Algorithmic encoding of protected characteristics and its implications on performance disparities

Ben Glocker¹,*, Charles Jones¹, Mélanie Bernhardt¹, and Stefan Winzeck¹

¹ Department of Computing, Imperial College London, London, SW7 2AZ, UK
*Corresponding author: b.glocker@imperial.ac.uk

It has been rightfully emphasized that the use of AI for clinical decision making could amplify health disparities. A machine learning model may pick up undesirable correlations, for example, between a patient's racial identity and clinical outcome. Such correlations are often present in (historical) data used for model development. There has been an increase in studies reporting biases in disease detection models. Besides the scarcity of data from underserved populations, very little is known about how these biases are encoded and how one may reduce or even remove disparate performance. There are concerns that an algorithm may recognize patient characteristics such as biological sex or racial identity, and then directly or indirectly use this information when making predictions. But it remains unclear how we can establish whether such information is actually used. This article aims to shed some light on these issues by exploring methodology allowing intuitive inspections of the inner working of machine learning models for image-based detection of disease. We also investigate how to address performance disparities and find automatic threshold selection to be an effective yet questionable technique, resulting in models with comparable true and false positive rates across subgroups. Our findings call for further research to better understand the underlying causes of performance disparities.

Introduction

Knowing what type of information is being used when machine learning (ML) models make predictions is of high relevance to all stakeholders, including clinicians, patients, developers, regulators, and policy makers. Some of the best performing ML models, however, seem rather opaque regarding the mechanism by which these models map input data (e.g., medical scans) to output predictions (e.g., clinical diagnosis). During training, these advanced models, commonly based on deep neural networks, construct a rather complex feature representation by successively applying non-linear transformations to the raw input data (e.g., image pixel values). The final representations in the penultimate layer of the neural network are then processed by the last layer which acts as the prediction layer. The prediction layer learns to assign weights to the individual features in the penultimate layer which are then aggregated to generate an output prediction (e.g., probability for the presence of disease). It is important to realize that this very last layer in itself constitutes a sub-model (typically in the form of a linear or logistic regression) that makes the ultimate decision what information is used for the output prediction. We will come back to this point later. The basic components of a typical deep neural network model are illustrated in Fig. 1.

A key issue with these ‘black box’ models is that it is difficult to know what type of information is being used. The reason is that the generated feature representations encoded in the penultimate layer do not carry semantic information. So, we cannot easily check whether a particular type of information, say patient’s racial identity, is being used by the prediction layer, even if we know that such information is present in the original input data, either explicitly, e.g., in tabular data, or implicitly, e.g., in medical scans. An important point we will make is that even the presence of such information in the learned feature representation of the penultimate layer is not a sufficient indication that this information is being used by the prediction layer. We will illustrate this with some concrete examples and a real world, clinical application of disease detection in chest X-ray.

Despite the opaque behavior of deep neural networks, these models are now ubiquitous, and have become the state-of-the-
art approach for most image-based prediction tasks. One may argue that some of the most exciting applications are the ones where the associations between the ‘raw’ inputs and output predictions are entirely unknown. Here, the technique of deep learning, thanks to its ability to learn complex feature representations, may help us to not only discover such associations but also to capture them in compact models such that we can make accurate predictions on new data.

### The unreasonable performance of deep learning

An intriguing example for the ‘power’ of deep learning is the discovery that cardiovascular risk factors can be accurately predicted from retinal fundus images, including age, biological sex, smoking status, systolic blood pressure and major adverse cardiac events. Additional biomarkers were discovered shortly after. Another remarkable finding was that deep neural networks were capable of predicting patients’ experienced pain from knee X-rays enabling an algorithmic approach for reducing unexplained racial disparities in pain.

A recent study further demonstrated that deep neural networks are also capable of recognizing a patient’s racial identity from chest X-ray and other medical scans with astonishing accuracy, in addition to a patient’s age and biological sex. The ‘reading race’ study is remarkable in multiple ways. Not only was the medical profession apparently unaware that it is possible to recognize racial identity from these scans, it also appears to be a task that expert radiologists are not capable of doing (or at least not trained for). The exact mechanism and types of imaging features that are being used for making these predictions are yet to be uncovered. But more importantly, these results may have profound implications in the context discussed earlier, that there is a real risk that ML models may amplify health disparities. Since it seems straightforward (given the very high accuracy) to extract features related to racial information from medical scans, any spurious correlations between race and clinical outcome present in the data could be picked up by a model that is trained for clinical diagnosis. Assuming that features predictive of race are easier to extract than features associated with pathology, the model is very likely to learn so called ‘shortcuts’, that would manifest an undesirable association in the model between the patient’s race and the prediction of disease. This emphasizes again the importance for being able to know what information is being used when an ML model makes predictions.

The ‘reading race’ study made some attempts at this, despite this not being the main focus of the work. More concretely, the authors tried to establish whether the feature representations that are learned by a deep neural network when trained on other tasks such as disease detection, are useful for predicting race. To cite from the paper, the authors “hypothesized that if the [disease] model was able to identify a patient’s race, this would suggest the models had implicitly learned to recognize racial information despite not being directly trained for that task.” The study found this to be the case, using a specific type of test that we will discuss below, and further concluded “these results suggest that even when race is poorly correlated with the outcome of interest, as is the case in pathology detection, deep learning models are likely to learn unintended cues related to race and incorporate these cues in their decision making.”

