Demographically-Informed Prediction Discrepancy Index: Early Warnings of Demographic Biases for Unlabeled Populations

Anonymous authors
Paper under double-blind review

Abstract

An ever-growing body of work has shown that machine learning systems can be systematically biased against certain sub-populations defined by attributes like race or gender. Data imbalance and under-representation of certain populations in the training datasets have been identified as potential causes behind this phenomenon. However, understanding whether data imbalance with respect to a specific demographic group may result in biases for a given task and model class is not simple. An approach to answering this question is to perform controlled experiments, where several models are trained with different imbalance ratios and then their performance is evaluated on the target population. However, in the absence of ground-truth annotations at deployment for a new target population, most fairness metrics cannot be computed. In this work, we explore an alternative method to study potential bias issues based on the output discrepancy of pools of models trained on different demographic groups. Models within a pool are otherwise identical in terms of architecture, hyper-parameters, and training scheme. Our hypothesis is that the output consistency between models may serve as a proxy to anticipate biases concerning demographic groups. In other words, if models tailored to different demographic groups produce inconsistent predictions, then biases are more prone to appear in the task under analysis. We formulate the Demographically-Informed Prediction Discrepancy Index (DIPDI) and validate our hypothesis in numerical experiments using both synthetic and real-world datasets. Our work sheds light on the relationship between model output discrepancy and demographic biases and provides a means to anticipate potential bias issues in the absence of ground-truth annotations. Indeed, we show how DIPDI could provide early warnings about potential demographic biases when deploying machine learning models on new and unlabeled populations that exhibit demographic shifts.

1 Introduction

Machine learning (ML) models are susceptible to exhibiting biases against certain subpopulations defined in terms of sensitive demographic characteristics such as gender, age, or race. Examples of such biases can be found in a variety of fields, including predictive policing Angwin et al. (2016), facial analysis Buolamwini & Gebru (2018), and healthcare Chen et al. (2019); Ricci Lara et al. (2022). Factors that contribute to biased models may include the data used for training and evaluation, as well as decisions made during the development process Suresh & Guttag (2019). As ML applications in the real world become increasingly widespread, it is important to evaluate models to ensure that they are not only accurate but also produce fair and ethical results.

In particular, under-representation of certain demographic groups has been identified as one of the main causes of bias when developing predictive systems. For example, gender imbalance in X-ray medical imaging datasets has been shown to have a significant impact on the performance of assisted diagnosis systems for thoracic diseases based on convolutional neural networks Larrazabal et al. (2020), and under-representation of ethnic groups has also been found to influence model performance for cardiac image segmentation Puyol-Antón et al. (2021). However, in other tasks, such data imbalance has not been associated with unequal performance. In Petersen et al. (2022) for example, the authors found that in the case of Alzheimer's disease prediction from brain magnetic resonance images (MRI), gender imbalance in the training dataset did not lead to a clear pattern of improved model performance for the majority group. A similar phenomenon was observed in Kinyanjui et al. (2020), where the authors studied under-representation of skin color when analyzing dermoscopic images for skin cancer detection, and did not observe such disparities. This observation was then challenged by another study Groh et al. (2021) which found disparities in performance arising from training a neural network on only a subset of skin types. In all, it is not always a fact that data imbalance will result in biased automated systems. To complicate matters further, even when the presence of biases can be assessed the during development of an automated tool, these properties may not transfer under distribution shifts Schrouff et al. (2022), for instance, once the model is deployed. This is a problem for fairness metrics which require ground-truth annotations, which are expensive to obtain and may not be available before deployment. Given these issues, a valid question that one may then ask is: can we anticipate whether models will exhibit biases with respect to data imbalance in terms of a particular protected attribute in the absence of ground-truth annotations?

Typical approaches to identify biases in ML models involve subgroup analysis and controlled experiments where both demographic and target labels are available Larrazabal et al. (2020); Buolamwini & Gebru (2018); Glocker et al. (2021). Model performance across demographic groups is commonly evaluated employing one or more metrics Corbett-Davies & Goel (2018) with the implicit assumption that the presence or absence of biases during development will be representative of the behaviour of these models when applied to previously unseen data at deployment. Recent findings regarding how fairness properties transfer across distribution shifts in real-world healthcare applications due to changes in geographic location, population demographics or even environmental conditions, warn us about the risks of this assumption Schrouff et al. (2022). A system that did not exhibit strong biases in the source population may begin to do so when the target population changes. This is particularly concerning in applications like healthcare, where collecting expert annotations on large datasets can be costly and time-consuming Ricci Lara et al. (2022), meaning that fairness metrics requiring labels may not be computed, with the result of biases going unnoticed. In this context, developing methods that can be used without the need for ground truth in the target population becomes highly relevant. In this paper, we are interested in exploring ways to anticipate potential bias issues that may arise in the context of a given task for a novel unlabeled target population. We do so by proxy: using an index that we call Demographically-Informed Prediction Discrepancy Index (DIPDI), which can be computed in the absence of ground truth annotations. We show in numerical experiments on both synthetic and real-world datasets that this index is indeed indicative of bias proneness, providing an early warning for potential fairness issues in these settings.

2 Related work

The implicit assumption that model assessment during development is representative of its behaviour at deployment is not unique to fairness studies. Indeed, anticipating whether a model will systematically fail or not when ground-truth annotations are not available is a current topic of interest in the field, and one way to tackle this issue is to look at predictive uncertainty Gal et al. (2016). Intuitively, if a well-calibrated model systematically makes highly uncertain predictions for certain individuals, then chances are that these predictions will have a higher failure rate for those individuals. In this context, recent studies have analyzed the relation between fairness and uncertainty, postulating that uncertainty estimates can be used to obtain fairer models, improve decision-making, and build trust in automated systems Bhatt et al. (2021). For example, Lu et al. (2021) analyzed how alternative uncertainty estimation methods can be used to evaluate subgroup disparities in mammography image analysis, while Stone et al. (2022) leveraged epistemic uncertainty estimates to mitigate minority group biases during training. The work Dusenberry et al. (2020) discusses the role of model uncertainty in predictive models for Electronic Health Record (EHR), and shows how it can change across different patient subgroups, in terms of ethnicity, gender and age, considering Bayesian and deep ensemble approaches for uncertainty estimation. Even though in this work we do not directly rely on the notion of uncertainty, our study is highly influenced by this idea, as it explores the use of output discrepancy for a set of models as a way of anticipating bias issues. This notion is closely related to ensemble variance, usually employed as a measure of uncertainty for ensemble methods Lakshminarayanan et al. (2017); Pividori et al. (2016); Larrazabal et al. (2021). Another important concept in our study is that of consistency Wang et al. (2020), defined as the ability of a set of multiple trained learners to reproduce an output for the same input Wang et al. (2020). According to this concept, model outputs are analyzed irrespective of whether they are correct or incorrect, and as such, it does not require ground-truth annotations to be computed. This idea will be central to our study.

