

Peer Review

Review of: "Fair Diagnosis: Leveraging Causal Modeling to Mitigate Medical Bias"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper addresses how sensitive attributes influence the predictions of medical image models (Abstract, Section 1). The authors use a structural causal model with sensitive attribute S , true diagnosis Y , input image X , and predicted diagnosis $\hat{Y}$ (Figure 1). The effect of S passing through the true diagnosis Y is treated as legitimate, while the direct path from S to the prediction is treated as bias to remove. Diagnosis Fairness is defined as $DE(S)=0$, a zero direct effect (Section 2.2, Figure 1).

In the method, a generator adds adversarial perturbation masks to input images, and a discriminator tries to predict the sensitive attribute from them given the diagnosis. An entropy regularizer makes the discriminator guess randomly, and the deployed classifier stays frozen after pretraining (Section 3.2, Figure 2). The metric Approximate Diagnosis Fairness (ADF) is the conditional mutual information $I(S; \hat{Y} | Y)$ (Eq. 24, Section 4.1.2), and Theorems 1 and 2 relate the generator objective and ADF to DF (Section 3.3). Experiments use MIMIC-CXR, CheXpert, and TCGA-LUAD (Section 4.1.1), with gender as the only sensitive attribute. The tasks are pneumonia classification on the two chest X-ray sets and early versus late stage lung cancer on TCGA (Appendix A.2). Baselines are Vanilla, Adversarial Debiasing (AD), and FAAP (Section 4.2). The authors conclude that the method gives better fairness than AD without sacrificing much accuracy, and better accuracy and fairness than FAAP (Section 4.3).

2. Whether the conclusion holds on the data

The most useful idea in this paper is how it treats the sensitive attribute. The effect that passes through the true diagnosis is allowed, and only the direct effect on the prediction is removed (Section 2.2, Figure 1). That fits clinical reality better than asking for equal rates everywhere. I also like that the deployed

classifier stays frozen, so the debiasing works without changing its parameters (end of Section 3.2). The evaluation is wide: three datasets, one of them a pathology staging task, three baselines, four ablation tables (Tables 2 to 5), and a Grad-CAM comparison (Figure 4).

Table 1 supports the headline claim. On MIMIC, Ours vs. Vanilla is ADF 2.25 vs. 10.59 (e-3), EO 2.02 vs. 5.10 (e-2), and ACC 84.26 vs. 85.46. On CheXpert, it is ADF 1.20 vs. 11.18, EO 0.98 vs. 5.05, and ACC 85.80 vs. 85.87. So the fairness gains are large, and the accuracy cost is small. EO and DP are standard metrics that don't depend on the method, and they improve in Table 1 as well. Section 4.3 says that on TCGA, ADF is "lower by 5 times" and EO and DP "approximately lower by 2 times." From Table 1, I get $ADF\ 7.21/1.46 \approx 4.9$, $EO\ 3.57/1.39 \approx 2.6$, and $DP\ 11.94/6.89 \approx 1.7$, so I would quote the ratios as computed.

The weaker part is the strength of the evidence. Every number comes from a single training run on one 90%/10% split (Appendix A.1), and I found no standard deviations, confidence intervals, seeds, or repeated runs in the text. Several accuracy gaps are within one point, for example, MIMIC ACC Ours 84.26 vs. AD 83.90 (Table 1), so the ranking of methods on accuracy can't be read from one run. I would add repeated runs with different seeds and report the spread. It would also help to give test set sizes, class prevalence, and whether the split is at the patient level. On MIMIC, Vanilla has ACC 85.46 but AUC 63.85 (Table 1), which suggests the accuracy is driven by the majority class. A per-group sensitivity or a balanced accuracy would say more here.

ADF is $I(S; \hat{Y} | Y)$ (Eq. 24), the prediction-level counterpart of the generator objective $I(S; X | Y) = 0$ (Eq. 8, Section 3.2), and Section 4.3 itself says the ADF result "aligns with the objective of our method". ADF is therefore closely tied to the training objective, which is why EO and DP carry the weight as independent evidence. I would say this openly in the paper and lead the fairness comparison with EO and DP. Also, only gender is used as the sensitive attribute in all three datasets (Appendix A.2), while the Abstract and Introduction speak of race and gender. I would either add a race experiment or narrow the claim in those two places.

Finally, there is one inconsistency I would ask the authors to resolve. The Table 1 caption says all models are ResNet50, but the "Ours" rows for MIMIC and CheXpert are identical to the Resnet18 rows of Table 2 (details in section 4 of this review). Until this is clarified, I can't be sure which configuration produced the headline numbers, so I would read the conclusion with some caution.

3. Reproducibility

I searched the text for the reproducibility details and found a good deal of them. Appendix A.1 reports $\eta = 0.2$, $\alpha = 1.0$, $\beta = 1.0$, a learning rate of 1×10^{-4} with step-based schedulers, and a 90%/10% train/test split. Appendix A.2 gives the task labels, the stage mapping for TCGA, and the gender coding (male = 0, female = 1). The computing environment in Appendix A.3 is stated better than in most papers: PyTorch, an AMD EPYC 7763 CPU, an NVIDIA RTX 6000 Ada GPU, Ubuntu 22.04.3 LTS, Miniconda 23.9.0, Python 3.10.13, and Docker 24.0.5 with the base image nvidia/cuda:12.4.0-cudnn8-devel-ubuntu22.04. Anyone who wants to rebuild the setup can start from this list, and I thank the authors for it.

For the code, Appendix A.3 says "We will provide ... upon publication" and names the code, the Conda environment, and the Docker container. Nothing is available yet for the Qeios version.