This rather daunting conclusion demands further investigation, as its implications are of high relevance for the safe and ethical use of AI. In the following, we will lay out why we believe that the methodology used to draw the above conclusion is insufficient to establish whether unwanted information is being used for making predictions. We will then investigate a different approach for model inspection that provides valuable insights about the algorithmic encoding of patient characteristics and its implications on disparate performance across patient subgroups.

### Disparities in disease detection models

The basis of our investigation is a real world, clinical application of image-based detection of disease. We study the behavior of deep convolutional neural networks trained for detecting different conditions using a publicly available chest X-ray dataset. The dataset contains detailed patient demographics including self-reported racial identity, biological sex, and age. We note that developing the disease detection model itself is not the primary concern of our study, nor do we claim any contribution in this respect. Here, we are studying a model that has been used in previous works. More details are provided under Methods.

For studying the effects of algorithmic encoding of patient characteristics, we trained multiple models for disease detection using different parts of the training data. Our main model uses the whole training set including all patients that identify as White, Asian, or Black. We trained two other models each using only a subgroup of patients to contrast our findings on the encoding of racial identity and biological sex. To minimize the effect of different amounts of training data on the model performance, we used the subgroups with the largest number of scans available which is the group of patients that identified as White, and the group of male patients. The models trained on subgroups are not exposed to the same variation in patient characteristics as the model trained on all data.

We focus our analysis on two tasks, one for detecting the presence of a specific pathology (‘pleural effusion’ label) and another aiming to rule out the presence of any disease (‘no finding’ label). These two labels are mutually exclusive which makes them suitable for our model inspection, discussed later. They also have very different and varying prevalence across subgroups (cf. Table S1), which makes them interesting to study in the context of performance disparities. The ‘no finding’ label is also the focus of a recent study which reported subgroup disparities assumed to be associated with underdiagnosis bias.

While the absolute performance of our disease detection models was not a focus of our study, we observed good classification accuracy for the two conditions with average AUC=0.87 which is comparable to previous work. A breakdown of disease detection performance for different patient subgroups and the different models is given in Table S2 with corresponding ROC curves shown in Fig. 2. We also report true positive rates (TPR) and false positive rates (FPR) using one fixed global threshold per model and task optimized over the whole patient population with a target FPR of 0.2. While AUC seems largely consistent...
The inter-relationship of prediction tasks

If we want to better understand what causes these disparities, we need to investigate what information is being used by the disease detection models when making predictions. It would be important to know whether the disparate performance on subgroups stems from the fact that a model is directly or indirectly using information about protected patient characteristics, especially when these are poorly correlated with clinical outcome.

One seemingly intuitive approach to investigate whether information is being used is to check if the model (or more precisely, the learned feature representation) trained for a primary task of disease detection can be used for a secondary task for predicting patient characteristics. Assuming the secondary task can be performed reasonably well, one may conclude that the two tasks are closely related.

Before we discuss this assumption in more detail, we first need to clarify what we mean, from a machine learning perspective, when we say that two tasks are related. Here, one may distinguish between three scenarios illustrated with the example of separating colors and shapes, as shown in Fig. 3. In scenario A, the two tasks are unrelated both on the feature- and the output-level; In this case, we can solve each task independently using a different set of features and no information about the other task is relevant nor helpful. In scenario B, the tasks are related on a feature-level but not on an output-level; Here, the other task is relevant nor helpful. In scenario C, the tasks are related both on a feature- and output-level, and it appears impossible to disentangle the tasks. Solving one task will, at least to some degree, also solve the other task. We can say that the information related to each task is correlated with the other task. In practice, there can be of course different degrees of correlation.

For the objective of establishing whether certain information (say shape) is being used for making predictions for a primary task (say classification of color), we can see that in scenario A the feature representation from the primary task (here just a single feature $x_1$) is not useful for solving the secondary task. We thus may safely conclude that shape has no role to play when classifying color and vice versa on this given example.

Fig. 2 Disease detection performance. a-c ROC curves for the performance on the ‘no finding’ label for three different disease detection models trained on all patients, White patients, and male patients. d-f ROC curves for the performance on the ‘pleural effusion’ label for the same three disease detection models. While AUC is largely consistent and ROC curves are very similar across subgroups and models, we observe a clear shift in TPR/FPR performance when using a single threshold that was optimized over the whole patient population for target FPR=0.2.
whether a disease detection model may have implicitly learned information about racial identity\(^6\). Here, we argue that SPLIT is insufficient as a test to confirm this. Considering the example of shape and color classification, SPLIT can only tell us whether we are either in scenario A, in which case the accuracy obtained with SPLIT would have to be very low as the feature learned for one task is not useful for the other task (we call this a negative SPLIT result), or we are in one of the other two scenarios, B or C. However, SPLIT is unable to distinguish between those two. In fact, even the absolute value of the observed SPLIT accuracy is uninformative as we can easily confirm for the given example in scenario B. Here, SPLIT would result in perfect classification of shape, despite shape being irrelevant for the classification of color. We conclude that obtaining a positive SPLIT result is a necessary condition for establishing that information is being used, but not a sufficient one. SPLIT is like a diagnostic test that has 100% sensitivity but unknown specificity.