Contributions: Here we present a methodology to understand whether biases with respect to a given demographic attribute are prone to arise in a new unlabeled dataset. We do so by analyzing the output consistency of a pool of models, where each model is *trained on separate demographic groups*, but is otherwise identical in terms of architecture, hyper-parameters and training scheme. We introduce a new index, DIPDI, based on the following hypothesis: if models specialized in different demographic groups produce inconsistent and highly discrepant predictions for the same test data, then the task under analysis is prone to be biased against that specific demographic attribute.

We validate our hypothesis using synthetic and real-world datasets, focusing on a simple task: age estimation from face photos and X-ray images, using three real-world datasets and considering different cases of demographic imbalance in the training data. Our results indicate that DIPDI can be used to anticipate potential bias issues in the absence of ground truth labels, and confirm the association between output discrepancy and bias proneness. We also assess the behaviour of DIPDI for unseen populations with demographic shifts, showing how it can be used to measure bias proneness in dynamic contexts. Moreover, since our metric does not require expert annotations to be computed, it could help to anticipate bias issues in real-world scenarios and give early warnings when deploying machine learning models on new, unlabeled populations.

3 Demographically-Informed Prediction Discrepancy Index (DIPDI)

3.1 Quantifying output discrepancy within and between demographically-informed sets of models

Given two sets of predictive models $\mathbb{A} = \{A_1, A_2\}$ and $\mathbb{B} = \{B_1, B_2\}$, we are interested in analyzing how the output discrepancy of models within the same set compares to the output discrepancy of models coming from different sets when they are evaluated on samples from an unlabeled dataset $\mathcal{D}$. For simplicity, we exemplify the calculation of DIPDI for regression models, though the method can be readily extended to other tasks. Here $A(\mathbf{x}_k) : \mathcal{D} \longrightarrow \mathbb{R}^+$, where $\mathbf{x}_k \in \mathcal{D}$ can be images or other types of data for subject k, and the output of $A(\mathbf{x}_k)$ is a positive real number (e.g. the problem of regressing age from an X-ray image or a face photography).

We then define an *output discrepancy* function $\mathcal{N}_{\mathcal{D}}(M_1, M_2)$, that takes as input two models M_1 and M_2, and returns a number representing how different their outputs are on average when evaluated on all samples from $\mathcal{D}$. If M_1 and M_2 are regression models as in our example, then we can simply define $\mathcal{N}_{\mathcal{D}}$ as

$$\mathcal{N}_{\mathcal{D}}(M_1, M_2) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_k \in \mathcal{D}} |M_1(\mathbf{x}_k) - M_2(\mathbf{x}_k)| \tag{1}$$

In other words, the output discrepancy is the average of the absolute difference between the predicted values of models M_1 and M_2 for all subjects in the dataset. It returns a number closer to 0 when the outputs of the two models *for every data sample* are similar, and higher if they tend to differ. Since we are interested in analyzing the *output discrepancy* for models within and between sets, we consider the following ratio as an indicator of relative output discrepancy:

$$\Phi_{\mathcal{D}}(\mathbb{A}, \mathbb{B}) = \log \left[\frac{\mathcal{N}_{\mathcal{D}}(A_1, B_1)\mathcal{N}_{\mathcal{D}}(A_2, B_2)}{\mathcal{N}_{\mathcal{D}}(A_1, A_2)\mathcal{N}_{\mathcal{D}}(B_1, B_2)} \right]. \tag{2}$$

This inter-model prediction discrepancy will be close to 0 when the output discrepancy for models within the same set (numerator) is similar to that of models coming from different sets (denominator), and it will be greater than 0 when the discrepancy for models coming from different sets is greater than that of models coming from the same set. When applied to models trained on different demographic groups, we refer to this diverse set of models as a demographically-informed pool and $\Phi_{\mathcal{D}}$ becomes our *Demographically-Informed Prediction Discrepancy Index (DIPDI)*. This will be the case, for example, when models in $\mathbb{A}$ are trained on male individuals while models in $\mathbb{B}$ are trained on female individuals.

3.2 DIPDI as a proxy for anticipating bias issues

Our goal is to anticipate whether biases may arise with respect to a particular protected attribute a in a novel dataset before annotated labels become available. The task at hand here is age regression and the protected attribute a indicates the gender of the patient, which for simplicity we take as *male* ($a = M$) or *female* ($a = F$). We create two sets of models (age regressors): $\mathbb{A}$, where models A_i are trained only on male patients, i.e. $a = M$; and $\mathbb{B}$, where models B_i are trained only on female patients, i.e. $a = F$. We say that this constitutes a demographically-informed pool of models, as each of them was trained on individuals from a particular demographic group characterized by the protected attribute a. Let us also have a fixed dataset $\mathcal{D}$ that will be used as the novel target population where potential biases would want to be flagged. $\mathcal{D}$ is a balanced dataset according to the protected attribute a (but unlabeled with respect to output class, i.e. without the reference age). In our example, this means that $\mathcal{D}$ is composed of 50% male and 50% female patients.

