On the Impact of Random Seeds on the Fairness of Clinical Classifiers

Silvio Amir and **Jan-Willem van de Meent** and **Byron C. Wallace**

Northeastern University

{s.amir,j.vandemeent,b.wallace}@northeastern.edu

Abstract

Recent work has shown that fine-tuning large networks is surprisingly sensitive to changes in random seed(s). We explore the implications of this phenomenon for model fairness across demographic groups in clinical prediction tasks over electronic health records (EHR) in MIMIC-III —— the standard dataset in clinical NLP research. Apparent subgroup performance varies substantially for seeds that yield similar overall performance, although there is no evidence of a trade-off between overall and subgroup performance. However, we also find that the small sample sizes inherent to looking at intersections of minority groups and somewhat rare conditions limit our ability to accurately estimate disparities. Further, we find that jointly optimizing for high overall performance and low disparities does not yield statistically significant improvements. Our results suggest that fairness work using MIMIC-III should carefully account for variations in apparent differences that may arise from stochasticity and small sample sizes.

1 Introduction

Fine-tuning pre-trained transformers (Vaswani et al., 2017) such as BERT (Devlin et al., 2019) has become the dominant paradigm in NLP, owing to their performance across a range of downstream tasks. Clinical NLP — in which we often aim to make predictions on the basis of notes in electronic health records (EHRs) — is no exception (Alsentzer et al., 2019). However, fine-tuning large networks is a stochastic process. Performance can vary considerably as a function of hyperparameter choice, and many parameter sets can yield the same validation accuracy (i.e., the model is not identifiable), and more generally the problem is *underspecified* (D'Amour et al., 2020). Recent work has demonstrated that the choice of random seeds alone can have dramatic impact on model performance in NLP and beyond, even when all
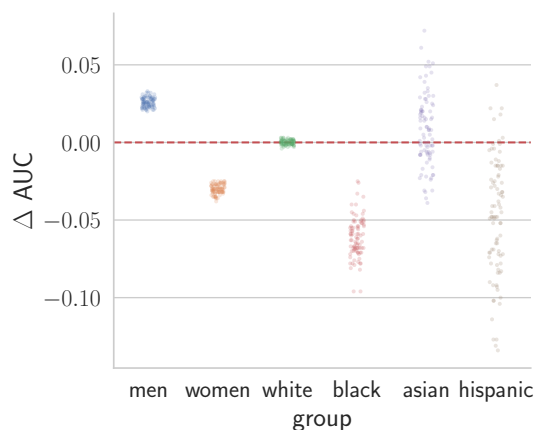

Figure 1: Differences (Δs) with respect to overall performance as a function of random seeds for demographic subgroups on the *Shock* phenotype classification task. Points represent results from pairs of seeds with similar (≤ 0.01 difference in AUC) validation performance to the best seeds.

other hyper-parameters are kept fixed (Phang et al., 2018; Dodge et al., 2020; D'Amour et al., 2020).

In this work, we explore the intersection of randomness and fairness in the context of clinical NLP. Fairness is a particularly acute concern in clinical predictive tasks, given the potential of such models to influence treatment decisions. This has motivated work investigating biases in predictive models trained over EHR (Zhang et al., 2020; Chen et al., 2018, 2019, 2020a; Pfohl et al., 2020; Chen et al., 2020b; Tripathi et al., 2020).

We investigate the impact of random seeds on the fairness of fine-tuned classifiers with respect to demographic characteristics such as gender and ethnicity. There are many definitions of algorithmic fairness which formalize different desired properties (Mehrabi et al., 2019). Following prior work, here we adopt a simple measure: The *mean differences in model performance across demographic subgroups* (Chen et al., 2019). We find that, on the popular MIMIC-III dataset (Johnson et al., 2016), seeds with comparable validation performance can

give rise to large variations in disparities across demographic subgroups (Figure 1).

2 Data and Methods

We investigate the variability of overall model performance and fairness across random seeds for a set of clinical prediction tasks derived from the Multiparameter Intelligence Monitoring in Intensive Care (MIMIC-III) set of Electronic Health Records (EHRs; Johnson et al. 2016). For each task, we train a classifier on top of the contextualized representations of a BERT (Devlin et al., 2019) model pretrained over EHR data (Alsentzer et al., 2019).

Following recent work exploring randomness and fine-tuning, we consider the seeds used to shuffle the training data and to initialize the model parameters independently (Dodge et al., 2020). Specifically, we generate $K = 1000$ pairs of shuffling and initialization seeds by sampling from a uniform distribution $\mathcal{U}(0, 10000)$. For each seed pair, we measure the overall performance as well as the performance for each demographic subgroup in terms of the Area Under the ROC Curve (AUC).

2.1 MIMIC-III

MIMIC-III is a database of deidentified EHR comprising over 40k patients admitted to the intensive care unit of the Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016). It comprises structured variables including vital sign measurements, lab test results, and medications. It also contains clinical notes (e.g., doctor and nursing notes, radiology reports, and discharge summaries), which are the focus of our analysis.

MIMIC-III contains demographic information, including potentially sensitive attributes such as ethnicity/race, sex, spoken language, religion, and insurance status (which may be seen as a proxy for socioeconomic status (Chen et al., 2019)). We are interested in the interaction between randomness and fairness in clinical predictions. Following recent prior work (Zhang et al., 2020) on the latter, we focus our analyses on two benchmark tasks proposed by Harutyunyan et al. (2019):

In-hospital Mortality (IHM)   Predict risk of in-hospital mortality based on the first 48 hours of an ICU stay.

Phenotype Classification (PC)   Classify which of 25 acute or chronic conditions (e.g., acute cerebrovascular disease, chronic kidney disease) are

present in a given patient ICU stay record. Similar to Zhang et al. (2020), we treat each condition as an independent binary classification task. Table A.1 in the Appendix enumerates the full set of conditions and their respective prevalences.

We extracted training and test datasets for these tasks using the same pre-processing pipeline as Zhang et al. (2020).[1] We kept the same data splits and reserved 20% of the training data as validation set per task. For each patient, we collected their clinical notes, as well as their gender and race/ethnicity (as recorded in the EHR). The notes were filtered according to the categories *Nurse*, *Physician* and *Nursing/Other* to avoid notes of poor semantic quality, as suggested by Zhang et al. (2020). Patients without relevant clinical notes were discarded, resulting in 11384/2591 and 22033/4919 training/test examples for the IHM and PC tasks, respectively. It should be noted that these datasets are highly imbalanced both in terms of labels and demographic distribution with 55% *Male*, 85% *White*, 9% *Black*, 3% *Asian* and 3% *Hispanic* patients. Table 1 shows the distribution of sample sizes across subgroups for each benchmark.

		Gender		Ethnicity			
		M	F	W	B	H	A
IHM	Train	6,262	5,122	8,044	1,081	353	251
	Test	1,438	1,153	1,860	226	84	49
	Val	1,580	1,268	2,027	2,64	85	62
PC	Train	12,372	9,661	15,652	2,049	728	485
	Test	2,752	2,167	3,552	451	159	95
	Val	3,029	2,480	4,002	521	196	120

Table 1: Sample sizes across subgroups for the in-hospital mortality and phenotype classification tasks.

