Detecting and Preventing Shortcut Learning for Fair Medical AI using Shortcut Testing (ShorT)

Alexander Brown¹, Nenad Tomasev², Jan Freyberg¹, Yuan Liu³, Alan Karthikesalingam¹, Jessica Schrouff¹*

ABSTRACT

Background
Machine learning (ML) holds great promise for improving healthcare, but it is critical to ensure that its use will not propagate or amplify health disparities. An important step is to characterize the (un)fairness of ML models - their tendency to perform differently across subgroups of the population - and to understand its underlying mechanisms. One potential driver of algorithmic unfairness, shortcut learning, arises when ML models base predictions on improper correlations in the training data. However, diagnosing this phenomenon is difficult, especially when sensitive attributes are causally linked with disease.

Methods
Using multi-task learning, we propose the first method to assess and mitigate shortcut learning as a part of the fairness assessment of clinical ML systems, and demonstrate its application to clinical tasks in radiology and dermatology.

Findings
We show that shortcut learning can be identified by varying the degree to which models may encode the sensitive attribute. We propose and demonstrate the utility of two types of mitigation strategies and their impact on model fairness and clinical prediction performance. Finally, our approach reveals instances when shortcutting is not responsible for unfairness, highlighting the need for a holistic approach to fairness mitigation in medical AI.

Interpretation
Identifying shortcut learning in ML models is difficult when sensitive attributes are causally linked to the clinical task. This paper is the first to propose a practical method for practitioners to assess and mitigate shortcut learning as a part of the routine fairness assessment of clinical AI systems.

Funding
This study was funded by Google.

¹Google Research, London, United Kingdom
²DeepMind, London, United Kingdom
³Google Research, Mountain View, CA, USA
*Corresponding author: schrouff@google.com
INTRODUCTION

Machine learning (ML) promises to be a powerful approach in many healthcare settings, with models being designed for a variety of diagnostic and prognostic tasks. A risk of harm from machine learning models is unfairness, as variation in model behavior for patients with different sensitive attributes (Figure 1a) has the potential to perpetuate or amplify existing health inequities. This has been observed in multiple clinical settings and remains a major topic of research. While the definition of what constitutes fairness may vary widely across fields, here fairness is defined as equal model performance across patient subgroups defined by sensitive attributes.

![Figure 1: Examples of correct and incorrect predictions influenced by shortcut learning.](image)

On the other hand, machine learning models may utilize information about sensitive attributes (such as age, sex, or race) to improve model performance in ways that may be justifiable where attributes correlate with disease risk in the deployment population. For instance, androgenetic alopecia is more prevalent in men and breast cancer more common in women; keloid scarring is more common in skin of color and melanoma more common in lighter skin tones. In such settings, ignoring or ablating attribute information may decrease clinical performance. However, the use of information about sensitive attributes can also be harmful - in particular due to the phenomenon of “shortcut learning.” This refers to ML models relying on spurious associations in training datasets to learn prediction rules which generalize poorly, particularly to new populations or new settings. Shortcut-based decision rules are also likely to amplify errors in ‘atypical’ examples, such as male patients with breast cancer, or melanoma in dark-skinned individuals. Reliance on shortcuts in model predictions is therefore likely to
be harmful, especially so when based upon sensitive attributes, as it risks exacerbating model unfairness, and furthering health disparities.

Concerns about ML models exploiting shortcuts based on sensitive attributes have been amplified by the observation that ML models can predict these attributes from clinical data, without the need for attributes to be directly inputted to the model. For example, models can be trained to predict sex or age from medical images \(^{11,12}\), and may even encode information about sensitive attributes when this was not the objective of ML training \(^{13}\).

However, the fact that a ML model encodes information about sensitive attributes does not necessarily mean that it uses this information to make clinical predictions \(^{14}\), or that such use results in shortcut learning. By intervening on the degree to which a model can encode a sensitive attribute, we demonstrate the first method to assess whether such encoding indicates the presence of shortcut learning, or is artefactual. We focus on age as a sensitive attribute, since aging is linked to disease risk across a wide variety of medical conditions.

The main contribution of this study is a novel approach that represents the first practically applicable framework for studying and mitigating shortcut learning in clinical ML models. This addresses an unmet need among practitioners when trying to develop fair and safe clinical AI.

**METHODS**

To identify shortcut learning, and how it relates to the encoding of sensitive attributes by the model, we define multiple quantities: (1) the encoding of the sensitive attribute, (2) fairness metrics, and (3) shortcut testing (ShorT), i.e. the correlation between the encoding of the sensitive attribute and fairness metrics. We demonstrate our proposed approach using binary prediction tasks in an open-source chest x-ray (CXR) dataset. The approach is then applied to a multiclass diagnosis task in dermatology.

**Datasets, tasks and models**

For CXR experiments, models were trained using the NIH Chest X Ray dataset \(^{15}\) to predict a single binary condition label (Figure 2a ‘Condition Prediction’). We investigated the “effusion”, “atelectasis” and “abnormal” findings. Model performance was estimated by computing the Area Under the Receiver Operator Curve (AUROC). The model included a “feature extractor” according to a deep learning architecture, followed by one or multiple “heads” that predicted outcomes of interest (see Supplement for details). We performed five repetitions of model training to assess result variability.

