learned for disease prediction using chest X-rays encode demographic information across age, race, sex, and the intersection of race and sex.
2. We show that encoding of demographic attributes is correlated with a greater fairness gap between demographic groups.
3. Applying a variety of existing machine learning methods for shortcut removal, we find that it is possible to achieve a fairer model with minimal loss in overall performance.
4. However, we find that these fairness interventions lead to worse calibration error, and the reduced fairness gaps in-distribution do not typically transfer to out-of-distribution external sites.

Ours findings underscore the need for broader evaluations across a wide range of metrics on both in-distribution and out-of-distribution data, as well as a careful consideration of the features that we want to integrate into clinical machine learning models [22].

2 Related Work

Spurious Correlations Spurious correlations, which is an instance of subpopulation shift [4], arise in a variety of real-world data settings [1]. For example, in the medical setting, chest X-ray models trained on multi-site data may use the site as a spurious correlation [23, 24]. Methods for tackling spurious correlations take several distinct approaches, including adversarial training [25, 26], robust optimization [27, 28], sample weighting [29, 30], final-layer retraining [3, 31, 32], data augmentation [33, 34], and weight averaging [35, 36].

Fair Medical Imaging There have been many prior works which demonstrate gaps in performance (typically measured using the false positive and false negative rates) between demographic groups in medical imaging tasks for various modalities, including chest X-rays [12, 13], MRIs [14], CT scans [37], and dermoscopic images [15]. Most relevant to this work is Seyyed-Kalantari et al. [16], which shows that chest X-ray models for predicting *No Finding* have higher false positive rate (i.e. underdiagnosis) for Black, female, and younger patients. Zhang et al. [38] applied various fairness algorithms to the same dataset, finding mixed results. Ktena et al. [39] used conditional diffusion models to generate synthetic images, finding improvements in both in-distribution and out-of-distribution fairness.

In comparison, our work approaches the fairness problem from the shortcut learning angle, which is a potential cause of the fairness gap due to the ability of deep models to predict demographic information chest X-rays [19, 20]. Our work is motivated by Glocker et al. [21], which shows that representations learned for disease classification contain demographic information, and Brown et al. [40], which proposes a test for shortcut learning in medical imaging. Compared with Brown et al. [40], our work (1) applies a wide range of algorithms, including the adversarial training approach examined in their paper, (2) examines trade-offs between fairness and a wide range of other metrics, and (3) examines model performance and fairness on external sites.

3 Experiments

We start by training DenseNet-121 [41] models (pre-trained on ImageNet [42]) on MIMIC-CXR [43], and evaluating on the same dataset (the in-distribution (ID) dataset). We examine four binary classification tasks, as they have been studied in prior work for potential biases [12, 16, 44]: *No Finding*, *Pneumothorax*, *Effusion*, and *Cardiomegaly*. We evaluate six algorithms: empirical risk minimization (**ERM**, [45]), resampling to equalize group size (**Resample**, [46]), GroupDRO (**GroupDRO**, [27]), domain adversarial training (**DANN**, [25]), domain adversarial training conditioned on the label (**CDANN**, [26]), and weight averaging (**MA**, [47]).

For each combination of task, algorithm, and demographic attribute, we conduct a random hyperparameter search [48] with 15 runs. Where applicable, we select the hyperparameter setting that