2.2 Fine-tuned Classifiers

We define text classifiers for clinical tasks that map clinical notes corresponding to individual patients to binary labels. We extract contextualized embeddings from notes using a pretrained Transformer encoder and then map these to outputs (predictions) via a linear layer. Transformers are feedforward networks and require fixed-length inputs. To handle longer sequences, we adopt an approach used in prior works (Huang et al., 2019; Zhang et al., 2020). Given an input sequence, we: (1) Extract N subsequences with sizes equal to that expected by the Transformer input layer; (2) Make individual

---

[1] https://github.com/MLforHealth/HurtfulWords

predictions on the basis of each subsequence, and; (3) Then aggregate them into a final prediction.

More formally, an encoder ϕ operates over inputs of size E with H-dimensional hidden layers. Given a patient's clinical notes $\mathcal{X}$, we extract a set of N subsequences of length E,

$$x = \{\{w_1^1, \ldots, w_E^1\}, \ldots, \{w_1^N, \ldots, w_E^N\}\} \subseteq \mathcal{X}$$

We construct a matrix $\mathbf{Z} \in \mathbb{R}^{H \times N}$ such that the nth column represents subsequence x_n, $\mathbf{Z}_{[:,\mathbf{n}]} = \phi(x_n) = \sum_j \mathbf{z}_j^{x_n}$ where $\mathbf{z}_j^{x_n} \in \mathbb{R}^H$ is the embedding produced by the last hidden layer of the encoder for token j in the context of x_n. We then use a linear layer followed by a sigmoid activation to produce a prediction vector $\tilde{\mathbf{Y}}$, encoding the class conditional probabilities for each subsequence. This vector is then used to calculate the final probability as

$$P(Y = 1|\tilde{\mathbf{Y}}) = \frac{\tilde{\mathbf{Y}}_{\max} + \tilde{\mathbf{Y}}_{\mathrm{mean}} N/c}{1 + N/c}, \quad (1)$$

where c is a scaling factor, which we set to $c = 2$, following Huang et al. (2019).

We implement classifiers with `PyTorch` using the Transformer encoders from the `huggingface`[2] library (Wolf et al., 2019). We initialize models to weights from ClinicalBERT (Huang et al., 2019), which was trained over scientific literature and clinical notes. We train classifiers on the most recent $N = 10$ subsequences of $E = 512$ tokens from the notes associated with each patient. We train using the ADAM optimizer (Kingma and Ba, 2014) for 500 epochs with early stopping. We set the learning rate to $\alpha = 0.01$, which we found to have the best validation performance on average across all tasks.

3 Results

We compare the overall performance with the performance for each subgroup as a function of random seeds. Figure 2 shows the overall performance (left) along with the gap between the best and worst observed subgroup AUCs (right), across tasks. We observe a large variance in both the overall performance and the gap. The former observation corroborates previous findings (Dodge et al., 2020).

To quantify how random seeds affect individual subgroups, we measure the the absolute differences (Δs) between overall and subgroup performances.

We then evaluate whether there are correlations between overall performance and subgroup Δs. Figures 3 and 4 present the results for the *Shock* phenotype classification task — one of the tasks with largest disparities observed in prior work (Zhang et al., 2020). Similar trends were found for the remaining tasks, and we report all results in the Appendix (Figures A.2-A.3 and A.4-A.6).

Figure 3 shows that the performance of all subgroups varies significantly across random seeds and that variances are higher for minority groups. Larger variations in minority subgroups are to be expected, as any empirical estimate will have a variance that is inversely proportional to the sample size of a group. In Figure 4, we observe that there seem to be two distinct clusters of seeds: One corresponding to high performing models (right of plots), and another to suboptimal models.[3] While the best performing models tend to have a lower variance of subgroup performance, there is otherwise no clear relationship between overall and subgroup performance. Indeed, we find that many models with similar overall performance correspond to widely different subgroup Δs, particularly for the minority groups.

To explore the implications of this phenomenon, we simulate a grid search over all the random seeds on the validation set. We select the best seed along with all other seeds with similar performance (i.e., within a difference of $\epsilon = 0.01$ absolute AUC). Figure 1 shows the test set subgroup performance Δs, for the best validation seeds, in the *Shock* phenotype classification task (see Figures A.7-A.8 for the other tasks). Figure 5 summarizes the overall performance (left) along with the subgroup performance gap (right) across tasks.

We can see that selecting seeds on the basis of overall performance helps to reduce the subgroup performance gap (compare the right subplots in Figures 2 and 5). However, the top performing models show disparities with respect to both gender and ethnicity, suggesting that these models maximize performance for some groups at the expense of others. Moreover, we find *multiple seeds with similar levels of validation performance that correspond to very different subgroup Δs*.

Since we have not encoded any model selection preferences into the pipeline this variance may reflect a form of *underspecification*. Can we define

---

[3]Dodge et al. (2020) also found that some seeds performed consistently well across all the evaluated tasks, while others always performed poorly.
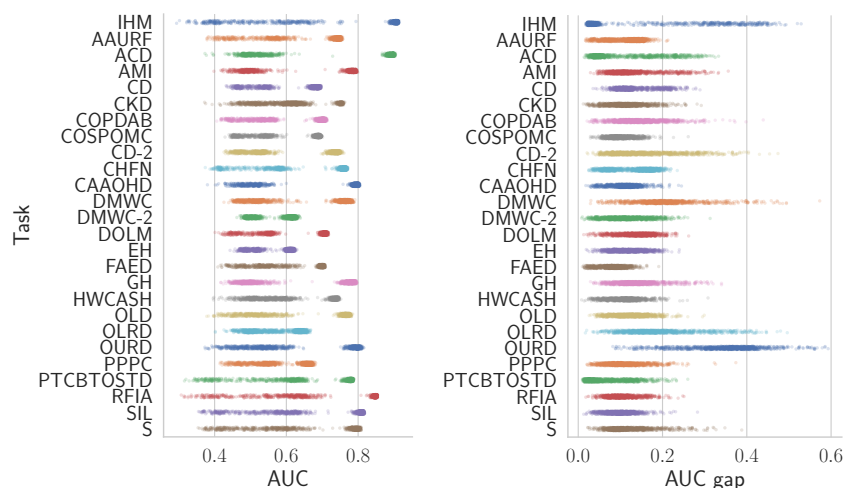
Figure 2: Variation of model performance across random seeds for all tasks (task details in Appendix). *Left:* Overall performance. *Right:* Gap between best and worst subgroup.
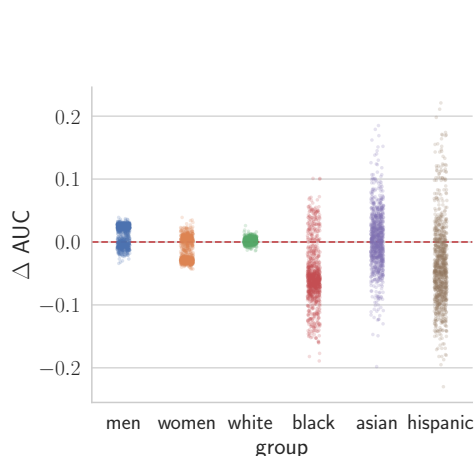