Some details that a reader would need to repeat the experiments are still missing. I could not find the random seeds or the number of runs, the batch size, or the number of epochs. The image resolution and preprocessing are not given, and neither are the architectures of the generator, the discriminator, and the MI estimator. The text does not say whether the split is made at the patient level, and the number of images and patients used is not reported. It would help to explain how ADF (Eq. 24) is estimated from the test predictions, and to list the settings of the AD and FAAP baselines. One more question concerns TCGA. Section 4.1.1 describes TCGA-LUAD by its gene expression, mutation, methylation, and clinical data, yet Figure 4 shows Grad-CAM heatmaps on TCGA-LUAD. The image modality used for TCGA is not stated, so I would add it to the paper.

4. What specifically to fix

1. Backbone mismatch. The Table 1 caption says "All models are pretrained and fine-tuned on ResNet50", and Section 4.6.1 repeats "For the remaining experiments, we use ResNet50". But the Table 1 "Ours" rows for MIMIC (84.26, 62.10, 2.02, 3.53, 2.25) and CheXpert (85.80, 64.99, 0.98, 1.16, 1.20) are identical to the Resnet18 rows of Table 2. The Resnet50 rows of Table 2 are MIMIC 84.67, 62.46, 1.78, 2.39, 1.19 and CheXpert 86.40, 66.34, 1.12, 1.41, 1.34. The Table 1 "Ours" row for TCGA (97.85, 98.20, 1.39, 6.89, 1.46) matches neither Table 2 Resnet18 (97.25, 98.20, 1.89, 7.06, 1.64) nor Resnet50 (97.87, 98.23, 1.40, 6.95, 1.51). Please state which backbone produced Table 1 and correct the caption or the rows.

2. Several default rows of the ablation tables do not agree with Table 1. The defaults are $\eta = 0.2$, $\alpha = 1$, $\beta = 1$ (Appendix A.1), so I would expect these rows to repeat the main result. Some do: the $\eta = 0.2$ rows of Table 3 for MIMIC and TCGA and the $\beta = 1$ TCGA row of Table 5 match Table 1 exactly. Others do not. In Table 3, $\eta = 0.2$, CheXpert gives 84.59, 63.11, 3.47, 2.64, 3.24, against Table 1 CheXpert "Ours" 85.80, 64.99, 0.98, 1.16, 1.20. In Table 4, $\alpha = 1$, MIMIC gives 86.46, 62.79, 4.50, 2.54, 1.62, against Table 1 MIMIC "Ours" 84.26, 62.10, 2.02, 3.53, 2.25. In Table 4, $\alpha = 1$, TCGA gives 98.70, 98.51, 1.89, 7.74, 3.42, while the Table 1 TCGA fairness values 1.39, 6.89, 1.46 appear in Table 4 at $\alpha = 4$. In Table 5, $\beta = 1$, MIMIC gives 84.98, 62.34, 1.39, 2.78, 1.87, and CheXpert gives 84.59, 65.13, 3.32, 2.02, 1.54. It would help to state, under each ablation table, the backbone and the fixed values of the other two hyperparameters.
3. Section 4.6.2 says that as α increases "we observe a gradual decrease in performance", but in Table 4 MIMIC ACC goes from 86.53 ($\alpha = 0$) to 86.92 ($\alpha = 3$), and CheXpert ACC goes from 85.41 ($\alpha = 0$) to 85.60 ($\alpha = 4$). Also, Table 4 MIMIC $\alpha = 0$ ACC of 86.53 is above Vanilla 85.46 in Table 1. A sentence explaining this would help.
4. In Table 5, $\beta = 0$ gives TCGA ACC 72.70 with AUC 92.01, and CheXpert ACC 71.15. Such a gap between accuracy and AUC deserves a comment (decision threshold, class balance).
5. On variance: all results are single runs on one 90%/10% split (Appendix A.1). I would repeat Table 1 with several seeds, or give bootstrap confidence intervals on the test set, and use a patient-level split.
6. Please report the test set sizes, the class prevalence per gender group, and a per-group sensitivity or balanced accuracy. On MIMIC, for example, Vanilla has ACC 85.46 with AUC 63.85 (Table 1), and these numbers would help the reader interpret that pair.
7. For TCGA-LUAD (Section 4.1.1, Figure 4), I would add the image modality, the number of cases, and how the images were prepared.
8. For ADF (Eq. 24), please describe how $I(S; \hat{Y} | Y)$ is estimated from the test predictions. I would also say openly in Section 4.3 that it is closely tied to the training objective $I(S; X | Y) = 0$ (Eq. 8), so that EO and DP carry the independent evidence.
9. The experiments use only gender as the sensitive attribute (Appendix A.2), while the Abstract mentions race and gender. Either add a second attribute or narrow the Abstract.
10. Figure 4 and Figure 3. The Grad-CAM comparison is qualitative; a quantitative overlap measure between the two heatmaps over the whole test set would make the claim of "clear separation" (Section 4.5) testable. For data utility (Section 4.4, Figure 3: "most tasks achieve an AUC of 65% and

an accuracy of 80%"), it would help to add the same downstream tasks without the MI pretraining as a reference. Finally, I would be glad to see the code released, as promised in Appendix A.3.

5. Recommendation

Major revisions. I think the central idea is sound and useful for clinical imaging: removing only the direct effect of the sensitive attribute while keeping its effect through the true diagnosis (Section 2.2). The fairness gains in Table 1 are large on EO and DP as well as on ADF, which is encouraging. Still, the headline numbers in Table 1 do not match the backbone stated in its caption (Table 1 vs Table 2), and several default rows of Tables 3 to 5 disagree with Table 1. It would also help to add repeated runs, since all results are single runs without variance (Appendix A.1). These points can be fixed with the authors' own logs and a few repeated runs, and I expect the main conclusion may well survive them. I would be glad to read a revised version.

Declarations

Potential competing interests: No potential competing interests to declare.