To confirm these shortcomings, we applied SPLIT for both race and sex classification from the disease detection models. We investigate different model backbones for producing feature representations on the chest X-ray data. In addition to ones obtained from our disease detection models, we also considered two backbones that were not trained for disease detection nor with any medical imaging data. One of them is based on random initialization of the network weights where the backbone then acts as a random projection of the input imaging data. This backbone is entirely untrained before applying SPLIT. The other backbone corresponds to a network trained for natural image classification using the ImageNet dataset\(^5\).

The results for applying SPLIT to race classification are summarized in Table 1. We find that SPLIT is unable to detect differences between the networks trained on all patients versus the model trained only on White patients (similar AUC). We also find similarly high SPLIT responses for the backbone trained on ImageNet, which indicates that it is possible to successfully train a prediction layer to recognize a patient’s racial identity from Chest X-ray using features tailored for ImageNet natural images. The training of the primary task of natural image classification, however, cannot have possibly learned to extract racial information as it has never seen any medical data, and yet, its features are useful for race classification when training a new prediction layer. In case of ImageNet, we are likely seeing an example of scenario B discussed earlier, where the same features

---

**Table 1 Results for applying SPLIT to race classification with four different neural network backbones.**

| Neural network backbone | White | Asian | Black |
|-------------------------|-------|-------|-------|
| Random weights          | 0.69  | 0.74  | 0.65  |
| ImageNet                | 0.76  | 0.79  | 0.77  |
| Disease detection (all patients) | 0.78  | 0.80  | 0.78  |
| Disease detection (white patients) | 0.77  | 0.79  | 0.79  |
| Race classification (all patients) | 0.96  | 0.97  | 0.97  |

AUC is determined in a one-vs-all approach for each racial group. For comparison, we also report the AUC in the last row for a neural network trained specifically for classifying race confirming the high accuracy observed in a recent study\(^6\).

---

**Fig. 3 Illustration of different inter-relationships of prediction tasks.**

a The two classification tasks of separating colors (blue vs orange) and shapes (crosses vs circles) are unrelated, both on the feature- and the output-level. The color classification can be performed by only considering feature \(x_1\) while shape information is irrelevant. Similarly, shapes can be classified using feature \(x_2\) with color being irrelevant. b The two tasks are related on a feature-level but not on their outputs. In both tasks, the features \(x_1\) and \(x_2\) need to be considered for classifying colors and shapes, however, shape information remains irrelevant for separating colors, and vice versa. While in both tasks the exact same features are being used, they are combined in different ways. c The two tasks are related both on a feature- and an output-level. Solving one of the tasks also solves the other. Shape and color information is highly correlated. The dashed green and gray lines indicate the optimal decision boundaries for color and shape classification, respectively.

In scenario C, we would find the opposite. The secondary task can be solved using the feature representation from the primary task (and vice versa). We may conclude that there is no way of disentangling shape and color information, and a model trained for one task may use the information related to the other task.

Scenario B is the most intriguing one. Here, we would find that the secondary task can be solved by using the features from the primary task. However, we need to learn a new set of feature weights specific to the secondary task, as the weights will be different than the ones learned for the primary task. So, while the information about one task is neither relevant nor helpful for the other task, we can still solve each task by using the features from the other task. This is an important observation which is relevant in the context of our real world application of disease detection, as discussed below.

**The supervised prediction layer information test**

The approach for testing whether we can solve one task using the features learned from another task can be easily implemented for deep neural networks. Using the notion of transfer learning\(^4\), we can first train a neural network for a primary task, ‘freeze’ the network’s parameters (i.e., its weights in what is often called the neural network ‘backbone’) and then replace the prediction layer with a new one (cf. Fig. 1). The new prediction layer is then trained specifically for the secondary task to learn a new set of weights assigned to the features in the penultimate layer. The features are generated by passing the input data through the frozen backbone from the primary task. We may then measure the accuracy of this new prediction layer on some test data. We might conclude that the two tasks are related and possibly even share information when the level of accuracy is reasonably high. In the following, we refer to this approach as the ‘supervised prediction layer information test’ or SPLIT.

SPLIT has been used in the ‘reading race’ study to assess whether a disease detection model may have implicitly learned
are useful for different tasks that are, however, unrelated on an output-level. We also observe higher than chance AUC for the random weights model which means that it even random data transformations retain some signal from the raw images about racial identity.

In Table S3, we additionally provide results for applying SPLIT to biological sex classification. For this, we replaced the backbone trained on White patients with the backbone trained on male patients. Similar to above, we obtain positive SPLIT results for all four backbones. We observe AUC above 0.90 for the backbones trained on ImageNet, all patients, and male patients, which further confirms that SPLIT is insufficient for concluding whether sex information is being used when making predictions for the presence of disease.

**Unsupervised exploration of feature representations**

In the following, we consider an alternative approach for inspecting whether a model may use information about patient characteristics for disease detection. Instead of training a new prediction layer as in SPLIT, we can directly inspect what information is captured (and how it is distributed) in the learned feature representations present in the penultimate layer of the neural network. We can then contrast this with the outputs of the prediction layer that has been trained for the primary task of disease detection. We recall that the prediction layer makes the ultimate decision what information to use. The difficulty is that the feature representations are typically high-dimensional. In a DenseNet-121\(^{16}\), a deep convolutional neural network widely used for image classification, the representations in the penultimate layer have 1024 dimensions. In order to explore what information is encoded in these representations and whether this information aligns with the primary task, we can make use of techniques for dimensionality reduction.