Figure 3: Differences relative to overall AUC as a function of random seeds for subgroups on the *Shock* phenotype classification task.
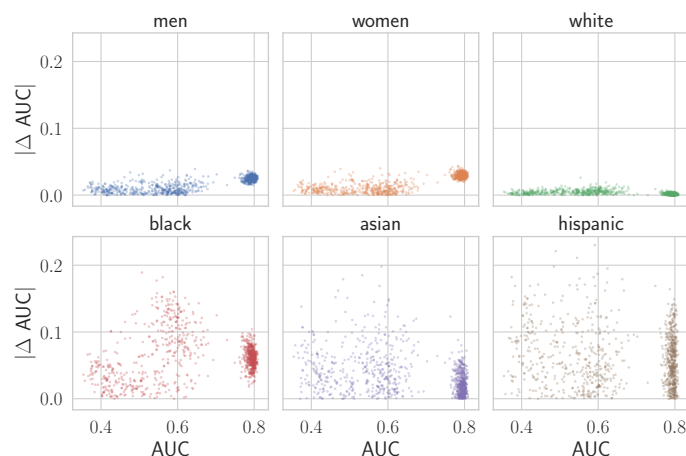

Figure 4: Correlations between overall performance and subgroup performance on the *Shock* phenotype classification task.

criteria that explicitly accounts for subgroup performance? We could then ask whether it is possible to maximize both fairness and overall performance with respect to random seeds. We repeated the grid search experiments with simple criteria that incorporate some notion of subgroup performance, such as selecting the seeds that: (a) maximize subgroup macro-average performance; (b) minimize the average subgroup Δ; and (c) maximize the overall performance minus the average subgroup Δ. To account for the effect the sample sizes on the apparent subgroup performance, we directly compare subgroup Δs for each random seed on the validation set and the test set. We find that correlations between validation set and test set fairness are either non-existent or very weak in most tasks.

These findings imply that the same pipeline may produce models with similar validation performance but very different levels of apparent 'fairness' as a result of varying the random seed alone. However, the fact that training-set and validation-set fairness are not reliable indicators of test-set fairness suggests that variance due to small subset sizes may be significant.

This is in some sense not surprising, given the combination of pronounced class imbalance and small subgroup samples in this data (see Tables 1 and A.1). To confirm this, we repeat the experiments on a subset of the test data containing the same number of examples (equal to the smallest subgroup) for *all* groups, including majority groups. Evaluating all subgroups using small samples yields similarly high variances in performance Δ (Figure 6 and Appendix Figure A.1), which con-
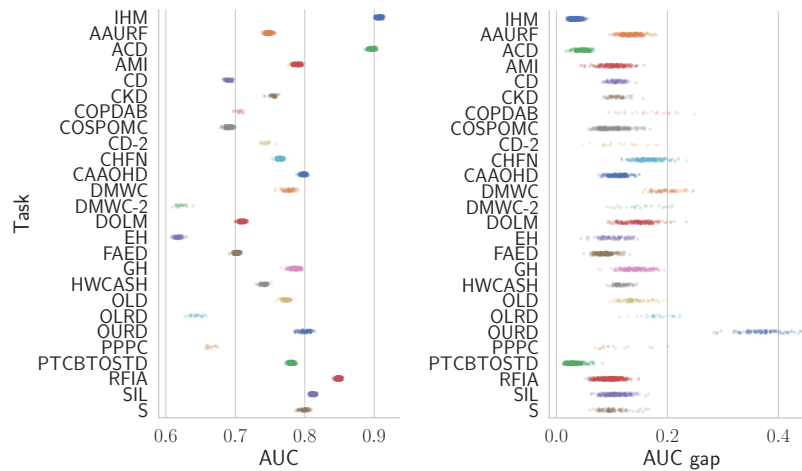
Figure 5: Variation of model performance for seeds with validation performance similar to the best seeds, for all tasks. *Left:* Overall performance. *Right:* Gap between best and worst subgroup.
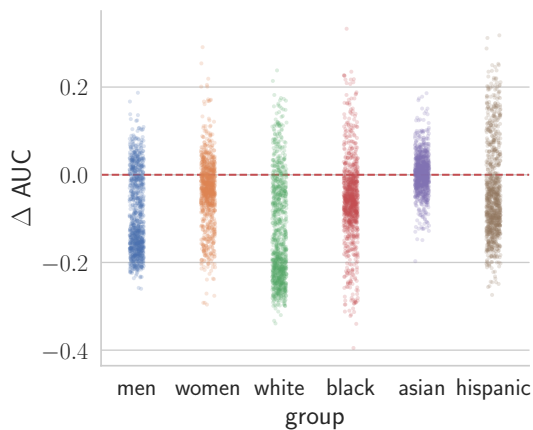

Figure 6: Differences (Δs) estimated from a balanced subset of the test data with equal sample sizes for all demographic subgroups. As in Figure 3, we show deviations relative to the overall AUC for the *Shock* phenotype classification task.

firms that the sample size is a significant factor in the variation of apparent model performance across random seeds.

These findings suggest that work investigating the fairness of fine-tuned classifiers should be careful to account for: a) model variability due the choice of random seeds; and b) variance in performance estimates due to small sample sizes. See Appendix Section B for an illustrative example. These observations are relevant for research using MIMIC-III, and for any corpora with similar properties, namely the combination of class imbalance and comparatively small subgroup sizes, which is likely to be present in EHR data where conditions are relatively rare and one is interested in fairness to minority groups (which are smaller by definition).

4 Conclusions

We have investigated the impact of random seeds on the fairness of fine-tuned pre-trained models for clinical tasks. Specifically, we measured gaps in performance across gender and racial subgroups as a function of the choice of random seeds for data shuffling and parameter initialization. In line with prior work, we found that classifiers trained on MIMIC-III data are often biased with respect to demographic subgroups. The contribution of this work is the empirical confirmation that choice of random seed alone significantly affects the apparent bias: Seeds that yield comparable performance in aggregate on the validation data correspond to very different performances on subgroups in test data. Our analyses corroborate Dodge et al. (2020)'s findings on the importance of carefully chosen random seeds, but also suggest that an equal amount of attention should be payed to the impact of these choices on model fairness.

However, interpretation of these results is complicated by sample size effects. While MIMIC-III is in itself a large dataset, it also exhibits significant imbalance, both in terms of subgroups of patients and the prevalence of medical conditions. These imbalances compound when considering subsets of patients in the context of specific prediction tasks, which often leads to small sample sizes for minority subgroups. While we observed higher apparent variances for demographic minorities, our results also suggest that these variances can in large part be explained by the smaller sample sizes. Indeed, we found the variances in subgroup performance to be inversely proportional to the size of the subgroup.

Ethical Considerations

Fairness has rightly been an issue of increasing concern within the NLP community. This issue is particularly important in clinical NLP, given the potential that such models may ultimately have on patient health. We have investigated the degree to which different subgroup performances may be observed even fixing the (aggregate) validation data performance; we find wide variances across subgroups. That said, this work also highlights inherent limitations of using MIMIC-III (the standard dataset for clinical NLP) to evaluate the fairness of models, given the relatively small samples of patients that belong to demographic groups of interest. We hope these contributions encourage continued research into fairness in the context of clinical NLP.

Acknowledgements

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pages 3539–3550.

Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2020a. Ethical machine learning in health. *arXiv preprint arXiv:2009.10576*.

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

John Chen, Ian Berlot-Attwell, Xindi Wang, Safwan Hossain, and Frank Rudzicz. 2020b. Analyzing text specific vs blackbox fairness algorithms in multimodal clinical nlp. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 301–312.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multi-task learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

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

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Stephen R Pfohl, Agata Foryciarz, and Nigam H Shah. 2020. An empirical characterization of fair machine learning for clinical risk prediction. *arXiv preprint arXiv:2007.10306*.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Sandhya Tripathi, Bradley A Fritz, Mohamed Abdelhack, Michael S Avidan, Yixin Chen, and Christopher R King. 2020. (un) fairness in post-operative complication prediction models. *arXiv preprint arXiv:2011.02036*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120.

A Clinical Prediction Tasks and Results

Table A.1 shows all the clinical prediction tasks and the respective prevalence. The following plots present the results for the Phenotyping classification tasks. Figures A.2 and A.3 show the performance Δs for each subgroup and the overall performance, as a function of the random seeds. Figures A.4 to A.6 show the overall performance against the Δ for each subgroup. Figures A.7 and A.8 show the subgroup performance Δs for pairs of seeds with validation performance similar to that of the best seeds.

Task	Description	Prevalence
IHM	In-Hospital Mortality	0.13
AAURF	Acute and unspecified renal failure	0.21
ACD	Acute cerebrovascular disease	0.07
AMI	Acute myocardial infarction	0.11
CD	Cardiac dysrhythmias	0.32
CKD	Chronic kidney disease	0.13
COPDAB	Chronic obstructive pulmonary disease and bronchiectasis	0.13
COSPOMC	Complications of surgical procedures or medical care	0.2
CD-2	Conduction disorders	0.07
CHFN	Congestive heart failure; nonhypertensive	0.28
CAAOHD	Coronary atherosclerosis and other heart disease	0.33
DMWC	Diabetes mellitus with complications	0.1
DMWC-2	Diabetes mellitus without complication	0.19
DOLM	Disorders of lipid metabolism	0.27
EH	Essential hypertension	0.41
FAED	Fluid and electrolyte disorders	0.25
GH	Gastrointestinal hemorrhage	0.07
HWCASH	Hypertension with complications and secondary hypertension	0.13
OLD	Other liver diseases	0.08
OLRD	Other lower respiratory disease	0.04
OURD	Other upper respiratory disease	0.04
PPPC	Pleurisy; pneumothorax; pulmonary collapse	0.08
PTCBTOSTD	Pneumonia (except that caused by tuberculosis or sexually transmitted disease)	0.14
RFIA	Respiratory failure; insufficiency; arrest (adult)	0.18
SIL	Septicemia (except in labor)	0.14
S	Shock	0.07

Table A.1: Clinical prediction tasks along with the respective prevalence

B Fine-tuning Experiments

To illustrate the impact that random seeds can have on measures of algorithmic fairness, we replicate the experiments concerning the fairness of fine-tuned clinical classifiers reported in (Zhang et al.,
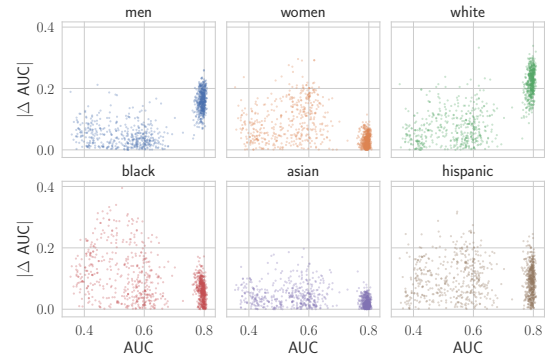

Figure A.1: Correlations between overall performance and subgroup performance on the *Shock* phenotype classification task, evaluated on a subset of the test data with equal sample sizes for all demographic subgroups.

2020)[4]. The study compares 3 measures of multigroup fairness (recall gap, parity gap and specificity gap) with respect to protected attributes such as the gender, language, ethnicity and insurance status. The experiments were conducted over 57 downstream classification tasks: including the IHM and PC tasks we considered in this paper, 3 additional tasks derived from logical ORs on subsets of the PC tasks, and a variation of PC using only the first note. Table 4 in (Zhang et al., 2020) reports the number of tasks with statistically significant gaps and the percentage of significant tasks which favor a subgroup.

We use the same experimental setup and implementation[5] to replicate the experiments for IHM and PC using only the first note (29/57 tasks). We repeat each experiment with 100 different random seeds (using the same seed to shuffle the data and initialize parameters) and compute the mean and standard deviation of each measurement across seeds (Table B.1). We find that in general the number of tasks with significant differences is roughly half of those reported by (Zhang et al., 2020), which was expected since we considered half of the tasks. However, we also observe that changes in a single random seed can affect the disparities across protected groups, both in terms of the number and the magnitude of the gaps. We see that there can be variations of up to two tasks with significant gaps and differences of up to 31% in the percentage of tasks favoring a specific group.

---

[4]see Section 5.2 and Tables 1 and 4.
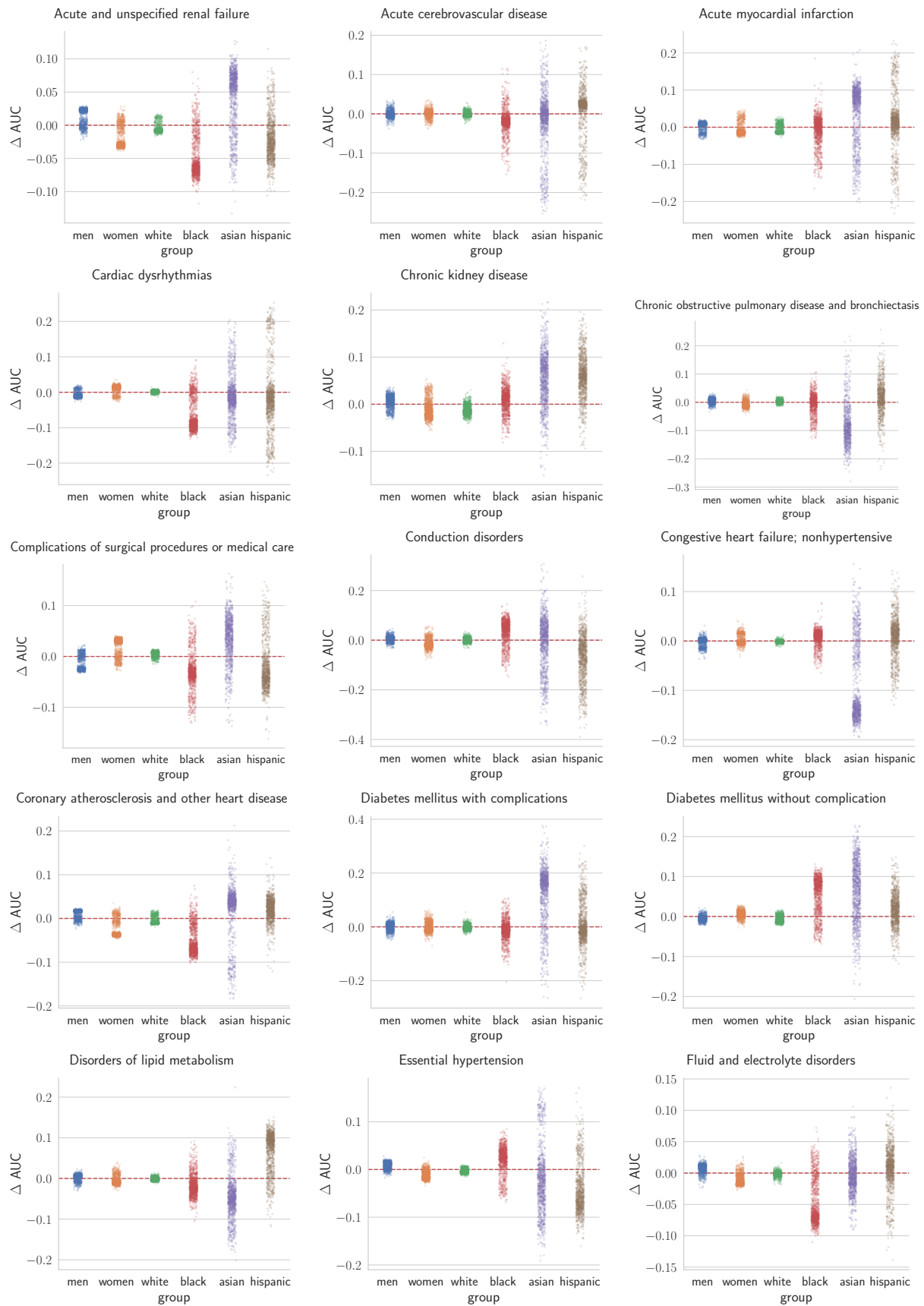[5]https://github.com/MLforHealth/HurtfulWords