**Assessing the encoding of demographic information**

Similarly to previous work \(^{13,14}\), we assessed the encoding of demographic information in the penultimate layer of the model by “transferring” the condition model to predict age. Once the condition model was trained (Figure 2a, ‘Clinical prediction’), we froze all weights in the feature extractor and trained a linear predictor for age (Figure 2a, ‘Attribute Transfer’). The performance of the transfer model was expressed as the Mean Absolute Error (MAE); this value was used as a measure of
the age information content of the final layer of the feature extractor, with lower values indicating a more accurate age prediction and hence more age encoding. Our results were unchanged when using MSE.

To contextualize the obtained MAE, we trained models to directly predict age from scratch (Figure 2a, ‘Demographic Prediction’). This provides an empirical lower error bound (LEB) for age for this dataset and model architecture. We estimated an empirical upper error bound (UEB) for age by calculating the error obtained as a result of predicting the mean age of patients in the training set for all examples in the test set (i.e. the baseline performance using the distribution of ages alone, without any image information, in this dataset).

**Assessing fairness for age as a continuous variable**

To assess the fairness of a model’s predictions according to age, we refer to the definitions in 7 and applied group fairness metrics, quantifying independence, separation and sufficiency. Similarly to 16, we expressed fairness as a function of a continuous attribute variable to avoid the need for quantizing the data. This was achieved by fitting a univariate logistic regression model on age against the label (Figure 2b). Since we do not wish to assume that the clinical task is independent of age, we focus on “separation” - the discrepancy of error rates across subgroups.

Separation was defined by fitting two LR models to the binarised model predictions, one for patients who do have the condition, and one for patients who do not, equivalent to modeling the effect of age on the True Positive Rate (TPR, sensitivity) and the False Positive Rate (FPR, 1 - specificity). Separation was then calculated as the mean of the absolute values of the two logistic regression coefficients. This resulted in a metric where a score of 0 indicates a perfectly fair model, and higher scores indicate that the TPR and/or FPR vary systematically with age. For small values of separation, our definition
approximates the fractional change in performance per year of life - while numerically small, the resultant discrepancy may be large over a clinically relevant age difference. This can be calculated as $e^{s\Delta a}$, where $s$ is the separation coefficient and $\Delta a$ is the age difference. A separation value of 0.01 will thus correspond to a 10.5% change in model performance per decade; 0.02 will correspond to a 22.1% change per decade. Other equivalent formulations could be considered, such as the odds ratio of the logistic regression model. For completeness, we report independence and sufficiency metrics in the Supplement.

ShorT: testing for shortcut learning

Our approach relies on the hypothesis that if the model is shortcutting, intervening on its encoding of the sensitive attribute should affect fairness metrics. We used multitask learning to alter the degree to which age is encoded in the penultimate layer of the condition prediction model (Figure 2a, ‘Multitask Prediction’). We trained models on both demographic information (here, age) and condition prediction tasks by using a common feature extractor with a separate head for each task. Varying the amount of age information encoded in this model was achieved by scaling the gradient updates from the age prediction head. Positive values of gradient scaling encourage the model to represent age more strongly in the final layer, whereas negative values decrease this representation by gradient reversal.

We measured the effect of varying the age information encoded in the model by computing model performance, age availability (MAE of the age transfer model, as described above) and fairness metrics for each value of gradient scaling. The presence of shortcut learning is then indicated by a significant relationship between age encoding and fairness metrics. Given our choice of MAE and our formulation of separation (lower is better), we expect a negative correlation (computed via Spearman correlation coefficient) to highlight shortcut learning.

Assessing the efficacy of ShorT to detect shortcut learning

To assess how our method behaves under different bias scenarios, we alter the correlation between the sensitive attribute and the condition label in the CXR training dataset. This was achieved by randomly subsampling the training set, with a probability determined as a function of the patient’s age and condition label (see Supplement). After resampling, we obtain two datasets: a “biased” dataset that introduces approximately a decade of age difference between the positive and negative classes, and a “balanced” dataset where the distributions of ages across classes are approximately matched. We expect the biased dataset to lead to strong shortcut learning, while the balanced dataset should lead to no shortcut learning.

RESULTS

Prediction models encode age and are unfair

For all binary CXR tasks, we trained condition and transfer models (Figure S1). Transfer models were able to predict age (Figure S1b; Effusion 11.9 ± 0.47 years; Atelectasis 11.3 ± 0.28; Abnormal 11.4 ± 0.44, age MAE on held-out test set) significantly better than chance (permutation test, p<0.0001 for all
models). They performed better than the upper error bound (UEB) of 13.6 years, but worse than the lower error bound (LEB) of 6.4 ± 0.23 years.

We then estimated algorithmic fairness as defined by our continuous version of separation. All tasks produced models with a bias in performance according to age based on separation (Figure S1c). The observed separation values, in the range of 0.01-0.02, correspond to around an 11-22% performance difference per decade of life, a discrepancy that we feel is likely to be unacceptable to users in the absence of other considerations.