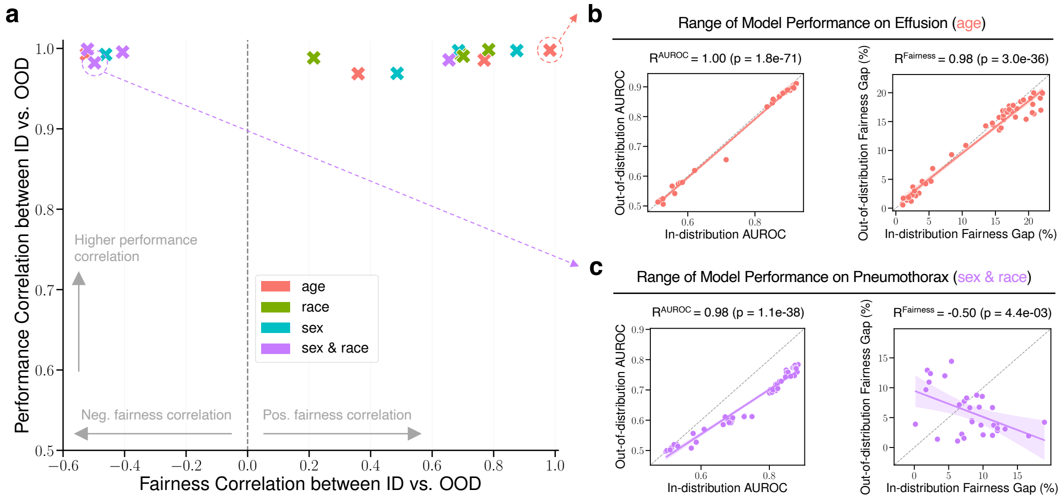

Figure 2: **Does fairness transfer under distribution shift?** We examine the transfer of *performance* (overall AUROC) and *fairness* between the ID (MIMIC-CXR) and OOD (all five other) datasets. **(a)** Pearson correlation coefficient of (ID vs. OOD) performance versus the Pearson correlation coefficient of (ID vs. OOD) fairness, where each point is a grid of trained models or a particular combination of task and attribute. We find that there is a high correlation between ID and OOD performance in all cases, but the correlation between ID and OOD fairness is tenuous. **(b), (c)** We show how two particular points in the first plot are obtained.

maximizes the worst-attribute validation AUROC. Confidence intervals are computed as the standard deviation across three different random seeds for each hyperparameter setting. We evaluate fairness as the False Positive Rate (FPR) gap for *No Finding*, and the False Negative Rate (FNR) gap for all other tasks (i.e. equal opportunity [49]), as these both correspond to underdiagnosis, which could lead to delays in treatment. For metrics where a binary decision is required, we binarize the score by selecting the threshold that maximizes the validation F1 score [50].

We then evaluate these models under domain shift, on CheXpert [51], NIH [52], SIIM [53], PadChest [54], and VinDr-CXR [55]. For convenience, we present aggregated results across the five sites as a single out-of-distribution (OOD) dataset. Dataset statistics can be found in Table A.1 and Table A.2.

4 Results

Disease Classification Models Encode Demographic Attributes and Are Unfair. We confirm that deep models trained for disease classification encode demographic attributes by training a linear attribute prediction head (i.e., logistic regression) on top of the feature extractor (weights frozen). Fig. B.1(a) shows that across different diseases and sensitive attributes, the penultimate layer of the models contain substantial information about demographic attributes, with attribute prediction AUC significantly higher than chance. In addition, we observe that these models are highly ***unfair*** across groups, where the fairness gaps can be larger than 20% (Fig. B.1(b)).

SOTA Algorithms Fix In-Distribution Fairness Gaps and Maintain Performance. In the ID setting (i.e., test on the same dataset), state-of-the-art robustness methods can effectively address fairness gaps while maintaining the overall performance (Fig. 1 and B.2). Specifically, ERM models exhibit large fairness gaps (e.g., models centered in the top right corner), whereas methods like GroupDRO and DANN can effectively close the gap while achieving similar AUC (e.g., the bottom right corner). We further plot the **Pareto front** that exploits the performance-fairness tradeoff across different diseases and attributes (Fig. 1 and B.2), where existing algorithms consistently balance the tradeoff, achieving high in-distribution fairness without losing overall performance for disease prediction.

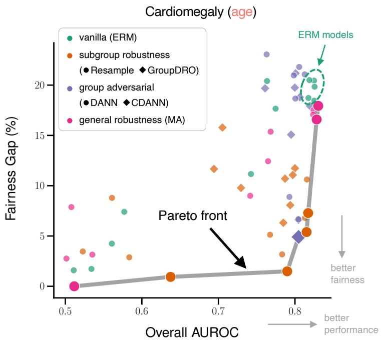

Figure 1: SOTA methods fix ID fairness gaps while maintaining performance.

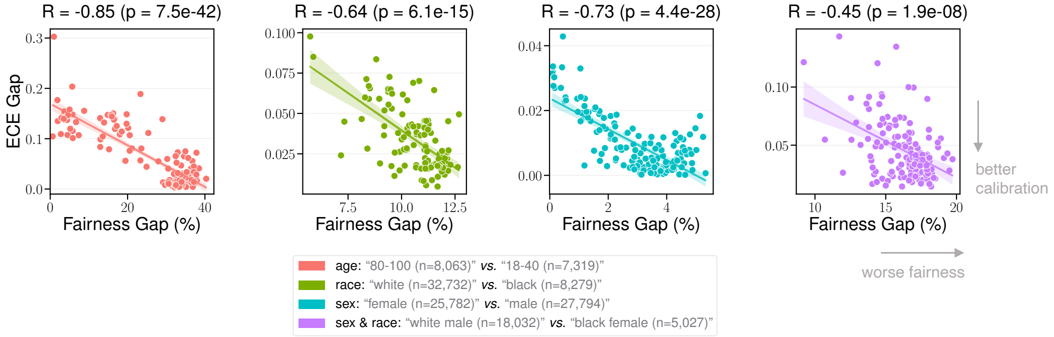

Figure 3: Inherent tradeoff between the fairness gap and the Expected Calibration Error (ECE) gap. Complete results for other metrics (e.g., WGA) are in Fig. B.3.