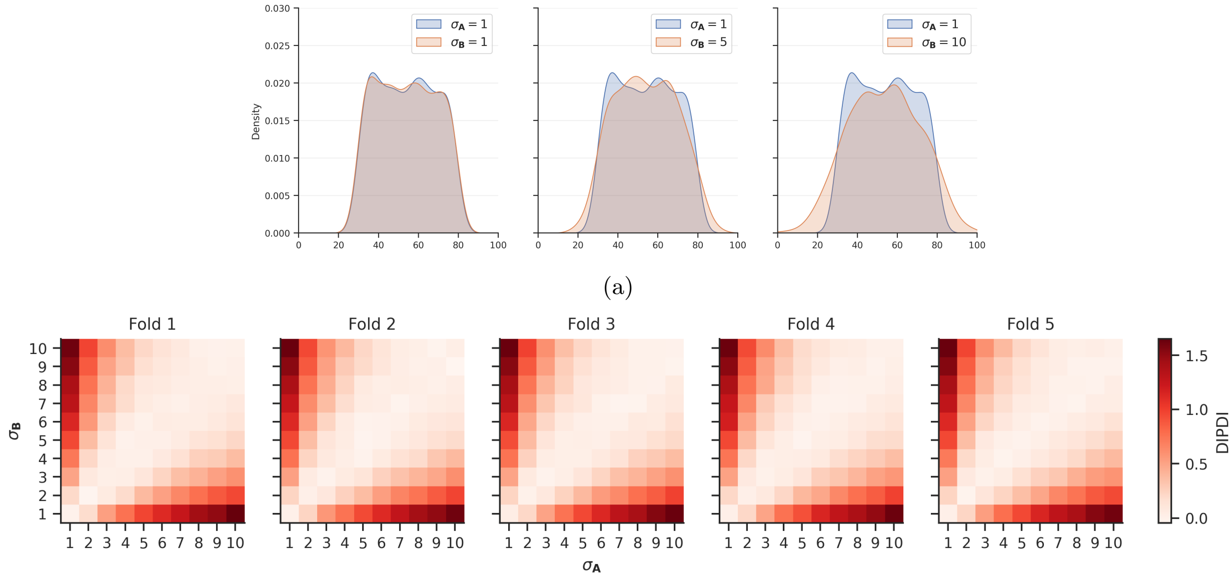
Our hypothesis is that for larger values of $\Phi_{\mathcal{D}}(\mathbb{A}, \mathbb{B})$, computed for a pool of models comprising sets $\mathbb{A}$ and $\mathbb{B}$, biases are more likely to emerge. In other words, we hypothesize that inconsistencies between the output discrepancy of models trained on highly unbalanced datasets with respect to the protected attribute a will tend to co-occur with potential bias issues. To confirm our hypothesis, we first look for biases with respect to a using ground-truth annotations in the target population (following a strategy similar to Larrazabal et al. (2020)), by computing performance gaps in terms of absolute error (using the ground-truth of each sub-population). Then we calculate DIPDI, which does not require ground-truth labels for $\mathcal{D}$, and verify if it produces results that are in line with the conclusions we drew when using the annotations. As a sanity check, we incorporate control experiments where we break the assumption that sets $\mathbb{A}$ and $\mathbb{B}$ are trained on different demographic groups, and show that in these cases DIPDI returns values close to 0.

4 Experimental validation

We start by verifying the behaviour of DIPDI under controlled conditions using synthetic data. Then, we perform a set of experiments to assess if age estimation from face and x-ray images is prone to be biased with respect to gender, particularly when a certain subgroup is underrepresented Larrazabal et al. (2020). We show that DIPDI anticipates potential biases against the minority group when training data is highly imbalanced in gender representation for the tasks of age estimation from X-ray and facial images (Section 4.3). To this end, we employ ground-truth annotations for the target population to compute performance gaps in terms of mean absolute errors for models trained with different imbalance ratios, in the different subgroups. Then, we proceed to compute DIPDI (which does not require ground-truth labels) in the target population. We show that bias gaps tend to occur for larger DIPDI values (Section 4.4). We conclude the study by showing how DIPDI can serve to anticipate potential bias issues at deployment in populations with demographic shifts, when target annotations are not yet available.

4.1 DIPDI on synthetic data

Before proceeding with the evaluation of DIPDI in real data, we verify its behaviour under controlled conditions using synthetic data. To do so, we simulate the predictions of two sets of models $\mathbb{A} = \{A_1, A_2\}$ and $\mathbb{B} = \{B_1, B_2\}$ when evaluated on samples from a synthetic dataset $\mathcal{D}$. We then systematically evaluate DIPDI in scenarios with different levels of disagreement between $\mathbb{A}$ and $\mathbb{B}$. The model discrepancy is here



(a)

(b)

Figure 1: (a) Synthetic data construction. Examples of predicted ages simulated for models in $\mathbb{A}$ and $\mathbb{B}$, for increasing levels of prediction disparities (left to right). (b) DIPDI on synthetic data. Both $\sigma_{\mathbb{A}}$ and $\sigma_{\mathbb{B}}$ range from 1 to 10. Each fold represents a new run of the experiment with different random seeds. The DIPDI index is computed by averaging 3 simulations for each $\sigma_{\mathbb{A}}$.

simulated by the addition of a stochastic value of varying size (disagreement level) to the output predictions (Figure 1a). Additionally, we analyze the stability of our method as a function of the number of data samples $N = |\mathcal{D}|$ in Appendix A.1.

We consider the task of age estimation, so the outputs of models in $\mathbb{A}$ and $\mathbb{B}$ are assumed to represent *predicted ages*. We start with a fixed sample $\mathcal{Y}$ drawn from a uniform distribution of ages between 30 and 80, representing the *ground-truth ages*, $y_k \in \mathcal{Y}$. We simulate synthetic predictions for the models in $\mathbb{A}$ and $\mathbb{B}$ by perturbing $\mathcal{Y}$ with Gaussian noise sampled from distributions $n_{\mathbb{A}} \sim \mathcal{N}(0, \sigma_{\mathbb{A}})$ and $n_{\mathbb{B}} \sim \mathcal{N}(0, \sigma_{\mathbb{B}})$. Thus, for a fictitious data sample k with ground-truth label $y_k \in \mathcal{Y}$, the synthetic model predictions are $A_i(\mathbf{x}_k) = y_k + n_{\mathbb{A}}$ and $B_i(\mathbf{x}_k) = y_k + n_{\mathbb{B}}$. Varying the standard deviations allows us to create scenarios where the predicted ages for the analysis groups are more or less similar, and then analyze the behaviour of DIPDI under different discrepancy ratios (see Figure 1a).

DIPDI values for different discrepancy scenarios are displayed in Figure 1b, considering $N = 1000$ and $\sigma_{\mathbb{A}}$ and $\sigma_{\mathbb{B}}$ values in the range [1-10]. Note that when the outputs of $\mathbb{A}$ and $\mathbb{B}$ are similarly perturbed (as shown on the diagonal of each image), then Φ is close to 0. However, when perturbations are sampled from a wider Gaussian in one set than the other (as shown outside the diagonal of each image), Φ tends to be higher than 0. This confirms the desired behaviour for our index: when intra-set predictions are more consistent than inter-set predictions, the index returns larger values.

