
Addressing fairness issues in deep learning-based medical image analysis: a systematic review

Zikang Xu, Jun Li, Qingsong Yao, Han Li, Mingyue Zhao & S. Kevin Zhou
(School of Biomedical Engineering, University of Science and Technology of China, Hefei, Anhui, China)

DOI: 10.1038/s41746-024-01276-5

Review article

npj Digital Medicine

Published: October 2024

Overview

- **Fairness in ML model** = attempts to correct algorithmic bias in automated decision processes
- **Group fairness in DL model** = equal utilities for all the subgroups in the test set

⇒ **sensitive attributes depending the imaging modality**: sex, age, race, skin tone, etc.

⇒ problem of performance consistency and reliability for clinical applications : most of the methods only focus on the **diagnosis performance while ignoring whether the algorithm has biased or unfair utilities towards different subgroups**

- **Goals of this review:**
 - 1) Introducing fundamental fairness concepts
 - 2) Categorizing existing studies about fair MedAI

4 main criteria of Fairness

Table 1 | Widely Used Criteria for Fairness

Metrics	Formula ^a	Explanation ^b
Demographic Parity (DP) ³³	$P(\hat{Y} = 1 A = 0) = P(\hat{Y} = 1 A = 1)$	The model outcome should not be affected by any sensitive attribute.
Accuracy Parity (AP) ¹¹⁸ Independence	$P(\hat{Y} = Y A = 0) = P(\hat{Y} = Y A = 1)$	The model should have an equal accuracy among subgroups. $R \perp A$
Equalized Odds (EqOdds) ¹¹⁹ Separation	$P(\hat{Y} = 1 A = 0, Y = y) = P(\hat{Y} = 1 A = 1, Y = y), y \in \{0, 1\}$	The model should have an equal TPR and FPR among subgroups. $R \perp A \mid Y$
Equal Opportunity (EqOpp) ¹¹⁹	$P(\hat{Y} = 1 A = 0, Y = 1) = P(\hat{Y} = 1 A = 1, Y = 1)$	The model should have an equal TPR among subgroups.

^a $Y, \hat{Y} \in \{0, 1\}$ denotes the ground truth label and model prediction, respectively. $A \in \{0, 1\}$ denotes the sensitive attribute. Note that Y and A can be easily extended to multi-class situations.

^bTPR true positive rate, FPR false positive rate.