Fairness Does Not Transfer Under Distribution Shift. When deploying AI models in real settings, it is crucial to ensure models can generalize to data from unseen institutions or environments. We directly test all trained models in the **OOD** setting, where we report results on external datasets that are unseen during model training. Fig. 2 illustrates that the *performance correlation* between ID and OOD is high across different settings, consistent with prior work [56]. However, the *fairness correlation* between ID and OOD does not show consistent pattern. This indicates that a model that is fair ID does not necessarily deliver fair outcomes when tested OOD. The observation holds across diseases and attributes.

Metrics Beyond Fairness. Finally, we demonstrate the inherent tradeoff between fairness and other important metrics. First, we show that enforcing fair predictions across groups can result in worse expected calibration error gap (**ECE Gap**, Fig. B.3) between attributes, a result that is consistent with previous work showing a theoretical impossibility between probabilistic equalized odds and calibration by group [57, 58]. Next, we explore the relationship between fairness and worst-group accuracy (**WGA**, Fig. B.3), a common metric for evaluating shortcut reliance in the spurious correlation literature [4] (where groups are defined as the product of the attribute and the label). We find that, surprisingly, fairer models exhibit *worse* WGA. We hypothesize that, though fair models encode less demographic information (Fig. B.1) and thus cannot rely as much on the shortcut, this regularization leads to a worse model for all, a phenomenon that has been observed in prior work [38, 59, 60, 61]. This finding uncovers the limitation of blindly optimizing fairness, where more realistic evaluations are needed for reliable medical AI models.

5 Discussion

Overall, our results present a cautious view on the efficacy and consequences of removing demographic shortcuts in disease classification models. Though removing shortcuts fixes ID fairness, the trade-offs with other metrics, as well as the lack of transfer to external domains, questions whether it provides any utility in the first place. These considerations demonstrate the complexities of the healthcare setting, where the relationship between the demographics and the label are complex, there could be mislabelling in both variables [62, 63], and distribution shifts between domains are difficult to quantify. This clearly contrasts with simple datasets for spurious correlations such as Waterbirds [64], where relying only on the invariant "bird" features over the spurious "background" features would improve WGA both in-distribution, and out-of-distribution when the set of possible backgrounds change [4].

In this work, we frame demographic features as "shortcuts" – nuisances [65] which should not be utilized by the model to make disease predictions. However, some demographic variables could be a direct *causal* factor in some diseases (e.g. sex as a causal factor of breast cancer). In these cases, it would not be desirable to remove all demographic reliance, but instead match the reliance of the model on the demographic attribute to its true causal effect [66, 67]. In addition, in the tasks we have examined here, demographic variables such as race likely have an indirect causal effect on disease (e.g. through socioeconomic status), though this effect certainly varies across geographic location. Whether demographic variables should serve as proxies for these causal factors is a decision that should rest with the model developers [22, 68].

References

[1] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[2] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020.

[3] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

[4] Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. *arXiv preprint arXiv:2302.12254*, 2023.

[5] Mark P Sendak, Joshua D'Arcy, Sehj Kashyap, Michael Gao, Marshall Nichols, Kristin Corey, William Ratliff, and Suresh Balu. A path for translation of machine learning products into healthcare delivery. *EMJ Innov*, 10:19–00172, 2020.

[6] Angela Zhang, Lei Xing, James Zou, and Joseph C Wu. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, 6 (12):1330–1345, 2022.

[7] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

[8] Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3529–3530, 2020.

[9] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[10] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[11] Phillippe Burlina, Neil Joshi, Katia D Pacheco, David E Freund, Jun Kong, and Neil M Bressler. Utility of deep learning methods for referability classification of age-related macular degeneration. *JAMA ophthalmology*, 136(11):1305–1307, 2018.

[12] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

[13] Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. Medfair: Benchmarking fairness for medical imaging. *arXiv preprint arXiv:2210.01725*, 2022.

[14] Carolina Piçarra and Ben Glocker. Analysing race and sex bias in brain age prediction. *arXiv preprint arXiv:2309.10835*, 2023.

[15] Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. Fairness of classifiers across skin tones in dermatology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 320–329. Springer, 2020.