4.2 DIPDI on real scenarios: datasets and experimental setup

We conduct experiments on the task of age estimation using convolutional neural networks (CNNs), employing three public databases: ChestX-ray14 Wang et al. (2017), UTKFace Zhang et al. (2017) and IMDB-WIKI Rothe et al. (2018). All experiments were performed using PyTorch Paszke et al. (2017) on an NVIDIA Titan X GPU.[1]

[1] Our code is publicly available at XXXXXXXXXXX

ChestX-ray14 dataset. The ChestX-ray14 dataset contains 112,120 high-resolution frontal-view radiographs of 30,805 unique patients with age and gender labels. Each image is annotated with up to 14 different chest disease labels extracted from radiology reports using natural language processing techniques. We use the ChestX-ray14 dataset to perform subgroup analysis in terms of gender and to evaluate DIPDI in models trained to perform age estimation from radiological images. We use here for gender the binary labels reported in the dataset, i.e. male and female. In order to perform these experiments, we collect the "healthy" subjects from the ChestX-ray14 dataset, i.e. those labeled as "No Finding", meaning that none of the 14 pathologies was diagnosed. We do so to avoid potential confounds arising for instance from varying disease prevalence across demographic sub-groups. Subsequently, one image per patient was randomly selected, resulting in a total of 22,850 images. This database was divided into 5 folds using a stratified cross-validation strategy, where each fold is balanced by gender. For each cross validation instance, one fold is used to evaluate the model and the remaining 4 folds are used to train the model, which are further sub-divided into training (90%) and validation (10%) subsets for hyper-parameter tuning and model selection.

For all experiments on ChestX-ray14 we used a DenseNet-121 Huang et al. (2017) that had been pre-trained on ImageNet Russakovsky et al. (2015). The last layer of the network was replaced with an adaptive pooling layer, followed by a single-output neuron layer to predict age. The models were trained for 50 epochs using the Adam optimizer Kingma & Ba (2014) with default parameters and the mean absolute error (MAE) loss function.

UTKFace dataset. The UTKFace dataset is a collection of over 20,000 facial images spanning ages from 0 to 116, annotated for age, gender, and ethnicity. It exhibits diverse variations in pose, facial expression, lighting, occlusion, and resolution. Images were filtered to include ages from 10 to 100 and followed the same training settings as applied in the case of ChestX-ray14. This dataset is utilized for subgroup analysis and assessing DIPDI in age estimation models.

We employed a VGG-16 architecture Simonyan & Zisserman (2014), pretrained on ImageNet, with the final layer replaced by adaptive pooling and a single-output neuron layer for age prediction.

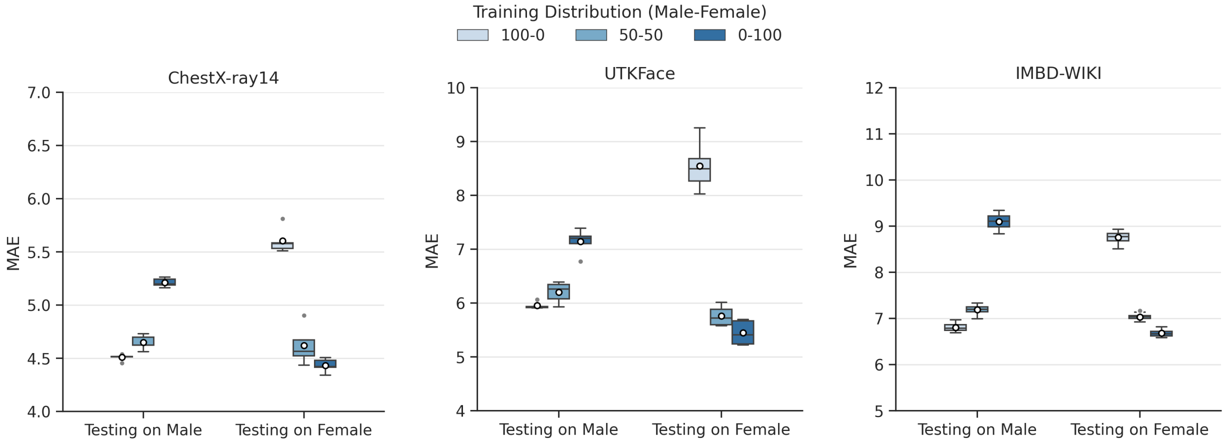
IMDB-WIKI dataset. The IMDB-WIKI dataset consists of 523,051 face images of 20,284 celebrities collected from IMDB and Wikipedia with age and gender labels. Age is estimated from the date of birth and the year when the photo was taken. The IMBD-WIKI dataset is used to perform subgroup analysis and to evaluate DIPDI for models trained to perform age estimation from facial images.

We used a VGG-19 architecture Simonyan & Zisserman (2014) pre-trained on ImageNet. We added a single-output neuron layer with ReLU activation and fine-tuned the last four layers. The models were trained with a MAE loss for 10 epochs using the Adam optimizer with default parameters.

4.3 Assessing the impact of gender imbalance for age estimation in a supervised setting

Age estimation from X-ray images. We analyze the impact of gender imbalance in age estimation from radiological images by performing a supervised subgroup evaluation. The aim is to understand if the age estimation task is prone to be biased with respect to gender if a certain subgroup is under-represented. We will then see if the proposed DIPDI can predict such behaviour without ground-truth annotations. We train models with different degrees of gender imbalance and then examine their performance separately in male and female subgroups. We consider three cases of gender imbalance in training: 100-0 (100% male), 50-50 (50% male and 50% female), and 0-100 (100% female). Each model was run 10 times with different random seeds. Importantly, the test set is fixed, and male and female subgroups in the test population are equal in size. This means that every model is evaluated on equal footing.

The MAE on the ChestX-ray14 is shown in Figure 2a. The results show that an imbalance in the protected attribute leads to a significant difference in performance across subgroups. For example, when testing on female subjects, models trained only on male (100-0) data have higher MAE than models trained on female images. The same happens when testing on female individuals: models trained only on female data (0-100) significantly outperform those trained on male data. Moreover, the differences between male and female subgroups are less significant when the training data is balanced (50-50). These results are consistent with previous observations reported by Larrazabal et al. (2020) in the context of disease prediction from X-ray



(a)

(b)

(c)

Figure 2: Mean absolute error (MAE) for age estimation on ChestX-ray14, UTKFace, and IMDB-WIKI by subgroup (male and female). Trained models with different degrees of male-female imbalance (100-0, 50-50, and 0-100) are shown in blue shades. Note how for a given test subgroup (e.g. male patients), the model trained on patients of the same gender significantly outperforms the other. This confirms that age estimation from both X-ray and face images is prone to gender bias if proper care is not taken when creating the training databases

images. Appendix A.2 contains supplementary results for ChestX-ray14 including additional statistics and metrics.