There are many algorithms readily available that we can apply. Here, we choose two widely used unsupervised methods while others may be considered. We employ principle component analysis (PCA) to capture the main modes of variation within the feature representations. We generate two-dimensional scatter plots for the first two PCA modes and overlay different types of patient information. Additionally, we employ t-distributed stochastic neighbor embedding (t-SNE)\(^{17}\), a popular algorithm for visualizing high-dimensional data. Again, we generate different scatter plots by overlaying different types of information. Additionally, we plot the output predictions of the primary task prediction layer. In our case, the output of the disease detection model has 14 dimensions (one output for each of the 14 conditions; cf. Methods section). We may either apply dimensionality reduction on the 14-dimensional outputs, or focus on specific conditions of interest. Here, we focus on the two tasks of classifying ‘no finding’ and ‘pleural effusion’. The two-dimensional logits (which are the unnormalized predictions for the two output classes) can then be directly visualized in a single scatter plot. Samples that are labeled neither ‘no finding’ nor ‘pleural effusion’ are indicated with the label ‘other’ in the plots.

For each scatter plot of PCA, t-SNE and logit outputs, we also show the corresponding (marginal) distributions that one obtains when projecting the data against the axes of the scatter plots. We then visually check if any patterns emerge in these visualizations or if any shifts occur in the marginal distributions when comparing different patient characteristics.

The PCA and t-SNE embeddings of the feature representations typically align well with the ground truth labels of the primary task of disease detection. This is because the features are learned to be discriminative with respect to that task. Data points with the same ground truth labels should obtain similar feature values, and visible patterns or even clusters will emerge in the scatter plots. Additionally, we will observe a separation in the marginal distributions along the dimension that best separates the data. The idea of the unsupervised exploration is then to inspect whether other types of information may show similar patterns which would indicate that the inspected information is related to the primary task. In such a case, we may have a strong indication that the model is under scenario C (cf. Fig. 3), and the inspected information may indeed be used for solving the primary task. If no patterns emerge, neither in the embeddings of the feature representations nor in the logit outputs, and there is no shift in the marginal distributions across subgroups, one may be carefully optimistic that the model is under scenario B. However, we have to remain careful as weak correlations between prediction tasks may not be detected due to limitations of these embedding plots\(^{18}\).

In Fig. 4, we show a variety of different plots produced for the disease detection model trained on all patients. We observe clear clustering of patients in the PCA and t-SNE embeddings and the logit outputs for both disease labels of ‘no finding’ and ‘pleural effusion’. We observe no obvious patterns in the scatter plots and no shifts in the marginal distributions for biological sex, neither in the feature embeddings nor in the logit outputs. This is despite the high AUC obtained with SPLIT (cf. Table S3). In contrast, we observe visible patterns and even clusters for age which is known to have a direct effect on the disease conditions, as patients with label ‘no finding’ are on average younger, and patients with ‘pleural effusion’ are on average older (cf. Fig. S1). We also observe a slight shift of logits for racial identity, mostly for the output of ‘pleural effusion’ for Black patients (visible in the marginal distributions). We also see a small shift in the marginal distributions for racial identity in the PCA embeddings for the second mode, while the disease classes separate predominantly in the first mode of PCA. This may indicate that race information has no strong effect on disease detection, in particular, when compared with the effect of age. The distribution shifts for Black patients in the logit outputs, however, will contribute to the observed shifts in TPR/FPR performance (cf. Table S2 and Fig. 2), and this may have to do with specific variations in the data of this subgroup, as well, as being an effect of the class imbalance across subgroups. We should also note that Black patients are on average 5-8 years younger than White and Asian patients in our study population (cf. Table S1), indicating that there is an interplay of multiple causes contributing to the disparate performance for disease detection.

Similar analyses can be performed for the disease detection models trained on subgroups (cf. Fig. S3 and Fig. S4). For
the model trained on White patients only, we observe a bigger shift in the logit distributions for Black patients, causing the increased shift in TPR/FPR performance. For the model trained on male patients only, we now observe a clear distribution shift for biological sex, causing a shift in TPR/FPR for female patients. These observations confirm the utility of unsupervised exploration of feature representations as a tool to identify potential causes of disparities in the disease detection performance across subgroups.

**Automatic selection of decision thresholds**

A likely cause of the observed shift in logit distributions and the resulting disparity in TPR/FPR performance is the underlying prevalence shift in the CheXpert dataset where the presence of disease varies significantly between groups. Prevalence shift, together with an imbalance between subgroups where the majority of patients identify as White, is known to cause threshold shift in classification models\(^{19}\). In this setting, a single decision threshold optimized over the entire patient population is unlikely to produce equitable TPR/FPR performance. Another reason that may explain the observed disparities is an underlying annotation shift where the diagnostic labeling may be biased, for example, due to systematic underdiagnosis of underserved populations\(^{13}\). Both prevalence shift and annotation shift can cause similar performance disparities in TPR/FPR, however, the two types of shifts require different strategies for mitigation\(^{20}\).

Importantly, disparities caused by prevalence shift may be corrected via post-hoc model calibration, while annotation shift can only be meaningfully addressed through careful re-annotation\(^{21}\).