Our findings demonstrate that CXR models do learn to encode age, despite not being trained to do so. In addition, the performance of the models varies systematically with age, exhibiting unfairness. However, it is not possible to infer from these observations alone that the encoding of age is a driver of age-related unfair performance - which would be required for shortcut learning.

**Intervening on attribute encoding affects fairness metrics**

![Figure 3: Intervening on age encoding using multitask learning. (a) The distribution of ages for positive (light green) and negative (gray) examples for Effusion in the training set. (b) The effect of altering the gradient scaling of the age prediction head on age encoding (as determined by subsequent transfer learning). For large positive values of gradient scaling (left), the models encoded age strongly, with a low MAE that approached the performance of a dedicated age prediction model (empirical LEB). For large negative values of gradient scaling (right), the age prediction performance of the multitask model approached the empirical UEB. Baseline models (with zero gradient scaling from the age head, equivalent to a single task condition prediction model) are shown in red. For each of 25 values of gradient scaling, the mean MAE of 5 replicates is shown, with error bars denoting 95% confidence intervals from bootstrapping examples. (c) Fairness and performance of all replicates. The degree of age encoding by the particular replicate is color-coded, with purple dots denoting more age information, and green dots less information, than the baseline model without gradient scaling (in red).](image)

In order to test the degree to which such encoding may drive unfairness via shortcut learning, we varied the scaling of the gradient updates from the age head in a multitask learning paradigm (ShorT, see Methods). We focus here on results for the Effusion label (Figure 3a); similar results were obtained for Atelectasis and Abnormal (see Supplement).
We were able to vary the amount of age information encoded in the feature representation across a wide range of MAE values, covering the region between the upper and lower bounds on age prediction error (Figure 3b). By plotting the fairness of the resulting models against performance, we compared the impact of altered age encoding on these critical properties (Figure 3c). The ideal model would have high performance and low separation, in the top left corner of the scatter plot.

In this case, we found that increasing age encoding relative to baseline (purple dots in Figure 3c) did not noticeably affect model performance. Reducing the age encoding (green dots) appeared to slightly improve fairness properties, but at the cost of reducing overall model performance. We quantitatively analyze this effect below.

**ShorT efficiently detects shortcut learning**

![Figure 4: Effect of dataset perturbation. Results are presented in the same manner as for figure 3, but for a subsampled dataset inducing a larger age disparity between classes ('biased' dataset, a-c), and a 'balanced' dataset (d-f).](image)

Using a “biased” dataset, we created a preferential ground for shortcut learning by introducing a gap of 11.2 years (Figure 4a) between classes. When trained on this perturbed dataset, condition models encoded age more strongly (9.18 years age MAE for models trained on biased dataset, vs 11.8 years for models trained on full dataset); however altering the gradient scaling of the age head in multitask models still resulted in a wide range of age encoding strengths (Figure 4b). Clinical task performance was similar, albeit slightly higher in the biased dataset (mean AUC 0.901 vs 0.882 in the original dataset). This is expected, since the separation by attribute creates further information that can be used to make more accurate predictions. However, the fairness of the models was degraded strikingly (Figure 4c). This disparity could be obviated to some degree by gradient reversal for age - with
separation approximately halved in models with poorer age representation, with only a slight decrease in overall model performance (green cluster).

On the other hand, removing age differences due to prevalence (the “balanced” dataset, Figure 4d-f) resulted in models that performed at a similar level to the baseline model (mean AUC 0.883 in the balanced dataset, 0.882 original) and were fairer than models trained on the original dataset. We found similar results for Atelectasis and the Abnormal labels (Supplement).

When quantifying the dependence of unfairness on age encoding using the original dataset, we found a small but statistically significant correlation ($p=-0.224, p=0.0156$, Figure 5a), indicating the presence of shortcut learning. This was amplified in models trained on the biased dataset ($p=-0.668, p=8.11e-17$, Figure 5b), indicating the presence of significantly stronger shortcutting. Conversely, the correlation coefficient in the balanced dataset was not significant (Figure 5c, $p=0.116, p=0.218$), indicating no systematic impact of age encoding on fairness in models trained on this dataset (for details, see Supplement). Based on these results, our approach seems to efficiently detect shortcut learning.

**Shortcut learning cannot be identified by attribute encoding alone**

On the other hand, we found that the amount of age information encoded in the model bears little relation to the fairness of the model when comparing datasets (Figure 5, compared directly in Supplement). For an age MAE of 8.5-9.5 years, models trained on a balanced dataset were almost
perfectly fair, with an average separation coefficient of 0.0016 (range 0.0003-0.0035, n=24), corresponding to an average 1.6% disparity in performance over a decade of life. In contrast, models trained on a biased dataset had a mean separation coefficient of 0.0384, (range 0.0335-0.0461, n=43), corresponding to an average 47% disparity over a decade of life. Thus, in this case, it is clear that the performance of an attribute transfer model alone is insufficient to make any predictions regarding the fairness of the model. Rather, testing directly for the effect of encoding on fairness, reveals the presence of shortcut learning.