Figure A.2: Differences relative to overall performance as a function of random seeds for each subgroup
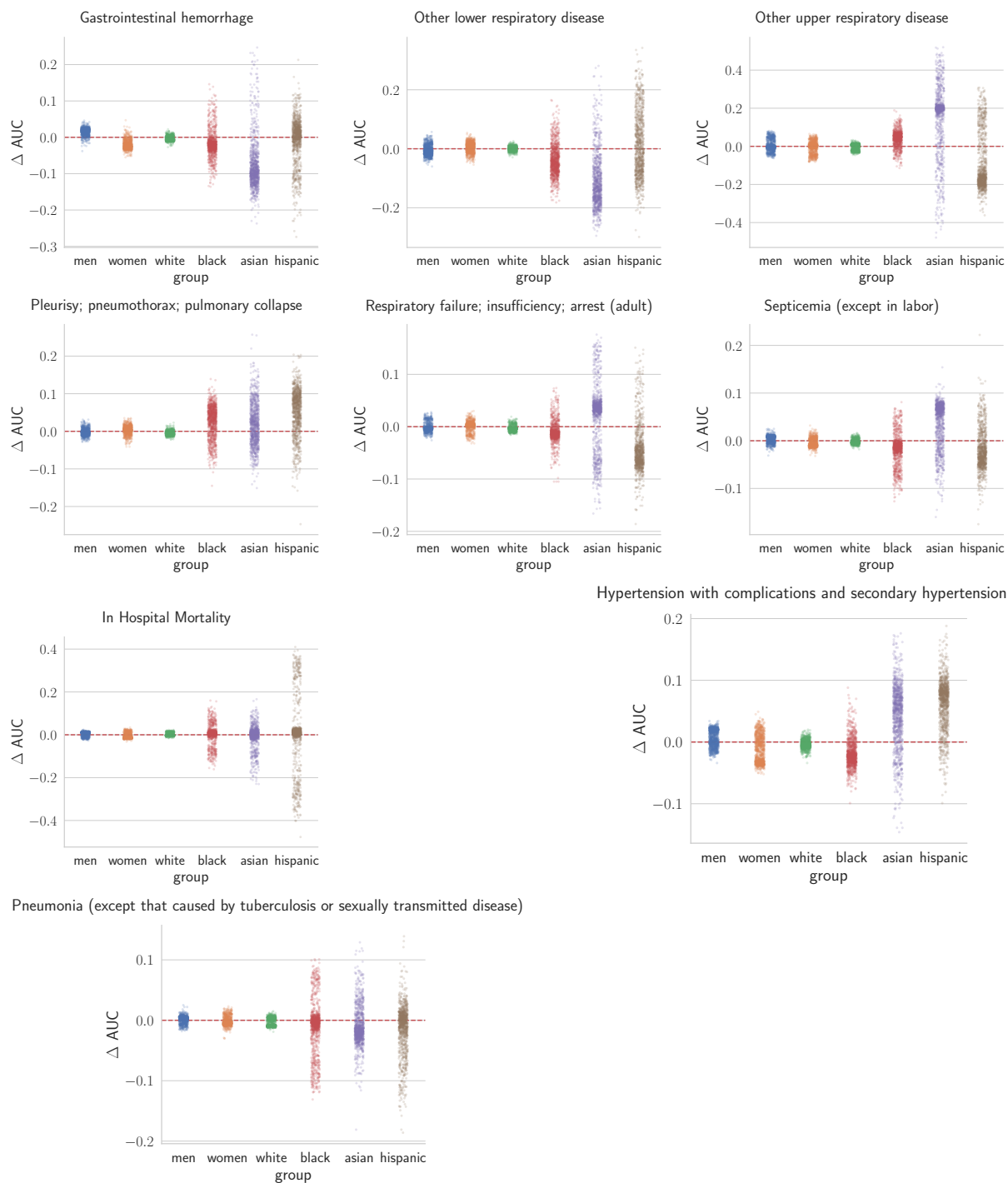
Figure A.3: Differences relative to overall performance as a function of random seeds for each subgroup
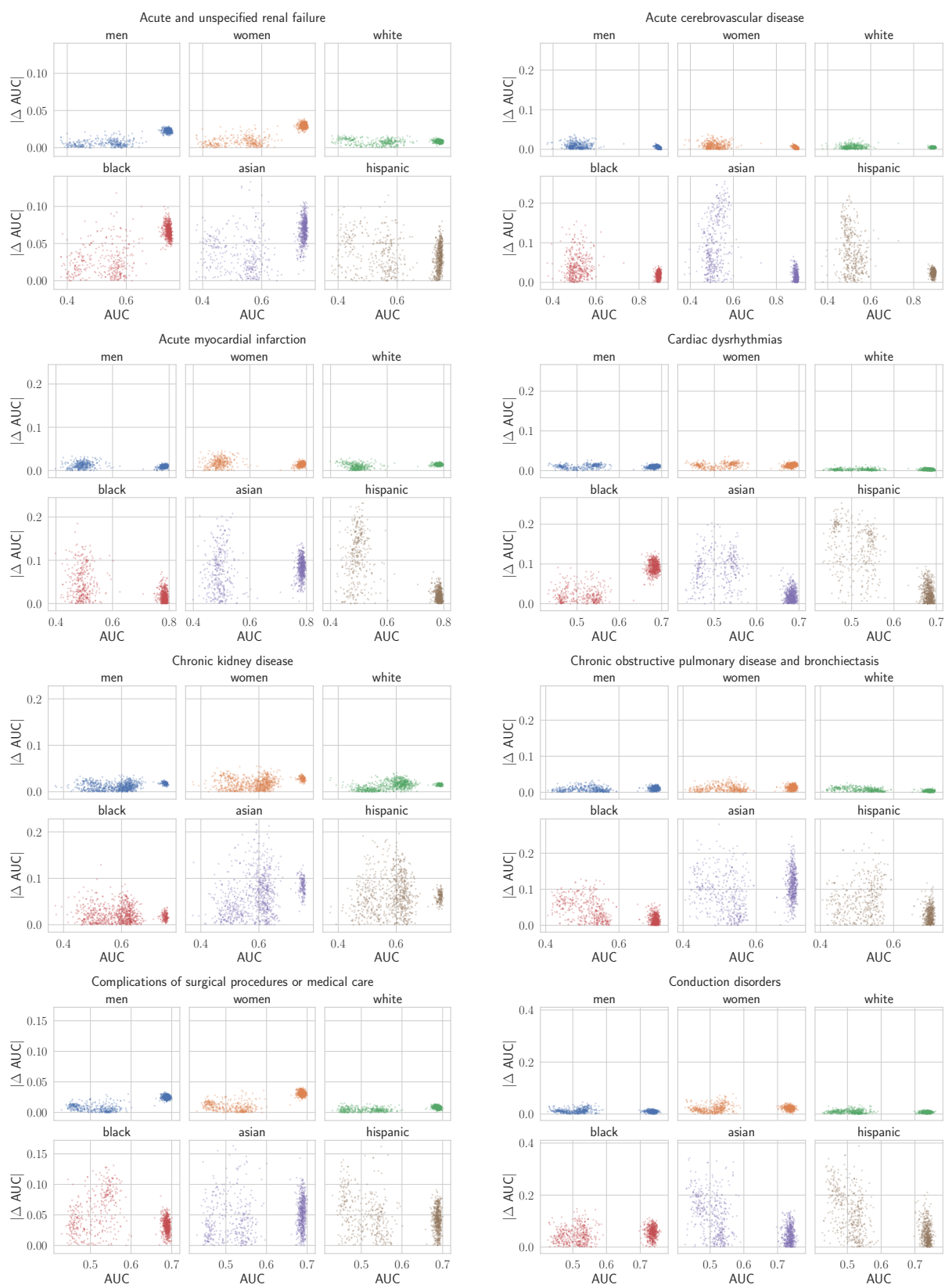
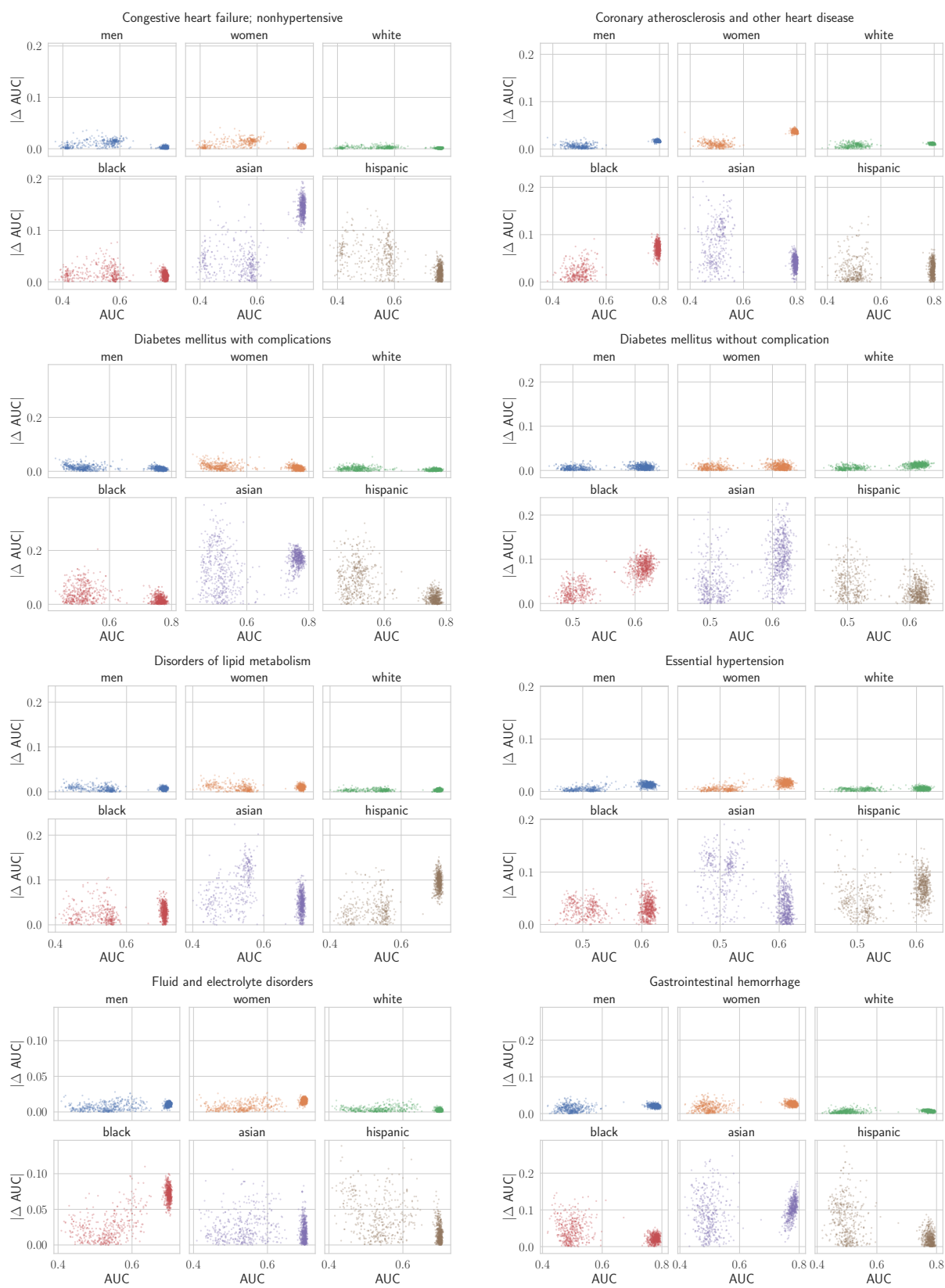Figure A.4: Correlations between overall performance and subgroup performance Δ