One possible way to distinguish between bias from prevalence versus annotation shift is to inspect the ROC curves together with the AUC performance metric. While ROC-AUC analysis

---

**Fig. 4 Unsupervised exploration of feature representations trained on all patients.**

a Scatter plots and marginal distributions for the first two modes of PCA applied to feature representations of the chest X-ray test data obtained with the neural network backbone trained for disease detection using data from all patient groups. A random set of 3,000 patients is shown with 1,000 samples from each racial group. Different types of information are overlaid in color from left to right including presence of disease, biological sex, age, and racial identity. b Scatter plots and marginal distributions for the t-SNE embedding. c Scatter plots and marginal distributions for the logit outputs produced by a prediction layer trained for disease detection on top of the backbone. No obvious patterns or distribution shifts emerge for biological sex, while we observe clear patterns for age and possibly small shifts in logits for racial identity. This indicates that age and possibly race are affecting disease detection leading to disparate performance across subgroups.
in itself is clinically not very meaningful, visually inspecting the ROC curves and comparing AUC across subgroups may give an indication whether we are under prevalence or annotation shift. Under prevalence shift, the underlying mechanism of mapping imaging features to prediction targets remains valid, independent of the varying prevalence. Thus, we may observe largely similar ROC curves and comparable AUC across subgroups, even if TPR/FPR are shifted. In annotation shift, however, the mechanism of mapping imaging features to prediction targets changes. Patterns that are labeled as disease in one group may be labeled as ‘no-finding’ in another group. For annotation shift, we may expect to see differences in the ROC curves and AUC metrics across subgroups, unless the model has learned separate mechanisms for each group by distinguishing between patients based on visual traits related to patient characteristics. In this case, we may find patterns in the feature embedding plots related to patient characteristics. It should be noted, however, that subtle differences due to systematic labeling biases may remain undetected. Also, different sources of biases will lead to mixing effects which are even more difficult to interpret.

Given that the AUC and ROC curves in this study are largely consistent across groups (cf. Fig. 2), and no obvious patterns emerge from the unsupervised feature exploration other than clusters related to age, the disease detection models itself should have the capability to perform similar across biological sex and racial identity. However, to obtain consistent TPR/FPR performance, this would require model calibration based on patient-specific adjustment of the decision threshold. In practice, it may be difficult to manually choose thresholds on the basis of protected characteristics. The information would need to be available when running the disease detection model. More importantly, using protected characteristics such as racial identity to adjust for clinical decision making raises ethical concerns. Here, we explore a surprisingly effective (albeit questionable) strategy involving automatic threshold selection leveraging the ability of deep neural networks to accurately predict patient characteristics from medical images.

We report results of such a strategy for the two cases of classifying patients either by biological sex or racial identity. A set of five thresholds is predetermined, one for each group of White, Asian, Black, female and male patients using a target FPR=0.2. At test time, when predicting the presence of disease, we also predict sex and race for the same input scan with FPR=0.2. At test time, when predicting the presence of disease detection models, demonstrating this approach with worked case-examples for both racial identity and biological sex. We also note that differences in performance across groups appear to be related to logit shifts caused by prevalence shift. Finally we consider whether at the technical level this could be mitigated by a strategy which leverages the prediction of patient characteristics to automatically select group-specific decision thresholds reducing the disparate disease detection performance in the considered application and study population.

These findings raise a number of important questions regarding the validity and safety of using group-specific decision thresholds in clinical decision making. The evaluated approach relies on the accurate prediction of protected characteristics, and these predictions will not always be correct. It also remains unclear which characteristics would define meaningful subgroups. In fact, it may be questionable whether such a subdivision of the patient population is ethical. There are rightful arguments that an approach that indirectly uses protected characteristics for decision making may even be dangerous as it could be misused, willingly or unwillingly, to further amplify disparities instead of mitigating them. A major concern here relates to the underlying assumption that the ‘ground truth’ diagnostic labels are largely correct, and that there was no difference in the clinician diagnostic behavior across subgroups, such as systematic underdiagnosis. However, we know that historic data may be biased and that health disparities can be an effect of differences in diagnostic behavior. In this case, adjusting the thresholds could be harmful, as the initially observed disparities that we are trying to correct for would be a result of a flawed performance evaluation on a biased dataset. In fact, if the ML model is unbiased, accurately separating disease conditions equally well across subgroups, we would expect to see disparities when using a biased test set for evaluation. Under the assumption that the test set is unbiased, and if we have reasons to believe that learned disease features should generalize across groups, the automatic, group-specific threshold selection may be a practical approach for reducing performance disparities, and better than relying on a single fixed global threshold that is optimized for the majority of patients and known to disadvantage minority groups.

Here, we would like to highlight that the selection of operating points in form of decision thresholds that dichotomize otherwise continuous predictions always requires human oversight, scrutiny, and transparency. A detailed discussion is beyond the scope of this work and we refer to the reflections on this topic in previous work. We would hope that our findings may positively contribute to the discussion about what constitutes safe and ethical use of AI for clinical decision making. We
believe this is a very timely debate, when an increasing number of reports show that AI may exacerbate health disparities, while there is a lack of solutions how to tackle these.