Beyond shortcut learning: Acne prediction in a Dermatology model

Figure 6: Unfair model performance not resulting from shortcut learning in a dermatology classifier - despite strong attribute-condition correlation. (a) Age distribution for examples with Acne (light green) and all other conditions (gray) in the training set. Note the significantly lower mean age for patients with Acne, as would be expected clinically. (b) AUC vs separation plot for Acne. AUC is binarised by using the score for Acne vs all; separation is calculated on a binarised prediction (top1). (c) ShorT analysis does not demonstrate that separation is significantly correlated with age encoding in this example.

Lastly, we applied our approach to a multiclass prediction model in dermatology (see Supplement) similar to that published in 8, focusing on the binarized Acne label, which is strongly correlated with age (18.6 year difference in mean age between patients with Acne vs other conditions; Figure 6a). The multiclass model encoded age strongly, with a mean age MAE of 9.58 years, compared to an LEB of 7.32, and UEB of 13.29 years (Supplement). Condition models showed unfair predictions of Acne across ages (separation range 0.0576-0.755, Figure 6b, red dots). This corresponds to a separation differential of up to 53% per decade.

However, despite the bias in the training set, strong age encoding in the model, and unfair performance, we found that varying the amount of age encoding did not result in a systematic change in fairness properties (Figure 6c, \(\rho=-0.177\), \(\rho=0.0746\)). Although the models are considerably less sensitive for Acne for older patients, the cause for this does not appear to be shortcut learning. There are a variety of other mechanisms which can lead to unfair performance, discussed below. However, multitask learning may still prove valuable in identifying models with high performance and better fairness properties.
DISCUSSION

Shortcut learning poses significant challenges for machine learning in healthcare, where predictions based on spurious correlations raise concerns regarding safety and fairness. However, identifying whether shortcut learning is responsible for model unfairness is difficult when sensitive attributes may be causally linked to the clinical task. In this paper, we propose the first practical method to directly test for the presence of shortcut learning during routine fairness assessment of clinical AI systems.

Before applying our method, it is critical to select an appropriate fairness criterion. The choice of metric will depend upon the particular clinical task, and wider societal context. Selection of an appropriate metric requires a deep understanding of how bias and inequity may be present in clinical environments and datasets. We recommend consultation with subject experts, patient groups, and literature review to identify plausible links between sensitive attributes and the clinical prediction target. This is likely to be improved by participatory problem formulation, but remains an open problem in the field. Although demonstrated in the context of separation, our framework is equally applicable to other fairness metrics.

We first replicated previous results, and demonstrated that clinical AI models were able to encode information about age, despite not being explicitly trained to do so. Therefore, providing the model with information about the age of a patient (as an auxiliary input) is not required for encoding to occur.

Such encoding is often assumed to demonstrate that the model has learnt to represent attributes so as to use them as shortcuts for predictions. However, whilst the presence of encoding is necessary for shortcut learning to occur, it does not provide conclusive evidence that models are basing diagnostic decisions on the encoded information using shortcut learning. Our results demonstrate that the degree to which models encode sensitive information was not predictive of the fairness of the models, in either the CXR or dermatology tasks. To our knowledge, this has not previously been empirically demonstrated.

We therefore developed a method to directly test whether unfair performance is driven by the encoding of the sensitive attribute. We varied the strength of this encoding using an additional demographic prediction head with variable gradient scaling. Models whose predictions did rely on the sensitive attribute should be systematically affected by changes in the degree of sensitive attribute encoding; models in which the encoding is incidental should not be affected in the same manner. We note that even if shortcut learning is not detected, incidental encoding may present intrinsic ethical concerns and potential for misuse.

Our proposed mitigation approaches (subsampling, gradient reversal) eliminate correlations between sensitive attributes and outcomes, or mitigate their effect on model training. This may seem counterintuitive, particularly where sensitive attributes are thought to be causal drivers of disease. However, our framework allows practitioners to identify when such mitigation is desirable by analyzing consequences on the trade-off between model performance and fairness.

Shortcut learning is usually assumed to involve the encoding of, and reliance upon, ‘spurious’ or non-causal correlations. As we show, this need not be the case. Shortcut learning may also occur in
cases where a sensitive attribute is strongly linked to the clinical task via known, biologically plausible mechanisms. Models may use attribute-clinical correlations appropriately, but these correlations do not generalize to deployment environments. It is also possible that models over-weight the importance of the attribute at the expense of clinical evidence that is more directly predictive, resulting in ‘stereotyping’. We chose to focus our study on age for two reasons. Firstly, age is known to be strongly linked to disease risk across a variety of conditions. Secondly, age is grounded as an objective attribute, rather than being socially constructed\(^{22}\). These two considerations suggest that age information may be useful in a disease prediction task, making disentangling shortcut learning from appropriate use of input features a more difficult task. For attributes not known to be linked to disease risk, or where attributes are considered to be social constructs, any use of the attribute in the prediction task may represent a shortcut.