[16] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

[17] Adewole S Adamson and Avery Smith. Machine learning and health care disparities in dermatology. *JAMA dermatology*, 154(11):1247–1248, 2018.

[18] Melissa D McCradden, Shalmali Joshi, Mjaye Mazwi, and James A Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5):e221–e223, 2020.

[19] Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, et al. Reading race: Ai recognises patient's racial identity in medical images. *arXiv preprint arXiv:2107.10356*, 2021.

[20] Jason Adleberg, Amr Wardeh, Florence X Doo, Brett Marinelli, Tessa S Cook, David S Mendelson, and Alexander Kagen. Predicting patient demographics from chest radiographs with deep learning. *Journal of the American College of Radiology*, 19(10):1151–1161, 2022.

[21] Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. *Ebiomedicine*, 89, 2023.

[22] Vinith M Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. When personalization harms: Reconsidering the use of group attributes in prediction. *arXiv preprint arXiv:2206.02058*, 2022.

[23] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

[24] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

[25] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[26] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018.

[27] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[28] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33: 8847–8860, 2020.

[29] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

[30] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

[31] Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Overparameterisation and worst-case generalisation: friend or foe? In *International Conference on Learning Representations*, 2020.

[32] Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. *arXiv preprint arXiv:2210.11466*, 2022.

[33] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pages 25407–25437. PMLR, 2022.

[34] Zongbo Han, Zhipeng Liang, Fan Yang, Liu Liu, Lanqing Li, Yatao Bian, Peilin Zhao, Bingzhe Wu, Changqing Zhang, and Jianhua Yao. Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. *Advances in Neural Information Processing Systems*, 35:37704–37718, 2022.

[35] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

[36] Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022.

[37] Yuyin Zhou, Shih-Cheng Huang, Jason Alan Fries, Alaa Youssef, Timothy J Amrhein, Marcello Chang, Imon Banerjee, Daniel Rubin, Lei Xing, Nigam Shah, et al. Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr. *arXiv preprint arXiv:2111.11665*, 2021.

[38] Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. Improving the fairness of chest x-ray classifiers. In *Conference on Health, Inference, and Learning*, pages 204–233. PMLR, 2022.

[39] Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Alan Karthikesalingam, et al. Generative models improve fairness of medical classifiers under distribution shifts. *arXiv preprint arXiv:2304.09218*, 2023.

[40] Alexander Brown, Nenad Tomasev, Jan Freyberg, Yuan Liu, Alan Karthikesalingam, and Jessica Schrouff. Detecting shortcut learning for fair medical ai using shortcut testing. *Nature Communications*, 14(1):4314, 2023.

[41] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

[42] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[43] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[44] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[45] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

[46] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022.

[47] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[48] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[49] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[50] Zachary C Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. Optimal thresholding of classifiers to maximize f1 measure. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II 14*, pages 225–239. Springer, 2014.

[51] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[52] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[53] Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, Paras Lakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation, 2019. URL https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation/.

[54] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

[55] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

[56] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021.

[57] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

[58] Lydia T Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pages 4051–4060. PMLR, 2019.

[59] Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10421, 2022.

[60] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[61] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[62] Rajiv Movva, Divya Shanmugam, Kaihua Hou, Priya Pathak, John Guttag, Nikhil Garg, and Emma Pierson. Coarse race data conceals disparities in clinical risk score performance. *arXiv preprint arXiv:2304.09270*, 2023.

[63] Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A Young, Andrew Y Ng, Matthew P Lungren, and Pranav Rajpurkar. Visualchexbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 105–115, 2021.

[64] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[65] Aahlad Puli, Nitish Joshi, He He, and Rajesh Ranganath. Nuisances via negativa: Adjusting for spurious correlations via data augmentation. *arXiv preprint arXiv:2210.01302*, 2022.

[66] Nitish Joshi, Xiang Pan, and He He. Are all spurious features in natural language alike? an analysis through a causal lens. *arXiv preprint arXiv:2210.14011*, 2022.

[67] Abhinav Kumar, Amit Deshpande, and Amit Sharma. Causal effect regularization: Automated detection and removal of spurious attributes. *arXiv preprint arXiv:2306.11072*, 2023.

[68] Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.

A Dataset Statistics

Table A.1: **Dataset statistics for six chest X-ray classification datasets.** We train models on MIMIC, and evaluate on the remaining datasets.