Sufficiency: $Y \perp A \mid R$

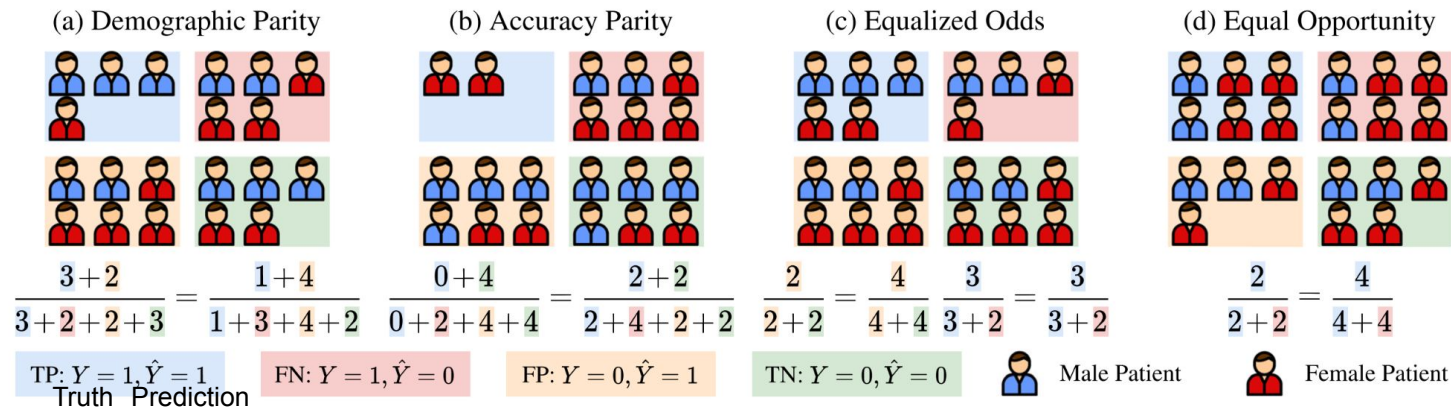


Fig. 1 | Ideal situations where various fairness criteria are satisfied. From left to right: **a** Demographic Parity, **b** Accuracy Parity, **c** Equalized Odds, **d** Equal Opportunity. The equations below compute the value of different criteria for the Male and Female groups.

Total fairness is not possible to achieve except in specific rhetorical cases

Overall fairness

- Proper fairness criteria depends on the specific task
- Overall fairness quantified by combining several group-wise fairness metrics
- Model might satisfy fairness constraints in one grouping scheme and not another :

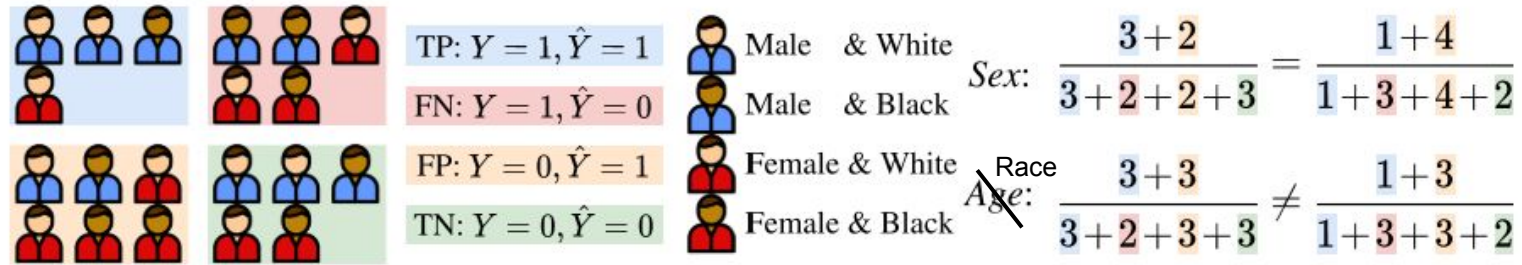


Fig. 2 | In a scenario involving two sensitive attributes, namely *sex* (male, female) and *race* (White, Black), demographic parity is achieved concerning *sex* but not *race*.