Where shortcutting does occur, multiple approaches can be used to mitigate its effects. In our work, we find that gradient reversal can ameliorate, but not obviate, the effect of a biased dataset. Balancing the dataset may be an effective strategy, where feasible; we find that balancing the data leads to fair model performance at no cost to overall performance for CXR. Other approaches \(^{24,25}\) could also be considered.

When applied to a dermatology application, our framework did not identify a pattern of shortcut learning, despite unequal model performance by age. This demonstrates a case in which model performance is unfair; the label is strongly correlated with the sensitive attribute, and the sensitive attribute is encoded by the model. However, the encoding does not appear to be the (main) source of the unfair performance. In this case, unfairness might be caused by other factors, such as:

- Different presentations of the same condition. For instance, the typical pattern of hair loss for females with androgenetic alopecia differs to that of males \(^{26}\). Where presentations differ, unfairness can occur due to an inadequate sample size for specific subgroups, or if the appearance of the condition is more “difficult” to identify for some groups. Potential solutions include obtaining more examples of these presentations, upweighting losses for difficult examples, or approaches such as focal loss \(^{27}\). There is also evidence that longer training times encourage the learning of more difficult examples \(^{28}\).
- Differences in the quality of the label, or the use of proxy labels to approximate underlying disease. This is extensively discussed in \(^{3,29}\), with potential mitigations.
- Differences in the quality or missingness patterns of the data. Multiple causes of unfairness in this regard are described in \(^{1}\).

Lastly, shortcut learning, where present, does not guarantee the absence of other sources of unfair performance.

Our work considered age as a single attribute of interest. In principle, this method may be readily applicable to an intersectional analysis \(^{30}\), although practically there may be challenges around model convergence. Finally, algorithmic fairness is a set of mathematical formulations, and model behavior should be considered in the broader context of health equity, the entire clinical system and its interaction with society, rather than just focussing on model behavior given a defined dataset. We
however believe that the identification and mitigation of shortcut learning, as demonstrated by our approach, paves the way for more fair medical AI.

ACKNOWLEDGEMENTS
We thank Alexander D’Amour, Oluwasanmi Koyejo, Stephen Pfohl, Katherine Heller, Patricia MacWilliams, Abhijit Roy, Vivek Natarajan, Charles Lau, Jon Deaton, Dale Webster, William Isaac, Shakir Mohamed, Danielle Belgrave, Greg Corrado and Marian Croak for their contributions to this work. We also thank Lisa Lehmann and Ivor Horn for helpful discussions on the ethical implications of this work.

AUTHOR CONTRIBUTION
A.B., N.T. and J.S. identified the research question and designed experiments. A.B., J.F. and J.S. performed experiments. Y.L. advised on the dermatology application and the model architecture. All authors contributed to the interpretation of the results. A.B., J.S., N.T. and A.K. wrote the manuscript.

DATA SHARING
The NIH CXR Dataset is provided by the NIH Clinical Center and is available at https://nihcc.app.box.com/v/ChestXray-NIHCC. The dermatology data is not available to the public. The code is proprietary.
REFERENCES

1. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann. Intern. Med.* 169, 866–872 (2018).

2. Leslie, D., Mazumder, A., Peppin, A., Wolters, M. K. & Hagerty, A. Does ‘AI’ stand for augmenting inequality in the era of covid-19 healthcare? *BMJ* 372, n304 (2021).

3. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 447–453 (2019).

4. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* 27, 2176–2182 (2021).

5. Samorani, M., Harris, S. L., Blount, L. G., Lu, H. & Santoro, M. A. Overbooked and Overlooked: Machine Learning and Racial Bias in Medical Appointment Scheduling. *M&SOM* (2021) doi:10.1287/msom.2021.0999.

6. Mulligan, D. K., Kroll, J. A., Kohli, N. & Wong, R. Y. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proc. ACM Hum.-Comput. Interact.* 3, 1–36 (2019).

7. Barocas, S., Hardt, M. & Narayanan, A. *Fairness and machine learning*. (fairmlbook.org, 2019).

8. Liu, Y. *et al.* A deep learning system for differential diagnosis of skin diseases. *Nat. Med.* 26, 900–908 (2020).

9. Huang, C., Wu, Z., Du, Y. & Ogawa, R. The Epidemiology of Keloids. in *Textbook on Scar Management: State of the Art Management and Emerging Technologies* (eds. Téot, L., Mustoe, T. A., Middelkoop, E. & Gauglitz, G. G.) 29–35 (Springer International Publishing, 2020).

10. Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nature Machine Intelligence* vol. 2 665–673 (2020).

11. Poplin, R. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* 2, 158–164 (2018).

12. Jabbour, S., Fouhey, D., Kazerooni, E., Sjöding, M. W. & Wiens, J. Deep Learning Applied to Chest
X-Rays: Exploiting and Preventing Shortcuts. in Proceedings of the 5th Machine Learning for Healthcare Conference (eds. Doshi-Velez, F. et al.) vol. 126 750–782 (PMLR, 2020).

13. Banerjee, I. et al. Reading Race: AI Recognises Patient’s Racial Identity In Medical Images. arXiv [cs.CV] (2021).

14. Glocker, B., Jones, C., Bernhardt, M. & Winzeck, S. Algorithmic encoding of protected characteristics in image-based models for disease detection. arXiv [cs.LG] (2021).