		MIMIC [43]	**CheXpert** [51]	**NIH** [52]	**SIIM** [53]	**PadChest** [54]	**VinDr** [55]
	Location	Boston, MA	Stanford, CA	Bethesda, MD	Bethesda, MD	Alicante, Spain	Hanoi, Vietnam
	# Images	357,167	222,792	112,120	11,582	144,478	6,354
	% Frontal	64.5	85.5	100.0	100.0	69.1	100.0
	Sample Image						
Sex (%)	**Male**	52.2	59.3	56.5	55.4	49.6	56.9
	Female	47.8	40.7	43.5	44.6	50.4	43.1
Race (%)	**White**	61.0	56.4	-	-	-	-
	Black	15.6	5.4	-	-	-	-
	Asian	3.1	10.5	-	-	-	-
	Other	20.3	27.8	-	-	-	-
Age (%)	**0-18**	-	-	4.8	5.0	3.7	21.8
	18-40	13.8	13.9	27.7	27.3	9.2	16.0
	40-60	31.1	31.1	43.9	42.9	26.5	27.1
	60-80	40.0	39.0	22.7	23.9	38.0	30.0
	80-100	15.1	16.0	0.9	0.9	22.6	5.1
Intersection (%)	**White Male**	33.8	34.1	-	-	-	-
	White Female	27.3	22.2	-	-	-	-
	Black Male	6.3	2.7	-	-	-	-
	Black Female	9.3	2.6	-	-	-	-
	Asian Male	1.6	6.0	-	-	-	-
	Asian Female	1.5	4.5	-	-	-	-
	Others Male	10.5	16.5	-	-	-	-
	Others Female	9.8	11.3	-	-	-	-
Task Prevalence (%)	**No Finding**	39.8	10.0	53.8	-	34.9	41.2
	Effusion	20.0	38.6	11.9	-	5.9	7.5
	Pneumothorax	3.4	8.7	4.7	28.4	0.3	0.7
	Cardiomegaly	14.9	12.1	2.5	-	9.5	22.6

Table A.2: Prevalences of the four diseases examined in this work for each demographic attribute in MIMIC-CXR and CheXpert.

		MIMIC				**CheXpert**			
		Cardiomegaly	**No Finding**	**Effusion**	**Pneumothorax**	**Cardiomegaly**	**No Finding**	**Effusion**	**Pneumothorax**
Sex (%)	**Male**	14.8	37.2	21.1	4.0	12.4	9.9	38.4	9.0
	Female	15.1	42.6	18.9	2.8	11.6	10.2	38.8	8.3
Race (%)	**White**	15.5	34.6	24.0	4.0	11.5	9.4	39.4	9.1
	Black	17.6	44.3	13.4	1.8	19.6	11.7	31.7	5.8
	Asian	16.6	36.0	24.2	5.4	12.7	10.4	40.5	9.8
	Other	11.1	52.5	12.6	2.5	11.7	10.8	37.6	8.0
Age (%)	**18-40**	6.8	64.0	8.1	3.6	9.1	20.5	27.0	12.5
	40-60	11.4	46.5	15.0	3.0	10.1	12.4	36.2	8.6
	60-80	17.6	32.5	23.9	3.8	12.4	7.0	42.3	8.9
	80-100	22.9	23.3	31.0	3.0	17.9	3.7	44.2	5.0
Intersection (%)	**White Male**	15.4	33.3	24.4	4.4	12.4	9.4	39.0	9.2
	White Female	15.5	36.3	23.5	3.5	10.2	9.4	39.9	9.0
	Black Male	16.7	41.0	13.9	2.2	18.2	11.9	31.6	6.8
	Black Female	18.3	46.6	13.0	1.5	20.9	11.5	31.8	4.7
	Asian Male	16.4	33.6	25.5	6.1	12.8	10.1	40.4	9.8
	Asian Female	16.9	38.6	22.7	4.7	12.5	10.7	40.5	9.8
	Others Male	11.5	48.1	14.2	3.3	11.4	10.4	37.6	8.7
	Others Female	10.7	57.2	10.9	1.7	12.1	11.4	37.6	7.0