In this context, it is worth highlighting that there is a very active branch of machine learning research aiming to develop methodology that can prevent (or at least discourage) the use of protected characteristics for decision making in the first place. Here, the goal is to learn fair representations that do not discriminate against groups or individuals. A popular approach in ‘fair ML’ is adversarial training, where a secondary task model is employed during training of the primary task. The secondary, adversarial model acts as a critic to assess whether the learned feature representations contain features predictive of subgroups. This is related to SPLIT, with the difference that the secondary task directly affects the learning of the feature representations by encouraging the active removal of predictive features during training. After training, these models should then yield negative SPLIT results. Other approaches focus on fair predictions by auditing and correcting performance disparities across subgroups during and even after the primary task model has been trained. These advances in fair ML are encouraging, in particular, in cases where we can identify the causes of disparities (e.g., biased training data) and we have reliable information about the protected characteristics. However, it is worth highlighting that many of these approaches can only provide soft constraints and cannot guarantee that the resulting model is fair. It remains to be seen how these approaches translate from the research environment to real world, clinical applications. Approaches for model inspection such as unsupervised exploration of feature representations, will remain important.

In conclusion, we would like to re-emphasize the need of rigorous validation of AI including assessment of performance across vulnerable patient groups. Reporting guidelines such as CONSORT-AI, STARD-AI and others, advocate for complete and transparent reporting when assessing AI performance. A detailed failure case analysis with results reported on relevant subgroups is essential for gaining trust and confidence in the use of AI for critical decision making. Disparities across patient groups can only be discovered with detailed performance analysis. We believe no machine learning training strategy or model inspection tool alone can ever replace the evidence gathered from well designed and executed validation studies and these will remain key in the context of safe and ethical use of AI. Additionally, new frameworks for auditing AI algorithms will likely play an important role for clinical deployment.

**Methods**

Our study sample contains a total of 127,118 chest X-ray scans from 42,884 patients taken from the publicly available CheXpert dataset. The scans are divided into three sets for training (76,205), validation (12,673) and testing (38,240). No scans from the same patient are used in different sets. The validation set is used for model selection, while the test set is the hold out set for measuring the test accuracy of different models, SPLIT and the unsupervised exploration of the feature representations. The study sample and splits are identical to the ones used in the ‘reading race’ study. A detailed description of the patient population is provided in Table S1. For measuring the performance of SPLIT we use the area under the receiver operating characteristic curve (AUC) metric with a one-vs-all approach for each racial group in the case of race classification.

We train deep neural networks for detecting 14 different conditions annotated in the CheXpert dataset. Similar to previous work, we use a multi-label approach as patients may have multiple conditions. The presence of each condition is predicted via a dedicated one-dimensional output which is passed through a sigmoid function to obtain predictions between zero and one. The simultaneous detection of the individual conditions uses a common feature representation obtained from a shared neural network backbone.

In our analysis, we focus on the labels ‘no finding’ and ‘pleural effusion’. In total, there were 10,916 (or 9%) scans labeled with ‘no finding’, and 51,574 (or 41%) scans that had a diagnosis of ‘pleural effusion’. In the test set, this amounted to 3,316 (9%) and 15,510 (41%) scans, respectively.

For the deep neural network models, we use a PyTorch implementation of a DenseNet-121. All chest X-ray images are resized using bi-linear interpolation to fit the required 224 × 224 pixel resolution. For the models trained for disease detection, we initialize the network with pre-trained ImageNet weights as provided in the torchvision module. We employ the Adam optimizer with default parameters and a learning rate of 0.001 for all our experiments. We use the same configuration when training the prediction layers in the SPLIT experiments. We select the model checkpoint with the best performance on the validation set. We also explored another network architecture, ResNet-34, which resulted in similar findings. Additional details about the implementation are provided in our code repository.

The input to PCA is a n-by-m feature matrix where n=38,240 is the number of scans in the test set, and m=1024 is the size of the output of the penultimate layer in the DenseNet-121. We use the PCA and t-SNE implementations provided in the Python-based machine learning library scikit-learn. We initialize t-SNE with PCA embeddings as this has been shown to yield better results. We use PCA embeddings that preserve 99% of the variance of the feature representations, which yields 172, 206, 129, and 450 PCA modes for the backbones trained on all patients, White patients only, male patients only, and ImageNet, respectively. Otherwise, we use the default parameters for t-SNE. To improve the visibility in the scatter plots, we randomly sample 1,000 scans from each racial group, so in total 3,000 samples are shown in the scatter plots.

**Ethical approval**

This research is exempt form ethical approval as the analysis is based on secondary data which is publicly available, and no permission is required to access the data.

**Data availability**

All data used in this work is publicly available. The CheXpert imaging dataset together with the patient demographic informa-
We would like to thank the authors of the ‘reading race’ study. All code that is required for replicating the results in this article is made available under open source Apache 2.0 license on our dedicated GitHub repository https://github.com/biomedia-mira/chexploration.

Code availability

All code that is required for replicating the results in this article is made available under an open source Apache 2.0 license on our dedicated GitHub repository: https://github.com/biomedia-mira/chexploration.

Author contributions

BG conceived and designed the study and conducted the experiments; BG, CJ, MB, and SW analysed and interpreted the results; SW performed data processing. All authors conceptualised, wrote, and co-edited the manuscript.

Acknowledgments

We would like to thank the authors of the ‘reading race’ study for the discussions and for making their code and data openly available. We are also very grateful to Xiaoxuan Liu and Alastair Denniston for their valuable feedback and helpful discussions during the preparation of this manuscript.

Funding sources

This work is supported through funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant Agreement No. 757173, Project MIRA, ERC-2017-STG). SW is supported by the UKRI London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare. CJ is supported by Microsoft Research and EPSRC through Microsoft’s PhD Scholarship Programme. MB is funded through an Imperial College London President’s PhD Scholarship.