15. Wang, X. et al. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. IEEE CVPR 2097–2106 (2017).

16. Estiri, H. et al. An objective search for unrecognized bias in validated COVID-19 prediction models. bioRxiv (2021) doi:10.1101/2021.10.28.21265629.

17. Ganin, Y. & Lempitsky, V. Unsupervised Domain Adaptation by Backpropagation. arXiv [stat.ML] (2014).

18. Raff, E. & Sylvester, J. Gradient Reversal Against Discrimination. arXiv [stat.ML] (2018).

19. Makar, M. et al. Causally-motivated Shortcut Removal Using Auxiliary Labels. arXiv [cs.LG] (2021).

20. Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T. & Prabhakaran, V. Re-imagining Algorithmic Fairness in India and Beyond. arXiv [cs.CY] (2021).

21. Martin, D., Jr, Prabhakaran, V., Kuhlberg, J., Smart, A. & Isaac, W. S. Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics. arXiv [cs.CY] (2020).

22. Krieger, N. Who and what is a ‘population’? Historical debates, current controversies, and implications for understanding ‘population health’ and rectifying health inequities. Milbank Q. 90, 634–681 (2012).

23. Kärkkäinen, K. & Joo, J. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. in 2021 IEEE Winter Conference on Applications of Computer Vision (WACV) 1547–1557 (2021).
24. Minderer, M., Bachem, O., Houlsby, N. & Tschannen, M. Automatic Shortcut Removal for Self-Supervised Representation Learning. in *Proceedings of the 37th International Conference on Machine Learning* (eds. Iii, H. D. & Singh, A.) vol. 119 6927–6937 (PMLR, 2020).

25. Du, M. et al. Towards Interpreting and Mitigating Shortcut Learning Behavior of NLU models. in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 915–929 (Association for Computational Linguistics, 2021).

26. Herskovitz, I. & Tosti, A. Female pattern hair loss. *Int. J. Endocrinol. Metab.* 11, e9860 (2013).

27. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal Loss for Dense Object Detection. *arXiv [cs.CV]* (2017).

28. Agarwal, C., D’souza, D. & Hooker, S. Estimating Example Difficulty Using Variance of Gradients. *arXiv [cs.CV]* (2020).

29. Obermeyer, Z. et al. Algorithmic Bias Playbook.
   https://www.ftc.gov/system/files/documents/public_events/1582978/algorimic-bias-playbook.pdf (2021).

30. Buolamwini, J. & Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *FAT* (2018).

31. Kolesnikov, A. et al. Big Transfer (BiT): General Visual Representation Learning. *arXiv [cs.CV]* (2019).

32. Blum, J. R., Kiefer, J. & Rosenblatt, M. Distribution Free Tests of Independence Based on the Sample Distribution Function. *Ann. Math. Stat.* 32, 485–498 (1961).

33. Gretton, A., Bousquet, O., Smola, A. & Scholkopf, B. Measuring Statistical Dependence with Hilbert–Schmidt Norms. *Plan. Perspect.* 63, 77 (2005).

34. Sricharan, K., Raich, R. & Hero, A. O., III. Empirical estimation of entropy functionals with confidence. *arXiv [math.ST]* (2010).

35. Miller, A. C., Gatys, L. A., Futoma, J. & Fox, E. B. Model-based metrics: Sample-efficient estimates
of predictive model subpopulation performance. arXiv [stat.ML] (2021).
Figure S1: Binary CXR prediction models. (a) Performance in terms of AUC for each task (Effusion, Atelectasis and Abnormal). (b) Age information encoded in each model in terms of age MAE. UEB: Upper Error Bound, LEB: Lower Error Bound, as determined experimentally. These bounds represent the limits for model age error in this dataset. (c) Fairness of each model in terms of separation, with 0 meaning a perfectly fair model (a separation value of 0.01 will correspond to a 10.5% change in model performance per decade; 0.02 will correspond to a 22.1% change per decade, see Methods). In all cases, each dot represents a different replicate of the model, with the average metric represented by a horizontal line.
Supplementary Figure S2

Figure S2: Fairness-Performance results for other CXR labels. Separation is plotted against AUC, with the age performance of each model represented by the color as in Figure 3. Similar patterns may be observed, whereby inducing a bias in the training dataset results in much more unfair model performance, which can be ameliorated by gradient reversal (center column, green dots), or exacerbated by increasing the age representation (purple dots). In contrast, balancing the training dataset results in baseline models which are considerably fairer, and gradient reversal results in degraded model performance without further fairness improvement.
Figure S3: ShorT analysis of original NIH and subsampled datasets for Atelectasis and Abnormal (the complement of the No Finding label). Biased datasets (middle column) cause significant dependence of fairness on age representation. In contrast, balanced datasets (right column), there is no such dependence. In the original dataset, there is no dependence of fairness on age representation for Atelectasis, however there is a significant positive correlation between fairness and age representation for the No Finding label. This implies that models which represent age more accurately (left) tend to be fairer (closer to 0 on the y axis). This may be explained by an underuse of age information for this particular dataset and task. For all plots, an AUC threshold was set at 0.7, with replicates with an AUC value less than this being excluded from the correlation analysis. We chose 0.7 as the threshold as the performance of baseline models was lower for Atelectasis and Abnormal labels (figure S1). One such replicate is not displayed on the Abnormal, Original Dataset plot, as it had a separation of > 0.1, and lies beyond the limits of the y axis.
Supplementary Figure S4