B Additional Experimental Results

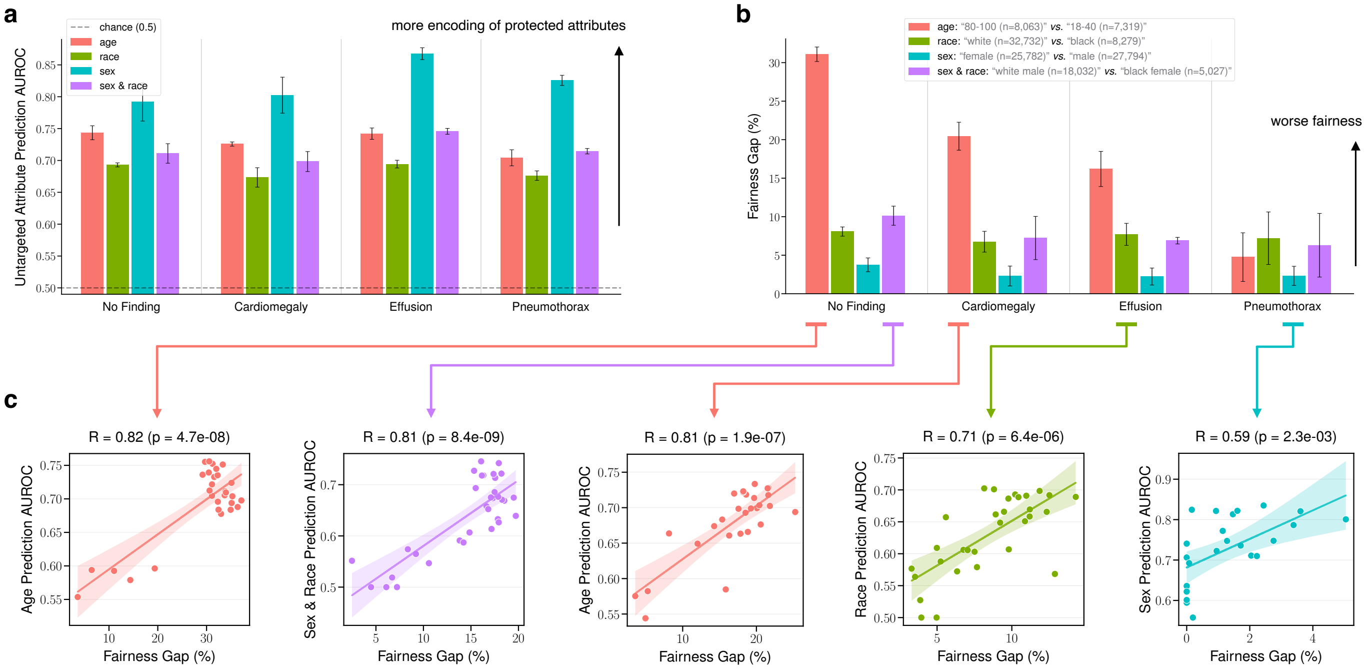

Figure B.1: We train ERM models on MIMIC-CXR to predict four different binary tasks. **(a)** We show the performance of a linear model that predicts the demographic attribute from frozen representations for the best ERM model, finding that ERM representations encode demographic attributes to a high degree. **(b)** We show the fairness gap, as defined by the FPR gap for *No Finding*, and the FNR gap for all other tasks for the best ERM model. We find that ERM models exhibit high fairness gaps, especially between age groups. **(c)** We examine the correlation between attribute prediction performance and fairness for all learned models (not only ERM), selecting models with overall validation AUROC ≥ 0.7. We find that there is a high correlation between the two.