Figure A.5: Correlations between overall performance and subgroup performance Δ
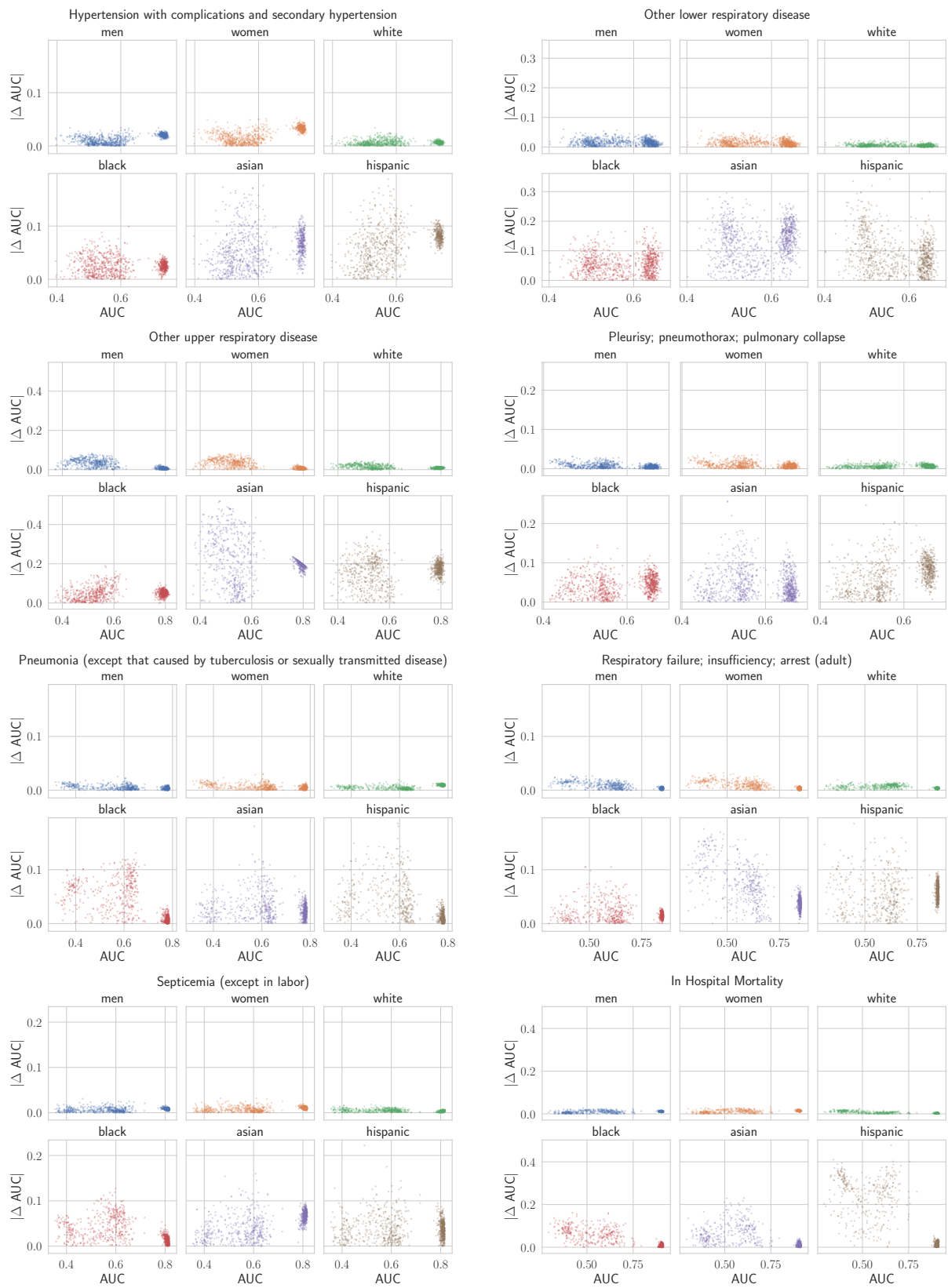
Figure A.6: Correlations between overall performance and subgroup performance Δ
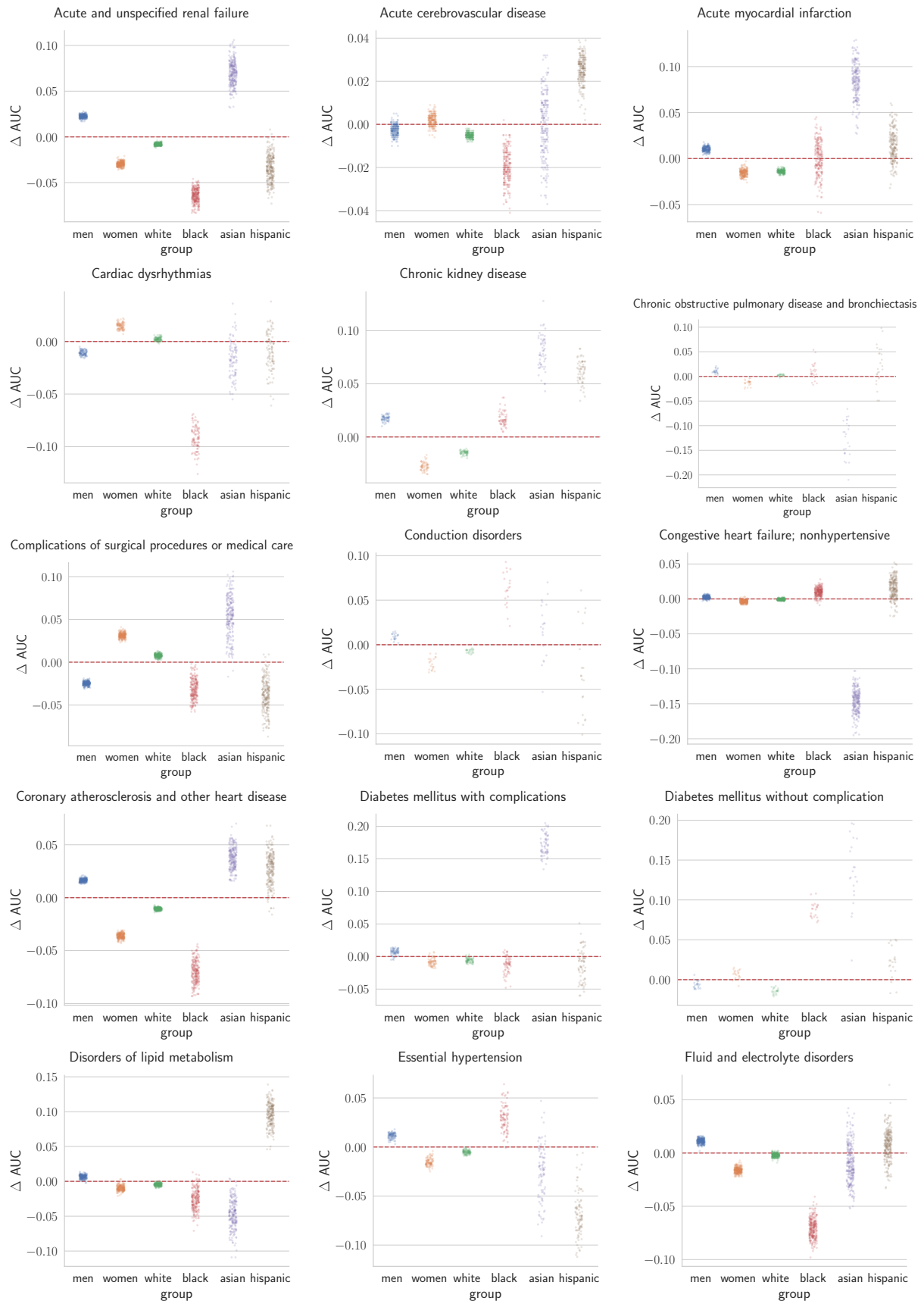
Figure A.7: Differences relative to overall performance as a function of random seeds for each subgroup. Each point