Competing interests

BG is part-time employee of HeartFlow and Kheiron Medical Technologies and holds stock options as part of the standard compensation package.

References

1. Esteva, A. et al. A guide to deep learning in healthcare. Nature Medicine 25, 24–29 (2019).
2. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015).
3. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering 2, 158–164 (2018).
4. Rim, T. H. et al. Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. The Lancet Digital Health 2, e526–e536 (2020).
5. Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S. & Obermeyer, Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nature Medicine 27, 136–140 (2021).
6. Banerjee, I. et al. Reading race: AI recognises patient’s racial identity in medical images. arXiv preprint arXiv:2107.10356 (2021).
7. Paul, H. Y. et al. Radiology “forensics”: determination of age and sex from chest radiographs using deep learning. Emergency Radiology 1–6 (2021).
8. Adamson, A. S. & Smith, A. Machine learning and health care disparities in dermatology. JAMA dermatology 154, 1247–1248 (2018).
9. 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).
10. Geirhos, R. et al. Shortcut learning in deep neural networks. Nature Machine Intelligence 2, 665–673 (2020).
11. Irvin, J. et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 590–597 (2019).
12. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y. & Ghassemi, M. CheXclusion: Fairness gaps in deep chest x-ray classifiers. In BIOCOMPUTING 2021: Proceedings of the Pacific Symposium, 232–243 (World Scientific, 2020).
13. Seyyed-Kalantari, L., Zhang, H., McDermott, M., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. Nature Medicine 1–7 (2021).
14. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? In Advances in Neural Information Processing Systems (2014).
15. Deng, J. et al. ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255 (IEEE, 2009).
16. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 4700–4708 (2017).
17. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. Journal of Machine Learning Research 9 (2008).
18. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. Nature Communications 10, 1–14 (2019).
19. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. Journal of Big Data 6, 1–54 (2019).
20. Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V. & Herrera, F. A unifying view on dataset shift in classification. Pattern recognition 45, 521–530 (2012).
21. Castro, D. C., Walker, I. & Glöckner, B. Causality matters in medical imaging. Nature Communications 11, 1–10 (2020).
22. Vyas, D. A., Eisenstein, L. G. & Jones, D. S. Hidden in plain sight—reconsidering the use of race correction in...
clinical algorithms. *New England Journal of Medicine* **383**, 874–882 (2020).

23. Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, 151–159 (2020).

24. Spiegelhalter, D. J. Probabilistic prediction in patient management and clinical trials. *Statistics in medicine* **5**, 421–433 (1986).

25. Ibrahim, H., Liu, X., Zariffa, N., Morris, A. D. & Denniston, A. K. Health data poverty: an assailable barrier to equitable digital health care. *The Lancet Digital Health* (2021).

26. Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, 325–333 (PMLR, 2013).

27. Edwards, H. & Storkey, A. Censoring representations with an adversary. In *International Conference on Representation Learning* (2015).

28. Madras, D., Creager, E., Pitassi, T. & Zemel, R. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, 3384–3393 (PMLR, 2018).

29. Kearns, M., Neel, S., Roth, A. & Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, 2564–2572 (PMLR, 2018).

30. Kim, M. P., Ghorbani, A. & Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 247–254 (2019).

31. Liu, X., Rivera, S. C., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine* **26**, 1364–1374 (2020).

32. Sounderajah, V. *et al.* Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ open* **11**, e047709 (2021).

33. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences* **117**, 12592–12594 (2020).

34. Wiens, J. *et al.* Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine* **25**, 1337–1340 (2019).

35. Raji, I. D. *et al.* Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44 (2020).

36. Schulam, P. & Saria, S. Can you trust this prediction? auditing pointwise reliability after learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1022–1031 (PMLR, 2019).

37. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).

38. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* **12**, 2825–2830 (2011).
A. Study population

Table S1 Characteristics of the study population.

| Attribute         | All       | White    | Asian    | Black    |
|-------------------|-----------|----------|----------|----------|
| Patients          | 42,884    | 33,338   | 6,642    | 2,904    |
| Scans             | 127,118   | 99,027   | 18,830   | 9,261    |
| Age               | 63 ± 17   | 64 ± 17  | 61 ± 17  | 56 ± 17  |
| Female            | 52,436 (41) | 39,735 (40) | 8,132 (43) | 4,569 (49) |
| No finding        | 10,916 (9) | 8,236 (8) | 1,716 (9) | 964 (10) |
| Pleural effusion  | 51,574 (41) | 40,545 (41) | 7,953 (42) | 3,076 (33) |

Training data

| Attribute         | All       | White    | Asian    | Black    |
|-------------------|-----------|----------|----------|----------|
| Patients          | 25,730    | 20,034   | 3,945    | 1,751    |
| Scans             | 76,205    | 59,238   | 11,371   | 5,596    |
| Age               | 63 ± 17   | 64 ± 17  | 62 ± 17  | 56 ± 17  |
| Female            | 31,432 (41) | 23,715 (41) | 4,754 (42) | 1,856 (33) |
| No finding        | 6,514 (9) | 4,910 (8) | 1,046 (9) | 558 (10) |
| Pleural effusion  | 31,015 (41) | 24,405 (41) | 4,754 (42) | 1,856 (33) |