Figure S4: Cross-Dataset Age comparison of fairness and performance at differing levels of age encoding in the Effusion task. (a) Fairness. Results from Figure 5 are displayed on a single graph, with replicates pooled according to degree of age encoding. For a given level of age encoding, models trained on the Balanced (solid line), Original (dashed line), or Biased (dotted line) datasets display vastly different fairness characteristics. (b) Age encoding vs Performance. At the same level of age encoding, performance is very similar for the Balanced and Original datasets, although the performance of the Balanced dataset drops off at higher age prediction errors. The Biased dataset results in a spuriously higher AUC due to cleaner class separation (see Figure S5).
Supplementary Figure S5

Figure S5: Cross-dataset performance and fairness for the effusion prediction task. AUC and Separation are shown for baseline models (without an age prediction head) trained on biased, original, and balanced datasets (x axis), tested on all three datasets. In-distribution results are located on the top-left to bottom-right diagonal. Note that the best performance is obtained in models trained on biased datasets, tested in-distribution; however, performance is degraded for out of distribution test sets, due to shortcut learning; this increase in performance is therefore spurious. Models trained on balanced datasets obtain similar performance results to those trained on the original dataset. However, separation is considerably improved in models trained on balanced data.
Figure S6: Effect of age head gradient scaling on age representation for the dermatology example in figure 6. Short models covered a range of age prediction errors, although there appeared to be a wider plateau in the middle of the range of age head scaling values, over which age prediction error was quite similar to baseline. This plateau, as well as the wide range between the “Clinical Label” and “No Information” upper error bounds, likely occurs due to the richer soft labels used in this example, as well as due to the stronger dependence between age and condition probability for many (but not all) dermatological problems.
Supplementary Methods

CXR Dataset

The NIH CXR Dataset is provided by the NIH Clinical Center and is available at https://nihcc.app.box.com/v/ChestXray-NIHCC. For CXR experiments, images were first downsized to 448x448 pixels. We select the “Effusion”, “Atelectasis” and “No findings” (which we report as “Abnormal” for semantic consistency) labels provided with data as our binary outcomes, focusing on Effusion.

Model architecture

The feature extractor was a ResNet 101x3 architecture initialized from BiT checkpoints 31. We used an Adam optimizer with a constant learning rate and optimal hyperparameters were determined for each class of model before training (see next section). Heads predicting binary outcomes (e.g. binary effusion label) were trained using a cross entropy loss, while heads predicting age used a mean squared error regression loss.

Hyperparameter tuning

All models were tuned for batch size, learning rate, weight decay, and dropout in the penultimate layer before training. The same parameters were applied to models trained on each label in the CXR task.

Table S1: Hyperparameters for CXR tasks.

|                      | Batch Size | Learning Rate | Weight Decay | Dropout |
|----------------------|------------|---------------|--------------|---------|
| Age Prediction       | 16         | $1 \times 10^{-5}$ | $1 \times 10^{-7}$ | 0       |
| CXR Prediction       | 16         | $4 \times 10^{-5}$ | $1 \times 10^{-6}$ | 0.1     |
| Age Transfer         | 8          | $3 \times 10^{-3}$ | n/a          | 0       |

After hyperparameters were identified, each model was trained from five different random seeds, with results presented in terms of average and standard deviation across seeds.

Multitask prediction

The demographic (age prediction) head in the multitask model uses two fully connected hidden layers between the feature extractor and the final age output layer, to provide the network with capacity during adversarial training. In order to approximately balance the losses between the age (mean square error) and condition (cross-entropy) heads, the regression loss is first down-weighted 100x. In all multitask experiments, 25 values were used for scaling of the gradient updates from the demographic head, ranging from -0.1 to +0.1. For each value of gradient scaling, 5 replicates were trained, resulting in 125 models per experiment.
**Subsampling of training data**

We aim to provide datasets with varying degrees of correlation between age and the condition distribution. We however desired to keep the original images rather than creating synthetic (and potentially out of distribution) examples, and selected subsampling to perform this operation. To this end, we use a logistic probability function, which defines the probability of an example being retained as a function of the age of the patient:

$$ p_{\text{retain}} = \frac{m}{1 + e^{-k(a-a_0)}} $$

Where $k$ is the slope of the function; $a_0$ is the midpoint of the probability function (the age at which the probability of being retained is 0.5); and $m$ is a scale factor that increases the probability of retaining examples. This defines a probability of retaining a positive example; for negative examples (patients without the condition), we use $1-p_{\text{retain}}$.