DP

Review scope

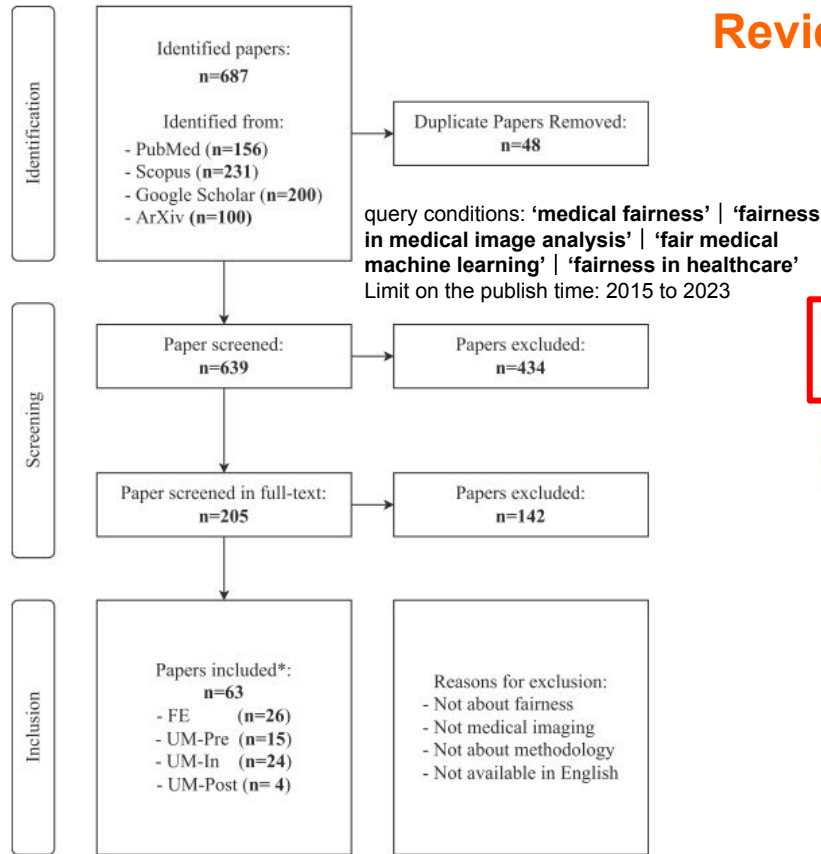


Fig. 3 | PRISMA diagram for this review. * denotes that six studies have been overcounted due to their involvement in research across multiple directions. FE fairness evaluation, UM unfairness mitigation, Pre pre-processing, In in-processing, Post post-post-processing.

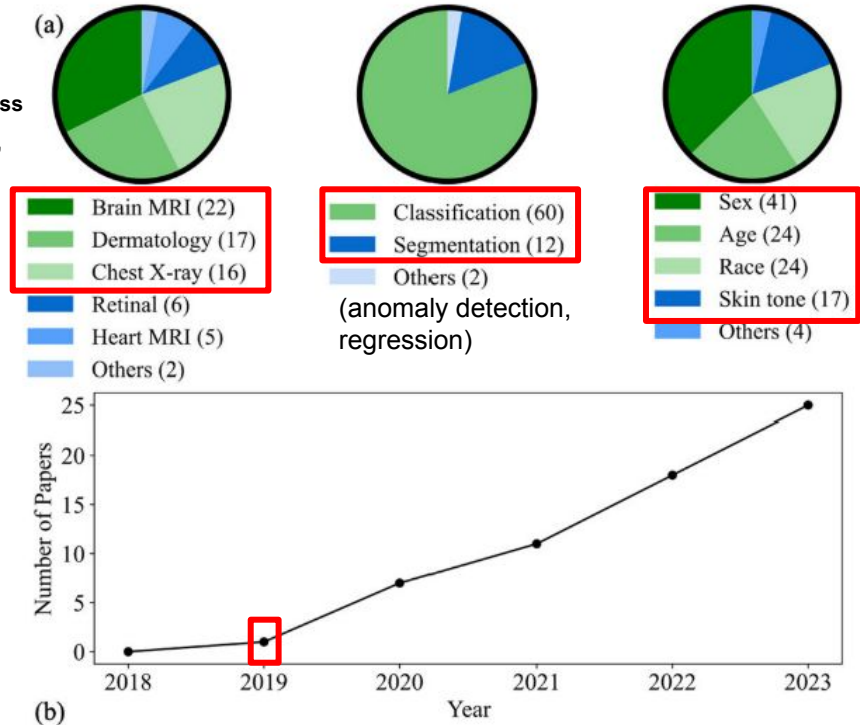


Fig. 4 | Summary of data extracted from studies in our systematic review.
a Annual trends in research on fairness in MedIA. **b** Prevalence of various medical imaging modalities, research tasks, and associated sensitive attributes.