Age estimation from face images. We also perform a similar analysis to study the impact of gender imbalance in age estimation from facial images. For UTKFace, we followed the same procedure as for ChestX-ray14. Figure 2b shows the MAE results for models trained with varying degrees of gender imbalance and tested on male and female subgroups over 5 folds. These results demonstrate a significant performance disparity across subgroups resulting from an imbalance in the protected attribute.

In the case of IMDB-WIKI, we train an ensemble of 5 models with different degrees of male-female imbalance and then evaluate their performance separately in male and female subgroups. We perform 20-fold cross-validation with a 60/20/20 ratio for the training, validation, and test sets.

The MAE is shown in Figure 2c for these experiments. We observe that the models perform best in the subgroup (either male or female) that is most represented in training, but their performance deteriorates in the other subgroup. We include additional results for UTKFace as well as a more fine-grained study with different imbalanced degrees for IMDB-WIKI in Appendix A.2.

4.4 Computing DIPDI for age estimation models without ground-truth

In the previous section we confirmed that age estimation is a task prone to be biased with respect to gender by computing error gaps between subgroups using ground-truth age annotations. Now, we want to study if it is possible to measure such bias proneness in a given population without ground-truth labels using DIPDI. We are interested in analyzing if output discrepancy for demography-aware model sets can be used as a proxy for anticipating potential fairness problems in specific ML tasks. To this end, we compute DIPDI in the same three settings where we explicitly evaluated biases in the previous section.

In all datasets, ChestX-ray14, UTKFace and IMDB-WIKI, we consider different scenarios of gender imbalance for the set of models $\mathbb{A}$ and $\mathbb{B}$ to be evaluated. First, we consider training populations consisting of only males (100-0), only females (0-100), and equal numbers of males and females (50-50). Two comparisons

$\mathbb{A}$, $\mathbb{B}$	ChestX-ray14	UTKFace	IMDB-Wiki
50-50, 50-50	-0.041 (0.056)	0.021 (0.030)	-0.106 (0.200)
100-0, 0-100	0.516 (0.078)	0.834 (0.107)	0.993 (0.309)

Table 1: DIPDI (mean ± std) for age estimation on ChestX-ray14, UTKFace and IMDB-WIKI according to group definition (rows) over 5 folds. Groups $\mathbb{A}$ and $\mathbb{B}$ represent two pairs of models. Each pair was trained on different male/female proportions: only males (0-100), only females (0-100), and equal males/females (50-50). Test folds are balanced with respect to the male-female ratio.

are made: *i*) 50-50 vs 50-50, representing cases where both $\mathbb{A}$ and $\mathbb{B}$ are groups of models trained with data from the same demographic sub-group; and *ii*) 100-0 vs 0-100, representing cases where $\mathbb{A}$ and $\mathbb{B}$ are groups of models trained with data from different demographic sub-group. To control for finite-size sampling variability, we split the training data into four random disjoint partitions, so that no data is shared between models even when they are trained for the same demographic sub-group. Then DIPDI is computed on the held-out test set (i.e. the unlabeled data $\mathcal{D}$), which is balanced by gender. Additional control experiments for DIPDI are included in Appendix A.3.

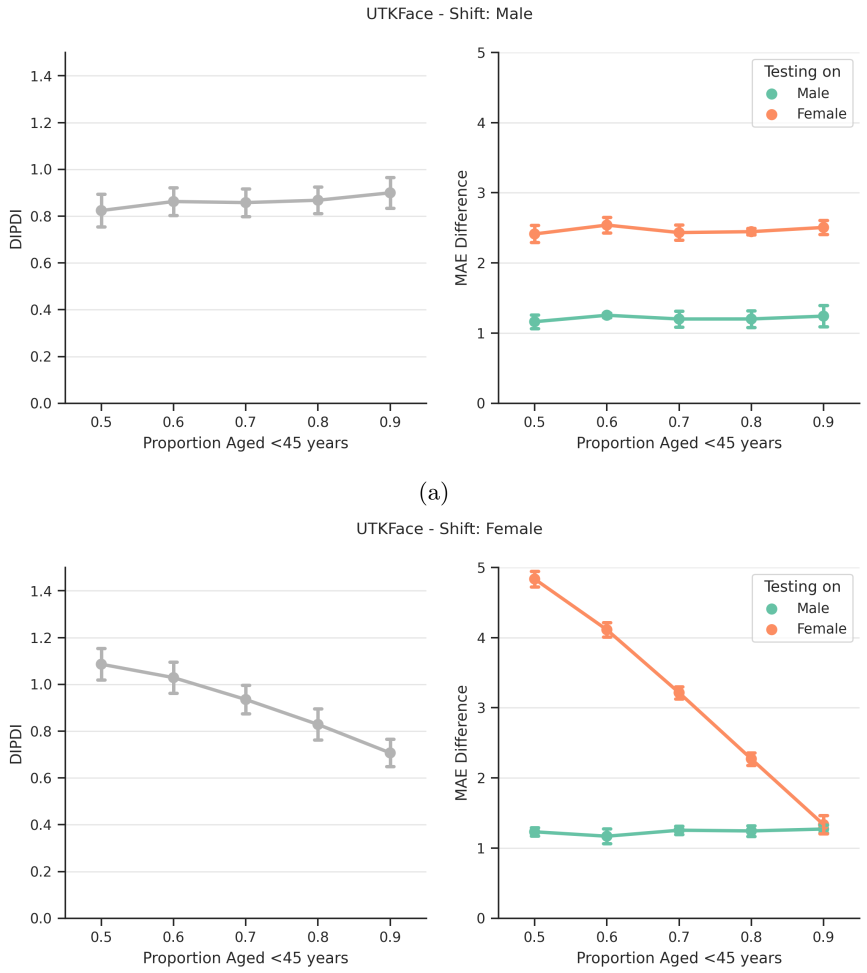
Table 4 shows DIPDI for age estimation on ChestX-ray14, UTKFace and IMDB-WIKI. Note that, in all scenarios, the index values are very close to 0 when comparing sets of models trained in the same population (row 1), but higher than 0 when comparing models from different populations (row 2), in line with the absence or presence of biases as a function of data imbalance shown in the previous section (recall Figure 2). Taken together these results demonstrate the co-occurrence between higher DIPDI and bias proneness: models that are less prone to bias, i.e. those coming from the same demographic population, produce more consistent outputs when evaluated on a target population. This output stability is clearly evidenced by index values close to 0 in row 1 for both datasets. In contrast, the index returns significantly higher values when it comes to models trained with different demographic subgroups, where biases are in turn more prone to appear, as shown in our previous supervised analysis (Section 4.3). Importantly, note that no labels were required in the target population $\mathcal{D}$ when computing DIPDI.