represents a run for a pair of seeds with validation performance similar to that of the best seeds.
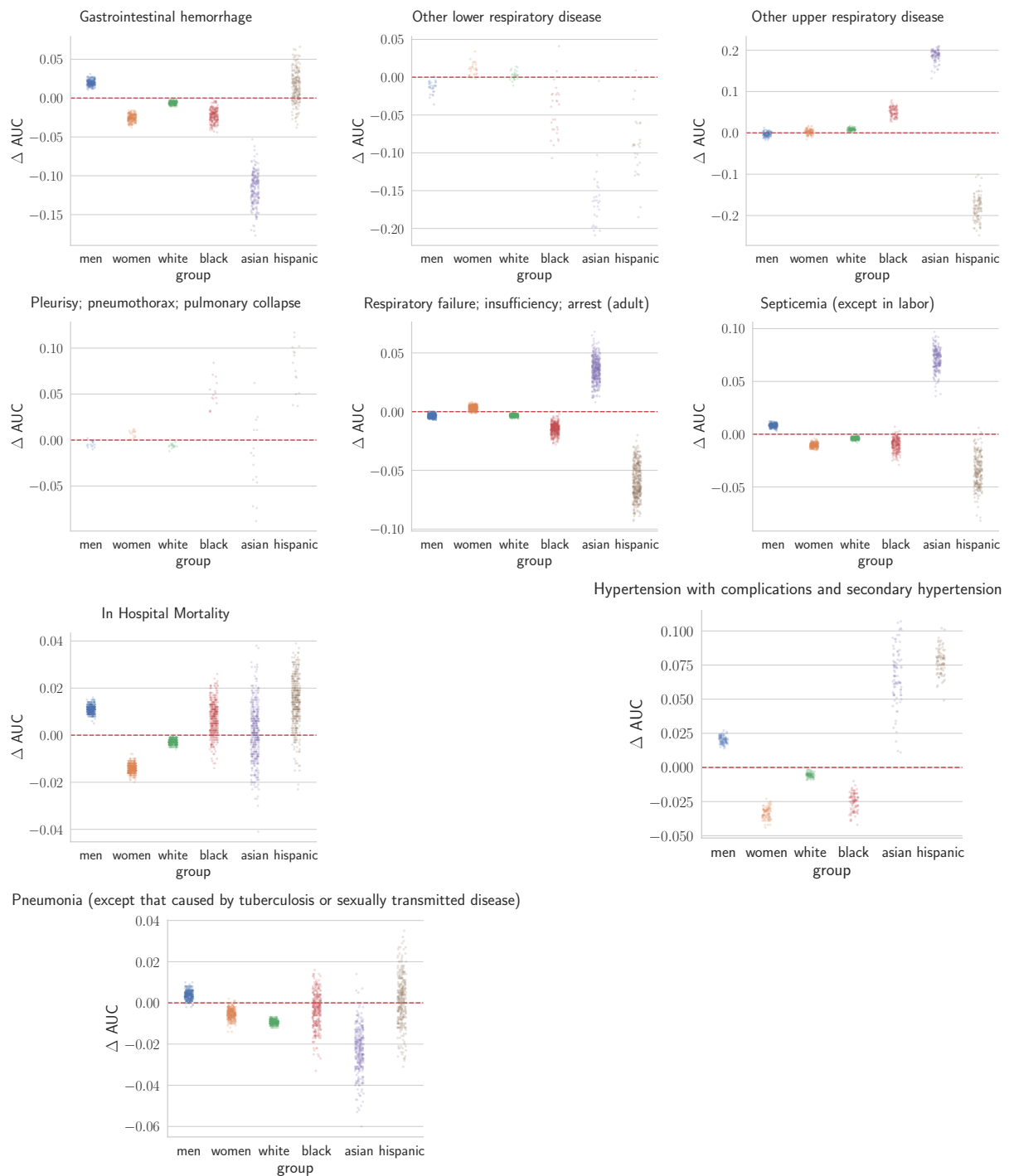
Figure A.8: Differences relative to overall performance as a function of random seeds for each subgroup. Each point represents a run for a pair of seeds with validation performance similar to that of the best seeds.