Overview of Studies selected

Table 2 | Overview of Studies in Fair Media

Research area ^a	Year	Citation	Imaging modality	Dataset	Task ^b	Sensitive attributes	Fairness metrics ^c
FE-benchmarking	2020	13	Chest X-ray	NIH Chest-XRay14, CheXpert	C	Sex	Δ AUC
	2020	5	Chest X-ray	MIMIC-CXR, ChestXray8, CheXpert	C	Age, Race, Sex, etc.	Δ AUC, Δ TPR
	2020	47	Dermatology	ISIC, SD-198	C	Skin tone	AP
	2021	54	Chest X-ray	CheXpert, MIMIC-CXR	C	Race, Sex	Δ AUC, Δ BCE, Δ ECE, Δ TPR, Δ TNR
	2021	2	Chest X-ray	MIMIC-CXR, CheXpert, ChestXray14	C	Age, Race, Sex	Δ FPR, Δ FNR
	2022	53	Multiple	Multiple datasets	C	Age, Race, Sex, Skin tone	Max-Min Fairness, Δ AUC, Δ BCE, Δ ECE, Δ TPR, Δ FPR, Δ FNR, EqOdds
	2022	8	Brain MRI	ADNI	S	Race, Sex	Δ DSC

^aFE fairness evaluation, UM unfairness mitigation.

^bC classification, S Segmentation, R Regression, AD anomaly detection.

^cTPR true positive rate, TNR true negative rate, FPR false positive rate, FNR false negative rate, PPV positive predictive value, NPV negative predictive value, AUC area under curve, ACC accuracy, BACC balanced accuracy, BCE binary cross-entropy, ECE expected calibration error, MAE mean absolute error, IOU Intersection over Union, DSC Dice similarity coefficient, RMSE Root Mean Squared Error, DPM Division form of DP, EOM Division form of EqOpp.

Results for MRI

- Petersen et al., 2022: performance of the under-represented group (gender) is not significantly different from that of the over-represented group, ADNI dataset
- Ioannou et al., 2022: unfairness evaluation should be conducted on different components respectively rather than on the overall utility (DSC)

Source of unfairness ?