4.5 Anticipating potential demographic biases in domain shift scenarios with DIPDI

We have highlighted in the previous section the role of DIPDI in identifying potential demographic biases in populations that lack ground-truth annotations. Now, we turn our attention to a new challenge: demonstrating how DIPDI can deal with domain shift scenarios even when ground-truth data is unavailable. Prior research has identified the vulnerability of fairness properties of machine learning models when deployed on datasets differing from those used during model development Schrouff et al. (2022). In this context, we leverage DIPDI as an unsupervised alternative to traditional fairness metrics for understanding bias in new datasets with population shifts.

Our experiment involves a target population that is always balanced by gender, and we introduce shifts by altering the age distribution within one gender group (either male or female, but not both). Specifically, we increase the proportion of individuals with ages exceeding a predefined limit (set at 45 in our experiments), while maintaining the age distribution within the non-shifted group. For each shift scenario considered, we calculate both the DIPDI and the MAE difference between male and female models when tested separately on male and female subsets. For the male subset, we calculate the MAE difference by subtracting the MAE of a female-trained model from that of a male-trained model. Similarly, for the female subset, we subtract the MAE of a male-trained model from that of a female-trained model. Note that the calculation of MAE requires access to ground truth annotations, whereas DIPDI does not.

Figure 3 presents the mean and standard error of DIPDI and MAE difference for age shift ratios ranging from 0.5 to 0.9 (a shift ratio of 0.9, for example, implies 90% of the subpopulation is under 45, and 10% is at or above 45). Note that when the shift affects the male subgroup (Fig. 3a), DIPDI tends to slightly increase, and a corresponding slightly increasing difference is observed for both male and female test groups. On the other hand, when the shift involves the female subgroup (Fig. 3b), the difference decreases for females



(a)

(b)

Figure 3: Mean and standard error of DIPDI and MAE difference within domain shift scenarios for male (a) and female (b) subgroups on the UTKFace dataset. The shift is induced by modifying the age distribution of individuals under 45 years at increasing ratios for male and female test groups independently. Note that in both cases of shift, the DIPDI tends to follow the behaviour of the difference curve, showing an increase with increasing biases and a decrease with decreasing biases.

while remaining stable for males. In this case, DIPDI exhibits a corresponding decrease as biases within the female group decrease. These experiments underscore the effectiveness of DIPDI in domain shift scenarios, particularly when ground-truth annotations are unavailable. An increase in DIPDI during deployment, compared to the development phase, can be interpreted as an indicator of intensifying bias proneness within one or both demographic groups, whereas a decrease in DIPDI implies a potential reduction in bias within these groups.

5 Discussion

In this work, we tackle the issue of anticipating potential demographic biases at deployment in the absence of ground truth annotations. Typical methods designed to assess fairness require access to such annotations, which may be available at training time but not when deploying models for previously unseen data. A prototypical example of this would be a model trained on a public dataset which will then be applied to a local population for which we do not have the corresponding annotations. Recent work has highlighted how distribution shifts may affect fairness Schrouff et al. (2022), resulting in a potential risk. While an explicit fairness metric may not be computed, we argue that we can employ the discrepancy of output predictions as a proxy to provide an early warning about potential demographic biases. We propose a concrete solution in terms of an index, DIPDI, whose value indeed provides a measure for the proneness towards biased solutions.

Intuitively, we can think about output discrepancies in a set of models as a notion of uncertainty, similar to that estimated via ensemble variance Lakshminarayanan et al. (2017). In that sense, our index quantifies the relative uncertainty estimated when using models trained with data from different demographic groups (numerator) and from the same demographic group (denominator). If both are similar (ratio equal to 1), then we get a DIPDI value close to 0 (log ratio 1) indicating that the problem shows no early signs of bias with respect to the analyzed demographic values. However, higher discrepancies (uncertainty) for models from different demographic groups will lead to DIPDI values significantly larger than 0, indicating bias proneness for the task under analysis. In particular, an increase in DIPDI from model development (training) to deployment could be interpreted as a red flag, triggering further detailed assessment. We showed that DIPDI can also be used to understand how fairness transfers across distributions, particularly in scenarios involving population shifts where age changes differently for male and female groups.

Finally, we note that while we have expressed DIPDI here as a global population average, the same reasoning could in principle be applied to population subsets defined by the intersection of multiple demographic traits, or even on a subject-by-subject basis. Such predictive discrepancies as captured by DIPDI could serve to flag subjects or sub-groups at higher risk of suffering biases, constituting another avenue of research to explore in future work.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of Data and Analytics*, pp. 254–264. Auerbach Publications, 2016.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401–413, 2021.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.

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

Michael W Dusenberry, Dustin Tran, Edward Choi, Jonas Kemp, Jeremy Nixon, Ghassen Jerfel, Katherine Heller, and Andrew M Dai. Analyzing the role of model uncertainty for electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 204–213, 2020.

Yarin Gal et al. Uncertainty in deep learning. 2016.

Ben Glocker, Charles Jones, Melanie Bernhardt, and Stefan Winzeck. Algorithmic encoding of protected characteristics in image-based models for disease detection. *arXiv preprint arXiv:2110.14755*, 2021.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1820–1828, 2021.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. Fairness of classifiers across skin tones in dermatology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI*, pp. 320–329. Springer, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

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

Agostina J Larrazabal, César Martínez, Jose Dolz, and Enzo Ferrante. Orthogonal ensemble networks for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pp. 594–603. Springer, 2021.

Charles Lu, Andreanne Lemay, Katharina Hoebel, and Jayashree Kalpathy-Cramer. Evaluating subgroup disparity using epistemic uncertainty in mammography. *arXiv preprint arXiv:2107.02716*, 2021.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Eike Petersen, Aasa Feragen, Maria Luise da Costa Zemsch, Anders Henriksen, Oskar Eiler Wiese Christensen, Melanie Ganz, and Alzheimer's Disease Neuroimaging Initiative. Feature robustness and sex differences in medical imaging: A case study in mri-based alzheimer's disease detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I*, pp. 88–98. Springer, 2022.