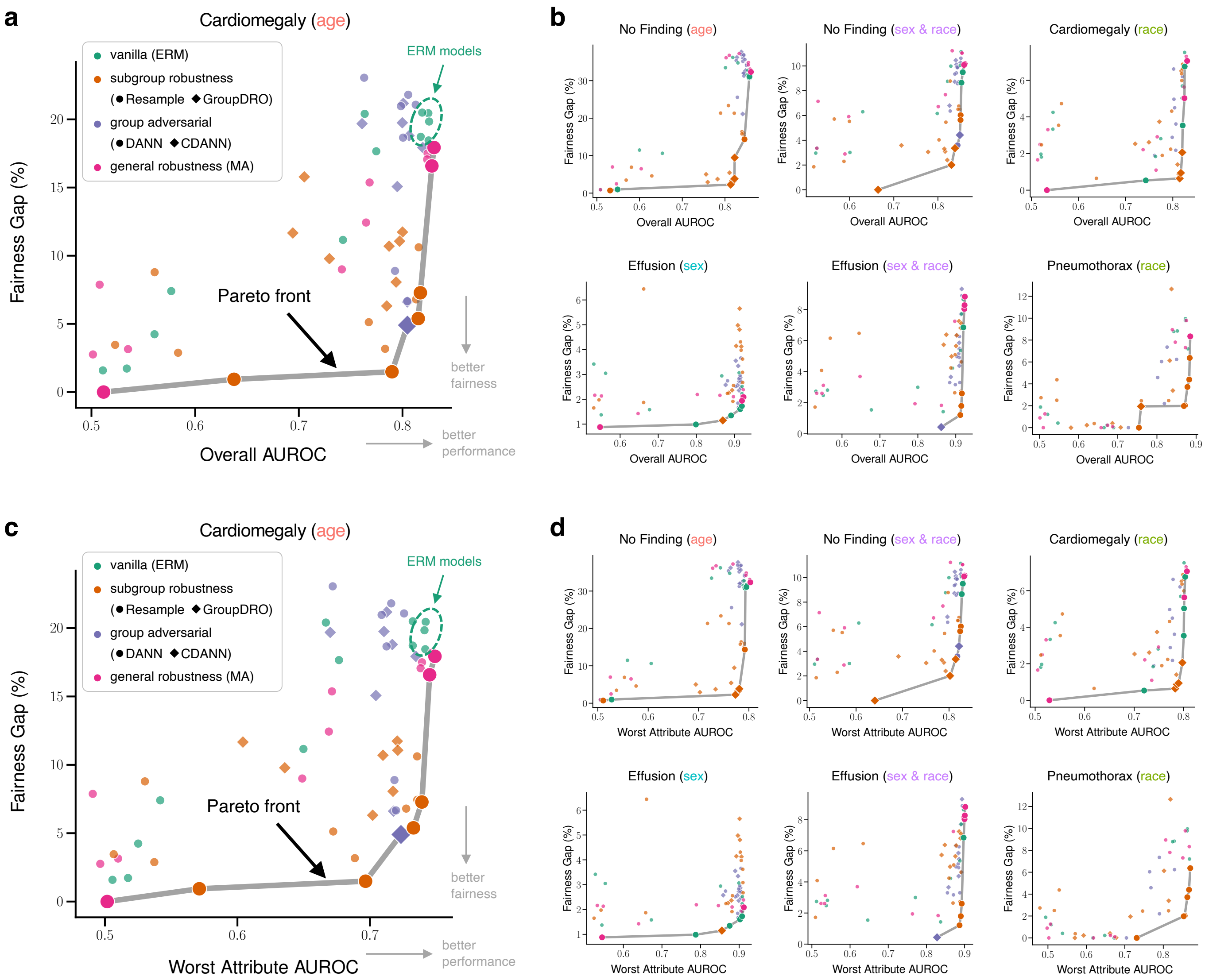

Figure B.2: We examine the trade-off between the fairness gap and two performance metrics (**(a)**, **(b)**: overall AUROC, **(c)**, **(d)**: worst-attribute AUROC) for all trained models. Each plot represents a specific disease prediction task (e.g., *Cardiomegaly*) with a specific attribute (e.g., *age*). In each case, we plot the Pareto front – the best achievable fairness gap with a minimum constraint on the performance. We find that for many tasks, it is possible to achieve a model that is fairer than ERM with minimal reduction of performance.

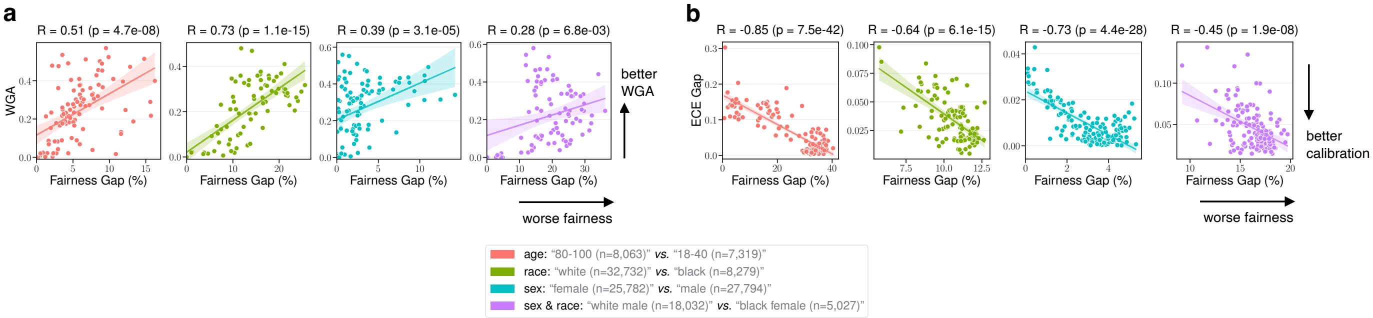

Figure B.3: For the *No Finding* task, we examine the trade-off between the fairness gap and **(a)** the Worst Group Accuracy (WGA), and **(b)** the Expected Calibration Error (ECE) gap. We find that enforcing fairness constraints lead to worsening of the other two metrics.