Importantly, whilst this perturbation increased or eliminated the correlation between age and disease, there remain examples of older patients and younger patients with and without the condition, and all retained examples were not modified in any way. Whilst this resampling strategy leads to a slightly lower number of training examples, it has the advantage of maintaining the marginal probability of diseases, and avoids creating synthetic examples which may not be realistic.

Table S2 reports the parameters used to generate subsampled training sets. Since the process is stochastic, these were obtained by trial and error.

**Table S2: Subsampling parameters for CXR datasets.**

|        | $k$     | $a_0$   | $m$  |
|--------|---------|---------|------|
| Biased | Balanced|         |      |
| Effusion       | 0.14    | -0.07   | 50   | 4   |
| Atelectasis   | 0.12    | -0.08   | 50   | 4   |
| Abnormal      | 0.14    | -0.065  | 50   | 4   |

The training sets generated using these parameters are described in Table S3. These perturbed datasets match the desired shift in ages only approximately due to stochastic errors.

**Table S3: Characteristics of the subsampled datasets.**

|        | Number of training examples | Positive Examples (%) of training set | Mean Age of Positive / Negative classes (years) |
|--------|-----------------------------|---------------------------------------|------------------------------------------------|
|        | Original | Biased | Balanced | Original | Biased | Balanced | Original | Biased | Balanced | Original | Biased | Balanced |
| Effusion | 65394    | 55634  | 61029    | 6731     | 5612   | 6421     | 51.3     | 55.8   | 50.1     | 48.2     | 44.6   | 50.0     |
Significance testing when comparing ShorT across datasets

Shortcut testing (ShorT) relies on calculating the correlation between the degree of age encoding and fairness metrics. To test that the ShorT statistics differ across datasets, we perform permutation tests of Spearman's rho across different versions of the training dataset. We calculate the true difference in correlation statistics, and compare it to an empirical null distribution of differences. The null distribution is simulated using bootstrapping. We combine the data points from the two groups, shuffle them, and randomly divide them into two groups. To calculate p-values, we compare the true difference to this null distribution.

For CXR, we find that differences are highly significant when comparing the original and biased datasets, and the original and balanced datasets (p = 1e-8; p = 1e-4, respectively), indicating that shortcutting happens significantly more with biased datasets, and significantly less with a balanced dataset.

Dermatology dataset and experiments

For dermatology experiments, models are trained to predict 26 skin conditions with an additional “other” category to capture the long tail of conditions, as a multiclass prediction task, as described in 8. Our approach differs slightly from previously published results, as we use a more modern architecture (ResNet 101x3 rather than Inception v4), and a slightly smaller training dataset. The commercial dataset used consists of teledermatology images with associated diagnoses obtained by labeling by multiple dermatologists. Unfortunately, this dataset is not available for public use.

We assess model performance for a single class by using binarised metrics. For AUC, we use the prediction score of the chosen class. For separation, we define positive predictions to be examples where the top ranking prediction score is for the chosen class. Using top-3 selection (i.e. a positive prediction is any example where the score for the chosen class is in the top-3 scores) did not change our results.

Demographics in dermatology dataset

Since the dermatological dataset is not publicly available, we report in Table S4 the basic demographics of the training dataset used. This dataset comprises 12,027 cases obtained from teledermatology clinics in California and Hawaii.

Table S4: Demographic distribution of the dermatology dataset.
### Attribute

| Attribute                  | Percentage in training set |
|----------------------------|----------------------------|
| **Race**                   |                            |
| American Indian / Alaska Native | 0.83                     |
| Asian                      | 11.6                      |
| Black / African American   | 5.99                      |
| Hispanic / Latino          | 41.9                      |
| Native Hawaiian / Pacific Islander | 1.52               |
| White                      | 35.5                      |
| Not Specified              | 2.71                      |
| **Gender**                 |                            |
| Male                       | 38.1                      |
| Female                     | 61.9                      |
| **Age**                    |                            |
| 18-19                      | 7.33                      |
| 20-29                      | 22.4                      |
| 30-39                      | 19.2                      |
| 40-49                      | 16.8                      |
| 50-59                      | 19.5                      |
| 60-69                      | 11.6                      |
| 70-79                      | 2.27                      |
| 80-89                      | 0.782                     |
| 90+                        | 0.191                     |

**Supplementary discussion**

In our analysis, we have chosen to preserve age as a continuous variable, using logistic regression analysis to characterize the fairness properties of the model. This avoids the need for arbitrary quantization of the data. However, it does assume that discrepancies, where observed, will be monotonic - with weaker performance for either older or younger patients. In cases where we may expect bimodal or more complex distributions of fairness properties it might be more judicious to examine the model outputs rather than rely on particular formulations of fairness metrics. Distribution-free approaches \(^{32-34}\), may be considered if no particular form of association can be expected, although these will in general be more limited in power and interpretability. Secondly, the use of a LR model requires a binarised outcome per example, and would be unsuitable for metrics such as prediction scores (continuous) or AUC (requires a set of observations). Alternative methods \(^{16,35}\) may
overcome some of these limitations, at the expense of interpretability. However, our framework does not require the use of a continuous attribute, and may be applied to binary or discrete variables, by substituting the model-based fairness metrics for conventional definitions.