M. Pividori, G. Stegmayer, and D.H. Milone. Diversity control for improving the analysis of consensus clustering. *Information Sciences*, 361:120–134, 2016.

Esther Puyol-Antón, Bram Ruijsink, Stefan K Piechnik, Stefan Neubauer, Steffen E Petersen, Reza Razavi, and Andrew P King. Fairness in cardiac mr image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 413–423. Springer, 2021.

María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. Addressing fairness in artificial intelligence for medical imaging. *nature communications*, 13(1):1–6, 2022.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Jessica Schrouff, Natalie Harris, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? *arXiv preprint arXiv:2202.01034*, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Rebecca S Stone, Nishant Ravikumar, Andrew J Bulpitt, and David C Hogg. Epistemic uncertainty-weighted loss for visual bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2898–2905, 2022.

Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2:8, 2019.

Lijing Wang, Dipanjan Ghosh, Maria Gonzalez Diaz, Ahmed Farahat, Mahbubul Alam, Chetan Gupta, Jiangzhuo Chen, and Madhav Marathe. Wisdom of the ensemble: Improving consistency of deep learning models. *Advances in Neural Information Processing Systems*, 33:19750–19761, 2020.

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

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5810–5818, 2017.

A Appendix

A.1 DIPDI on synthetic data

Figure 4 illustrates the stability of DIPDI with the target population dataset size ($N = |\mathcal{D}|$) based on its behavior under controlled conditions using synthetic data. To do this, we simulated predictions from two model sets, $\mathbb{A}$ and $\mathbb{B}$, and evaluated DIPDI in scenarios with varying levels of disagreement between these models, represented by stochastic discrepancies in their output predictions (see Figure 1 in the main manuscript). We found consistent stability in the DIPDI index across scenarios of similar, discrepant, and highly discrepant outputs, particularly when N exceeded 50.

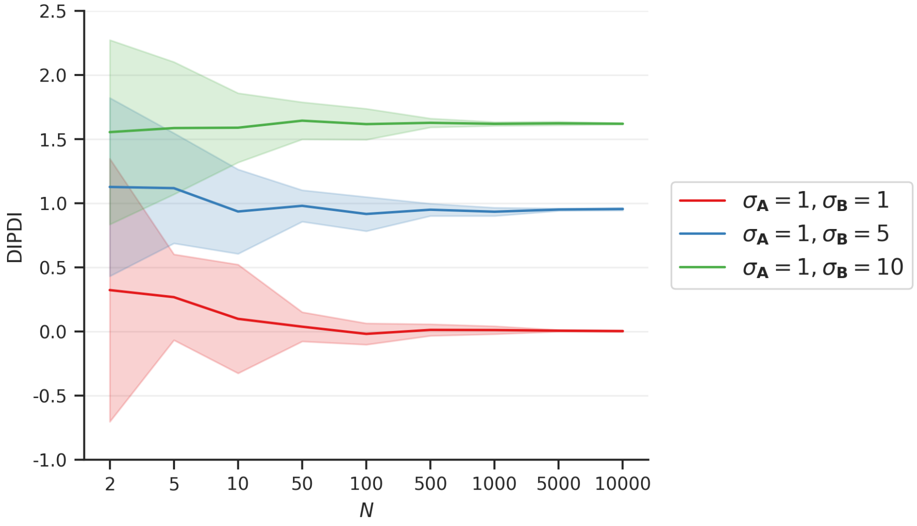


Figure 4: DIPDI stability in terms of the number of data samples N (the size of $\mathcal{D}$). Results are based on 10 runs for each pair of $\sigma_{\mathbb{A}}$ and $\sigma_{\mathbb{B}}$, and displayed for three scenarios: no discrepancy (red), lower discrepancy (blue) and higher discrepancy (green).

A.2 Controlled subgroup analysis for age estimation

In this section, we provide additional results on for ChestX-ray14, UTKFace, and IMDB-WIKI datasets, aiming to assess the impact of gender imbalance on age estimation, as discussed in Section 4.3 of the main manuscript.

A.2.1 Mean absolute error (ChestX-ray14 and UTKFace)

Tables 2 and 3 present results for ChestX-ray14 and UTKFace, reporting the mean absolute error (MAE) values for male and female subgroups under different scenarios of gender imbalance in the training data.

Training (Male-Female)	Testing on Male	Testing on Female
Male (100-0)	4.508 (0.028)	5.602 (0.108)
Mixed (50-50)	4.634 (0.061)	4.632 (0.177)
Female (0-100)	5.209 (0.038)	4.432 (0.057)

Table 2: Mean absolute error (MAE) (mean ± std) for age estimation on ChestX-ray14 by subgroup (male and female) across 5 folds, using trained models with different degrees of male-female imbalance: 100-0, 50-50, and 0-100.

Training (Male-Female)	Testing on Male	Testing on Female
Male (100-0)	5.950 (0.059)	8.534 (0.418)
Mixed (50-50)	6.200 (0.172)	5.757 (0.172)
Female (0-100)	7.139 (0.206)	5.446 (0.204)

Table 3: Mean absolute error (MAE) (mean ± std) for age estimation on UTKFace by subgroup (male and female) across 5 folds, using trained models with different degrees of male-female imbalance: 100-0, 50-50, and 0-100.

A.2.2 Cumulative score (ChestX-ray14 and UTKFace)

Figures 5 and 6 show cumulative score (CS) values for male and female subgroups in different gender imbalance scenarios in the training data. The CS quantifies the proportion of test samples (N) for which the absolute error e falls below a specified threshold of n years. This calculation is defined as follows:

$$CS(n) = \frac{N_{e \leq n}}{N},$$

where $N_{e \leq n}$ represents the number of test images for which the absolute age error is less than or equal to the corresponding threshold value.

A.2.3 Mean absolute error (IMDB-WIKI)

Figure 7 shows the results of the MAE corresponding to a 20-fold cross-validation for models trained with different degrees of gender imbalance and evaluated on males, females, and the whole population. This is a more fine-grained analysis than the one presented in Figure 2c of the main manuscript, aiming to observe the effect of gender imbalance in more detail.