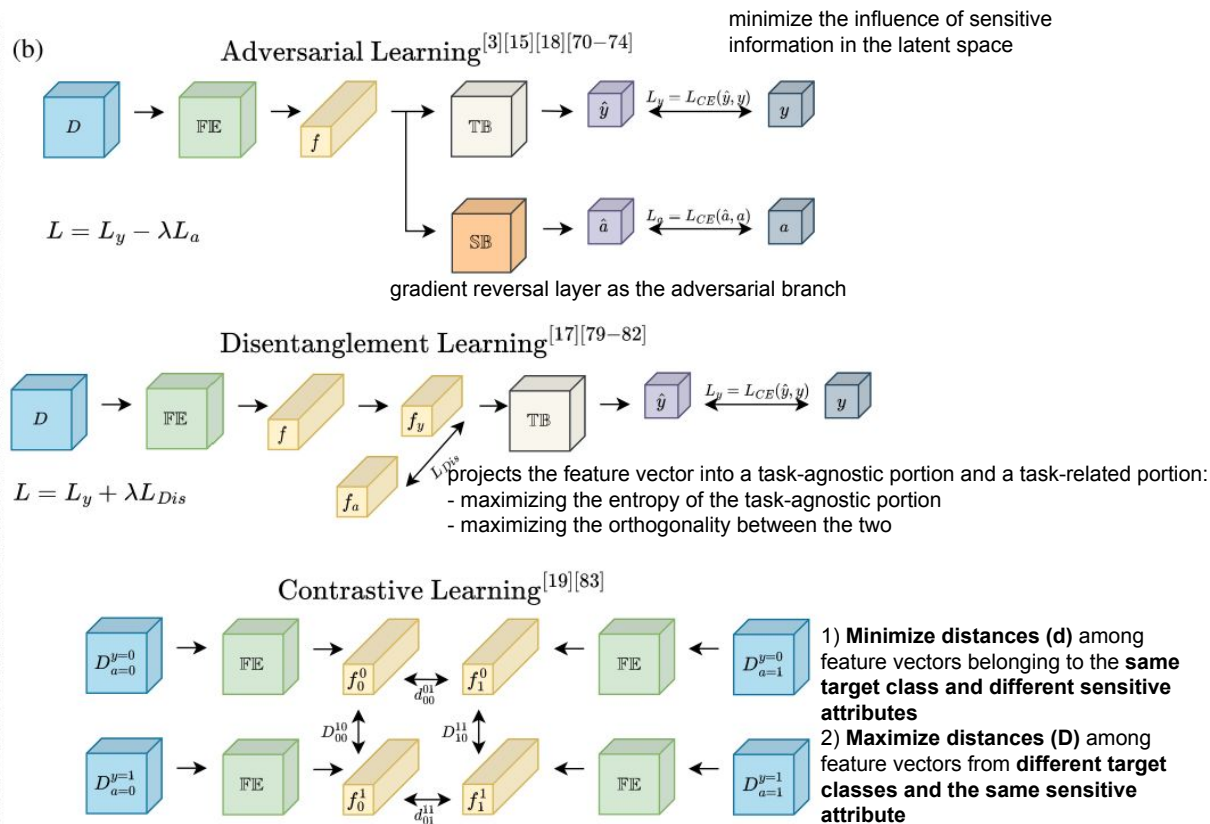
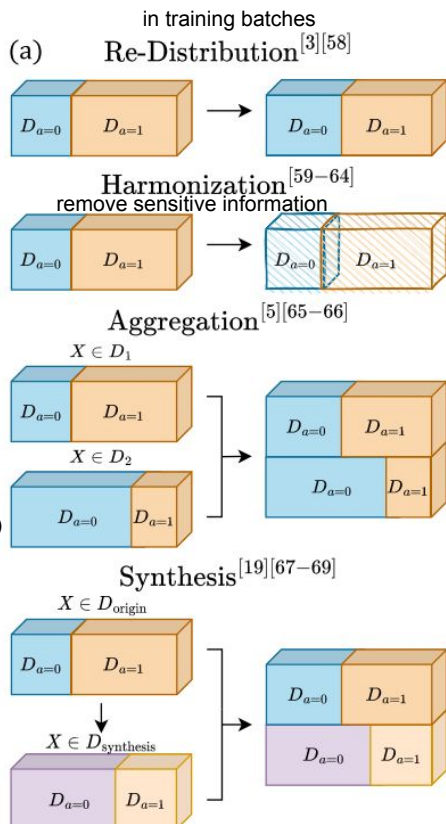
- Duet al., 2023: Total Intracranial Volume and normalized Whole Brain Volume (MRI reconstruction tasks)
- Petersen et al., 2022: Inherent differences in the pathology of the two sexes

Remark:

- DL models might be fair on sex and age, respectively, but can still be unfair when evaluated on the combination of sex and age

Unfairness mitigation

- pre-processing (a), in-processing (b), and post-processing (c)



(FairAdaBN: extra adapters, FairTune: fine-tuning ...)