		Recall Gap	Parity Gap	Specificity Gap
Gender	Male vs. Female (% Tasks Favoring Male)	3.0 ± 1.2 ($67.8 \pm 25.2\%$)	11.2 ± 2.0 ($39.0 \pm 8.1\%$)	9.5 ± 2.0 ($76.8 \pm 10.3\%$)
Language	English vs. Other (% Tasks Favoring Other)	3.1 ± 1.4 ($50.3 \pm 28.9\%$)	6.9 ± 1.9 ($5.4 \pm 7.0\%$)	4.2 ± 1.5 ($88.9 \pm 14.8\%$)
Ethnicity	White vs. Other (% Tasks Favoring White)	2.5 ± 1.5 ($93.0 \pm 22.0\%$)	7.3 ± 1.7 ($92.6 \pm 10.7\%$)	4.9 ± 1.5 ($11.1 \pm 12.1\%$)
	Black vs. Other (% Tasks Favoring Black)	3.8 ± 1.5 ($37.9 \pm 21.2\%$)	6.6 ± 1.7 ($65.6 \pm 15.0\%$)	3.6 ± 1.4 ($40.8 \pm 22.4\%$)
	Hispanic vs. Other (% Tasks Favoring Hispanic)	5.1 ± 1.6 ($8.4 \pm 10.9\%$)	7.6 ± 1.8 ($0.0 \pm 0.0\%$)	9.3 ± 1.8 ($99.1 \pm 8.9\%$)
	Asian vs. Other (% Tasks Favoring Asian)	6.1 ± 1.5 ($53.6 \pm 15.8\%$)	2.3 ± 1.3 ($77.1 \pm 31.0\%$)	3.6 ± 1.5 ($54.3 \pm 25.0\%$)
	Other vs. Other (% Tasks Favoring Other)	10.0 ± 1.7 ($6.0 \pm 4.9\%$)	2.6 ± 1.1 ($1.0 \pm 5.0\%$)	4.1 ± 1.2 ($94.5 \pm 12.5\%$)
Insurance	Medicare vs. Other (% Tasks Favoring Medicare)	15.0 ± 2.0 ($93.8 \pm 9.5\%$)	25.7 ± 2.5 ($92.2 \pm 8.6\%$)	23.9 ± 2.6 ($2.9 \pm 2.6\%$)
	Private vs. Other (% Tasks Favoring Private)	7.1 ± 1.4 ($10.5 \pm 9.0\%$)	19.5 ± 2.2 ($4.2 \pm 3.3\%$)	19.7 ± 2.4 ($95.5 \pm 9.0\%$)
	Medicaid vs. Other (% Tasks Favoring Medicaid)	9.0 ± 1.7 ($8.7 \pm 7.8\%$)	17.2 ± 2.1 ($12.0 \pm 3.3\%$)	15.0 ± 2.1 ($92.8 \pm 9.3\%$)

Table B.1: We replicated Zhang et al. (2020) analysis of multi-group fairness performance gaps for fine-tuned classifiers across gender, language, ethnicity, and insurance status. We evaluated 28 (out of 57) tasks and repeated the experiments with 100 different random seeds. We measured the average and standard deviation of the number of tasks with statistically significant differences, and the percentage of significant tasks which favor a subgroup.