Validation data

| Attribute         | All       | White    | Asian    | Black    |
|-------------------|-----------|----------|----------|----------|
| Patients          | 4,288     | 3,348    | 666      | 274      |
| Scans             | 12,673    | 9,945    | 1,009 (14) | 919 (7) |
| Age               | 62 ± 17   | 63 ± 17  | 62 ± 17  | 55 ± 16  |
| Female            | 5,030 (40) | 3,933 (40) | 667 (37) | 430 (47) |
| No finding        | 1,086 (9) | 817 (8)  | 175 (10) | 94 (10)  |
| Pleural effusion  | 5,049 (40) | 3,988 (40) | 738 (41) | 323 (35) |

Test data

| Attribute         | All       | White    | Asian    | Black    |
|-------------------|-----------|----------|----------|----------|
| Patients          | 12,866    | 9,956    | 2,031    | 879      |
| Scans             | 38,240    | 29,844   | 5,650 (15) | 2,746 (7) |
| Age               | 63 ± 17   | 64 ± 17  | 61 ± 17  | 57 ± 16  |
| Female            | 15,974 (42) | 12,087 (41) | 2,489 (44) | 1,348 (49) |
| No finding        | 3,316 (9) | 2,509 (8) | 495 (9)  | 312 (11) |
| Pleural effusion  | 15,510 (41) | 12,152 (41) | 2,461 (44) | 897 (33) |

Breakdown of demographics over the set of patient scans by racial groups and training, validation and test splits. Proportions across the splits are very consistent. Percentages in brackets are with respect to the number of scans. We also provide the number of unique patients for each group.

Fig. S1 Relationship of patient demographics and presence of disease. a Age distribution over racial identity grouped by biological sex. b Age distribution over racial identity grouped by presence of disease. c Number of scans for each race group grouped by presence of disease. d Number of scans for biological sex grouped by presence of disease.
B. Disease detection

Table S2 Disease detection results for deep neural networks trained on different patient groups.

| Training data | White | Asian | Black | Female | Male | TPR / FPR |
|---------------|-------|-------|-------|--------|------|-----------|
| All patients  | 0.87  | 0.88  | 0.86  | 0.87   | 0.88 | 0.79 / 0.20 |
| White patients| 0.86  | 0.87  | 0.88  | 0.86   | 0.87 | 0.78 / 0.19 |
| Male patients | 0.86  | 0.86  | 0.88  | 0.86   | 0.87 | 0.78 / 0.20 |

Group-specific thresholds

| All patients  | same as above | 0.79 / 0.20 |
| White patients| same as above | 0.78 / 0.20 |
| Male patients | same as above | 0.78 / 0.20 |

Pleural effusion

| All patients  | 0.86  | 0.88  | 0.86  | 0.87   | 0.86 | 0.77 / 0.21 |
| White patients| 0.86  | 0.87  | 0.85  | 0.86   | 0.86 | 0.76 / 0.21 |
| Male patients | 0.85  | 0.87  | 0.84  | 0.85   | 0.86 | 0.73 / 0.20 |

Group-specific thresholds

| All patients  | same as above | 0.76 / 0.20 |
| White patients| same as above | 0.76 / 0.20 |
| Male patients | same as above | 0.74 / 0.20 |

Disease detection performance reported separately for each race group and biological sex for the two labels of ‘no finding’ (top) and ‘pleural effusion’ (bottom).

Fig. S2 Disease detection performance with group-specific thresholds. a-c ROC curves for the performance on the ‘no finding’ label for three different disease detection models trained on all patients, White patients, and male patients. d-f ROC curves for the performance on the ‘pleural effusion’ label for the same three models. The group-specific threshold selection largely preserves TPR/FPR performance across subgroups.
C. SPLIT for sex classification

Table S3 Results for applying SPLIT to sex classification with four different neural network backbones.

| Neural network backbone               | White | Asian | Black |
|--------------------------------------|-------|-------|-------|
| Random weights                       | 0.84  | 0.80  | 0.85  |
| ImageNet                             | 0.92  | 0.90  | 0.91  |
| Disease detection (all patients)     | 0.94  | 0.90  | 0.93  |
| Disease detection (male patients)    | 0.93  | 0.89  | 0.91  |
| Sex classification (all patients)    | 1.00  | 1.00  | 1.00  |

AUC for classifying sex reported separately for each racial group. For comparison, in the last row we also report the AUC for a neural network trained specifically for sex classification.
D. Unsupervised exploration of feature representations trained on White patients

Fig. S3 Unsupervised exploration of feature representations trained on White patients. a,b Scatter plots and marginal distributions of the two-dimensional embeddings of PCA and t-SNE for the neural network backbone trained for disease detection on White patients only. c Scatter plots and marginal distributions for the logit outputs produced by a prediction layer trained for disease detection on top of the backbone. We observe clear distribution shifts for racial identity in the logit outputs, which affects disease detection leading to disparate performance for Black patients.
E. Unsupervised exploration of feature representations trained on male patients

Fig. S4 Unsupervised exploration of feature representations trained on male patients. a,b Scatter plots and marginal distributions of the two-dimensional embeddings of PCA and t-SNE for the neural network backbone trained for disease detection male patients only. c Scatter plots and marginal distributions for the logit outputs produced by a prediction layer trained for disease detection on top of the backbone. We observe clear distribution shifts for biological sex in the logit outputs, which affects disease detection leading to disparate performance for female patients.