Unfairness mitigation

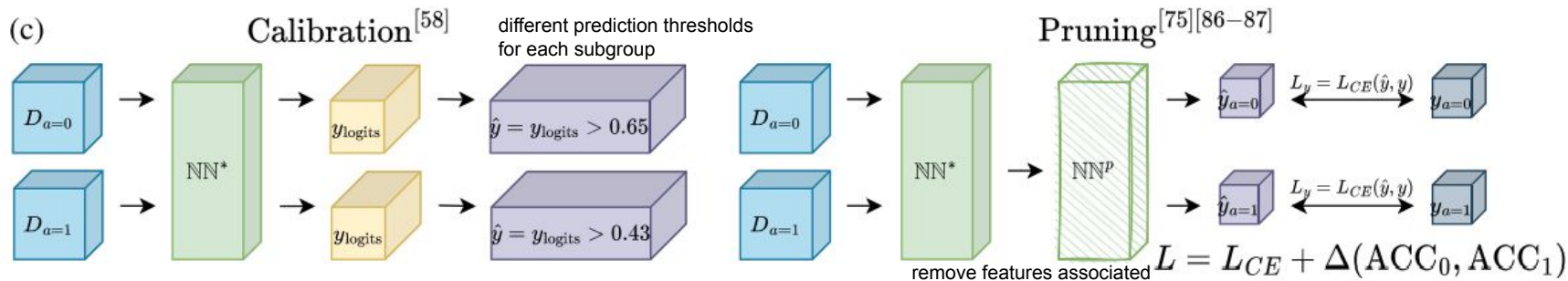






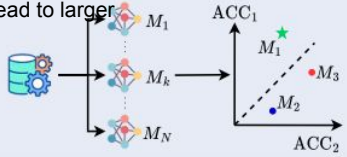
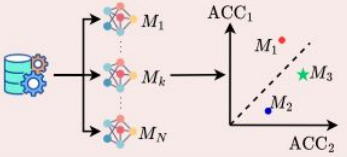
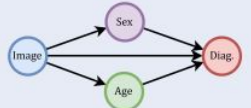
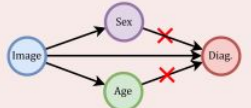


Fig. 6 | Schematic diagram of unfairness mitigation algorithms. **a** Pre-processing methods. D_1, D_2 : Two independent datasets; $D_{\text{origin}}, D_{\text{synthesis}}$: The original dataset and synthesized dataset. **b** In-processing methods. FE, TB, SB: Feature Extractor, Target Branch, and Sensitive Branch, which are three parts of an adversarial network; f : latent feature vector; y, a, \hat{y}, \hat{a} : the ground truth target task label, sensitive attributes, and their corresponding predictions generated by the neural network; \mathcal{L}_{CE} : Cross-Entropy loss, measuring the difference between the predicted label and

the ground truth label; \mathcal{L}_{Dis} : Disentanglement loss, for example, MMD-Loss¹⁷⁰, measuring the distance between two distribution; D_{00}^{10} : requiring the maximum distance between f_0^0 and f_1^1 ; d_{00}^{01} : requiring the minimum distance between f_0^0 and f_1^1 . **c** Post-processing methods. NN^* : a pre-trained and fixed Neural Network; y_{logits} : predicted probability of y , range from 0 to 1; NN^p : pruned NN^* ; $\Delta(ACC_0, ACC_1)$: difference between accuracy on subgroup test set $D_{a=0}$ and $D_{a=1}$.

Potential solutions for different sources of unfairness

	Sources of Unfairness	Potential Solutions
Data	<p>Skewed Distribution</p> 	<p>Data Augmentation</p> 
	<p>Anatomy Difference</p>  <p>Female Male</p>	<p>Causal Image Synthesis</p>  <p>Female "Male"</p> <p>$\text{Sex} : F \rightarrow \text{Sex} : M$</p>
	<p>Annotation Noise</p>  <p>HR 110 HR 110 High HR Normal HR</p>	<p>Multi-Annotator Involvement</p>  <p>HR 110 Normal HR High HR High HR</p>
Model	<p>Empirical Risk Minimization</p> <p>higher performance might lead to larger subgroup gaps</p> <p>ERM-Based Model Selection</p>  <p>M_1, M_k, M_N ACC_1, ACC_2</p>	<p>Dipper throated optimization</p> <p>DTO-Based Model Selection</p>  <p>M_1, M_k, M_N ACC_1, ACC_2</p>
	<p>Spurious Correlation</p>  <p>Sex Age Diag</p> <p>irrelevant correlations between the input image and the output diagnosis</p>	<p>Confounder Removal</p>  <p>Sex Age Diag</p>