A.3 Computing DIPDI for age estimation models without ground-truth

Table 4 presents additional control experiments for DIPDI on ChestX-ray14, UTKFace and IMDB-WIKI datasets as discussed in Section 4.4 of the main manuscript. We evaluated DIPDI in the same scenarios of gender imbalance for the set of models $\mathbb{A}$ and $\mathbb{B}$ as previously examined biases: only males (100-0), only

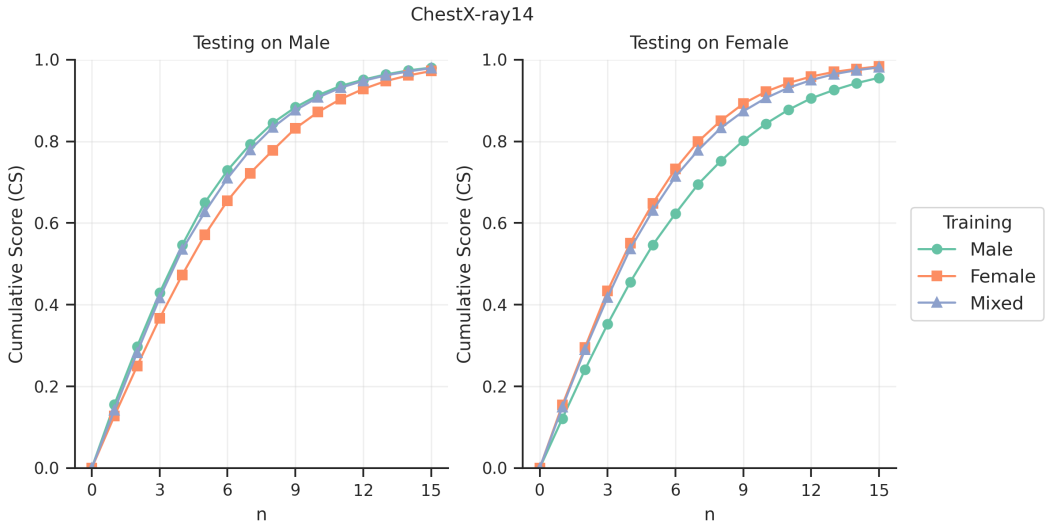


Figure 5: Cumulative score (CS) for age estimation on ChestX-ray14 by subgroup (male and female) aggregated across all 5 folds, using trained models with different degrees of male-female imbalance: 100-0, 50-50, and 0-100. The age threshold n ranges from 0 to 15 years.

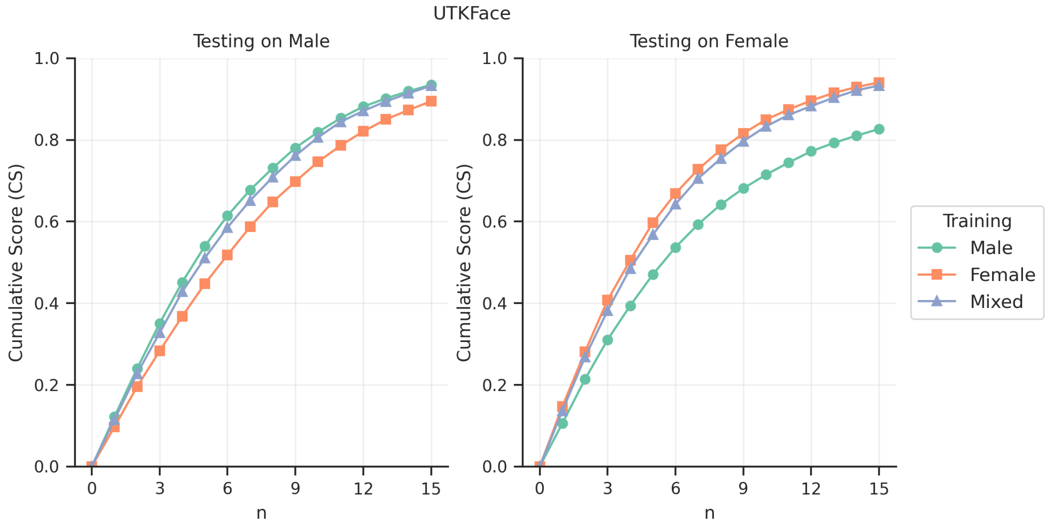


Figure 6: Cumulative score (CS) for age estimation on UTKFace by subgroup (male and female) aggregated across all 5 folds, using trained models with different degrees of male-female imbalance: 100-0, 50-50, and 0-100. The age threshold n ranges from 0 to 15 years.

females (0-100), and equal numbers of males and females (50-50). We observe that models from the same population (rows 1 to 3) tend to produce DIPDI values close to 0, whereas models trained with different demographic subgroups (row 4) exhibit higher DIPDI values, indicating a greater propensity to bias.

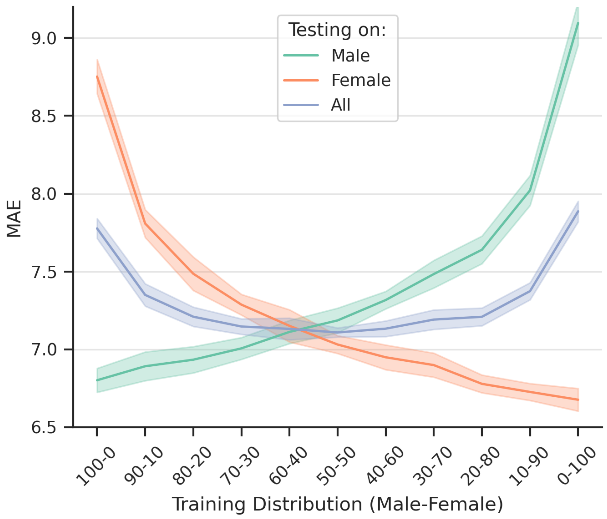


Figure 7: Mean absolute error (MAE) for age estimation on IMDB-WIKI. Trained models with different degrees of male-female imbalance and evaluated on males (green), females (orange), and the whole population (purple).

$\mathbb{A}$, $\mathbb{B}$	ChestX-ray14	UTKFace	IMDB-Wiki
100-0, 100-0	0.004 (0.038)	-0.007 (0.115)	0.053 (0.110)
50-50, 50-50	-0.041 (0.056)	0.021 (0.030)	-0.106 (0.200)
0-100, 0-100	0.031 (0.081)	0.013 (0.180)	-0.104 (0.154)
100-0, 0-100	0.516 (0.078)	0.834 (0.107)	0.993 (0.309)

Table 4: DIPDI (mean ± std) for age estimation on ChestX-ray14, UTKFace and IMDB-WIKI datasets according to group definition (rows) over 5 folds. Groups $\mathbb{A}$ and $\mathbb{B}$ represent two pairs of models. Each pair was trained on different male/female proportions. Test folds are balanced with respect to the male-female ratio.