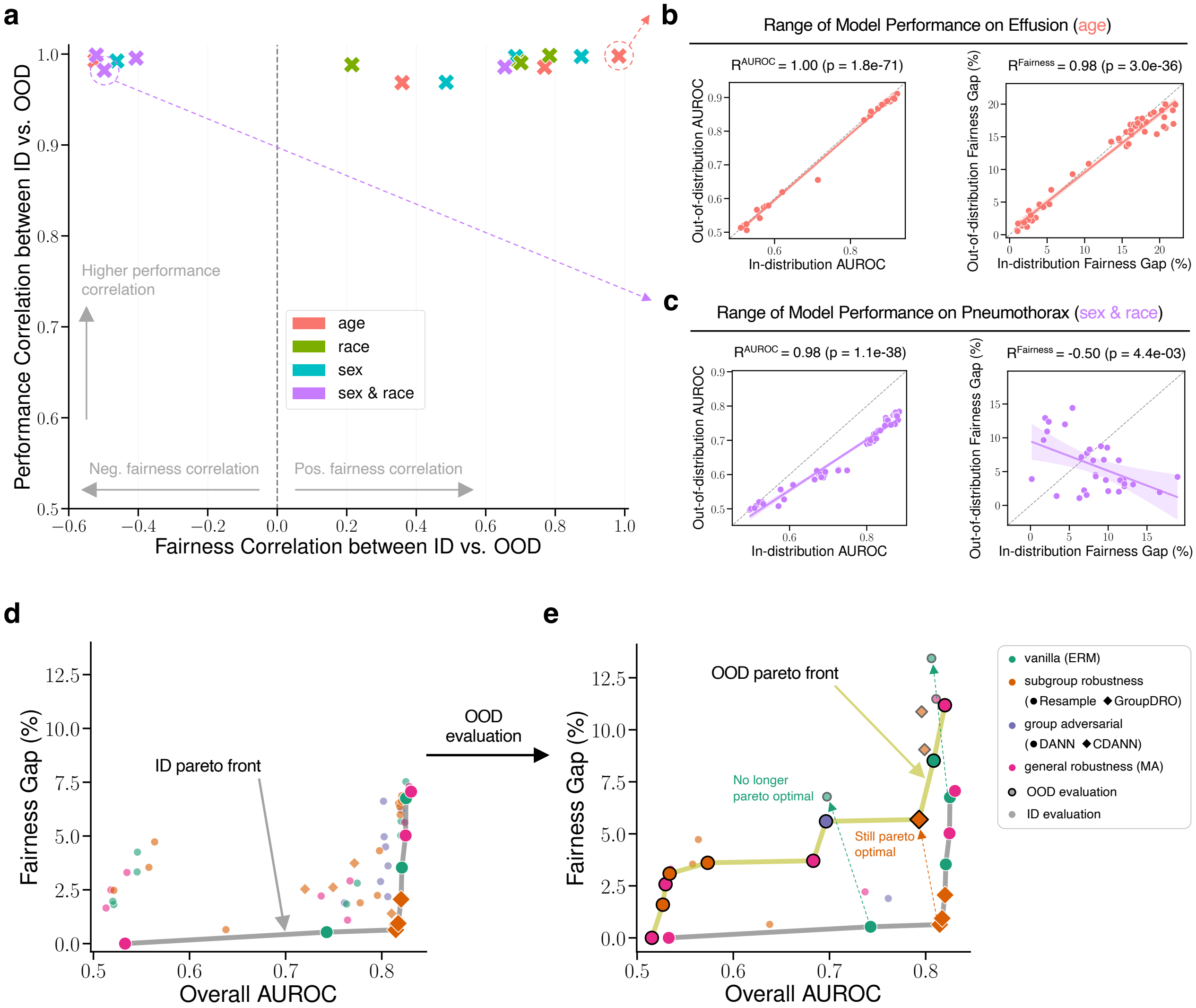

Figure B.4: We examine the transfer of performance (overall AUROC) and fairness between the ID (MIMIC-CXR) and OOD (all five other) datasets. **(a)** We plot the Pearson correlation coefficient of (ID vs. OOD) performance versus the Pearson correlation coefficient of (ID vs. OOD) fairness, where each point is a grid of trained models or a particular combination of task and attribute. We find that there is a high correlation between ID and OOD performance in all cases, but the correlation between ID and OOD fairness is tenuous. **(b), (c)** We show how two particular points in the first plot are obtained. **(d)**, **(e)** We show the transformation of the ID Pareto front to the OOD Pareto front, for *Cardiomegaly* prediction and using *race* as the attribute.