Potential solutions for different sources of unfairness

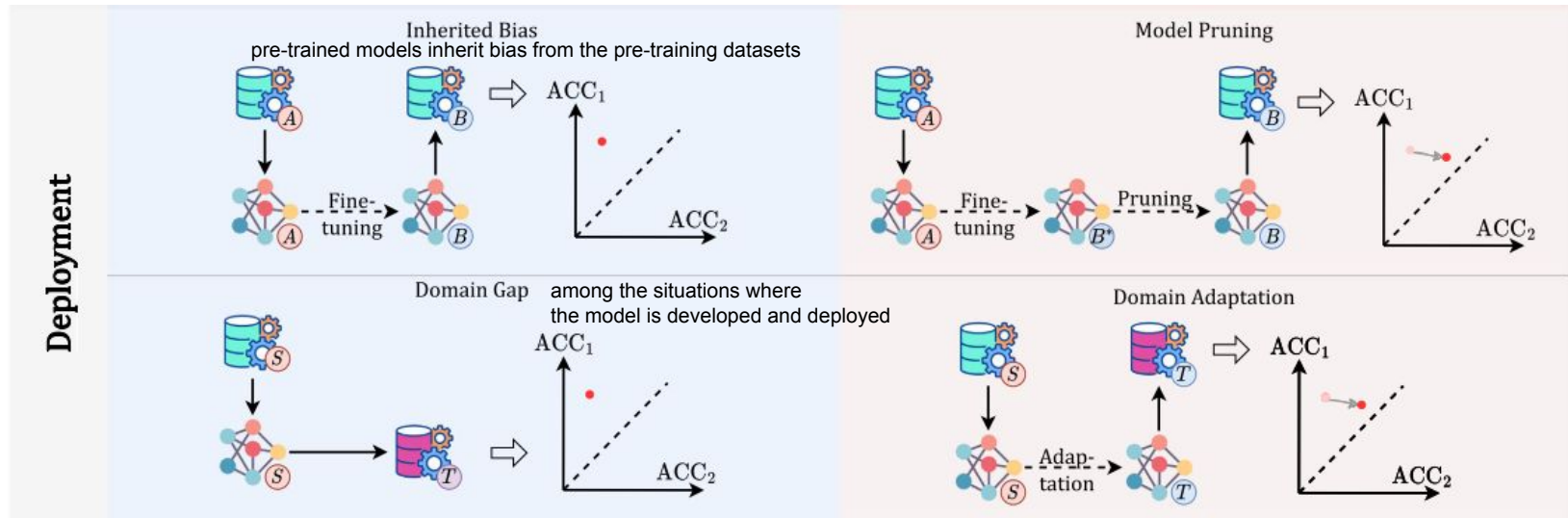


Fig. 7 | Sources of unfairness and potential solutions. From top to bottom: skewed data distribution → aggregate data from multiple datasets; anatomy difference between subgroups → ; annotation differences for each subgroup → using causal image synthesis methods to transfer the input to the synthesis ones with opposite attribute; annotation noise → involve multi-annotators to stabilize annotation; ERM-based model selection which chooses models with the highest overall

performance → DTO-based model selection which consider both performance and fairness; spurious correlations between sensitive attributes and diagnosis → removing the effects of the confounder; inherited bias from the pre-train dataset → pruning the pre-trained with fairness constraints; domain gaps between the source dataset and target dataset → using domain adaptation methods to transfer models.

Difference in paradigm

AI scientists : does the model have a biased outcome on this attribute?

Clinicians : will the anatomical difference between the male and the female affect the diagnosis difficulty?

Clinicians regard some of the metrics as unreasonable:

demographic parity requires that the patients from each subgroup should have the same probability of being predicted as ailing (many illnesses are proven to be related to age or sex)

When anatomical differences: good to have **unequal diagnosis precision** due to **reasonable medical prior**

Fairness metrics usually fluctuate along the training procedure and are hard to converge

Most of the **large-scale foundation models and language models** suffer from different levels of unfairness, due to domain gap, annotation noise, spurious correlation, or inherited bias from the training set

Conclusion

- **AI scientists, ethicists, and clinicians** should cooperate to identify sources of unfairness, and develop strategies to mitigate them
- government can also **incorporate fairness considerations into clinical AI guidelines** to improve the awareness of fairness in the whole pipeline

Thank you for your attention !