Quantifying Fairness in a Multi-Group Setting and its Impact in the Clinical Setting

Anonymous Author(s)
Affiliation
Address
email

Abstract

In the machine learning fairness literature, the majority of fairness definitions are formalized for the binary case where there are only two protected groups. Such formalization allows for simple demonstrations of fairness, and enables straightforward proofs. However, in the clinical setting, many prediction tasks involve multiple protected groups. Moreover, current frameworks often do not explore the effects of comparing multiple groups, leading to unexpected consequences. As such, to audit models before deploying machine learning algorithms in healthcare, definitions of fairness must be defined and expanded robustly beyond the binary paradigm. We analyze different methods of expanding fairness definitions from the binary cases to the multi-group case of multiple protected groups. We also highlight edge cases where such expansions are actually unfair when assessed using another metric. We perform empirical analyses on clinical classification tasks to assess the likelihood of such edge cases happening in healthcare classification problems.

1 Introduction

There is a plethora of fairness definitions in the current literature in fair machine learning, including demographic parity [14], equalized odds [7], and individual fairness [4]. However, the vast majority of definitions only formalize definitions for the *binary case*. In this work, the term *binary case* is used to refer to binary classification settings where there are only two demographic groups. Examples include White vs Black recidivism prediction [12], loan repayment prediction [7], and college admission [5]. The term *formalized* is used to describe mathematical expansion of stated definitions which is usually followed by examples. Thus far, researchers in the machine learning and health community have not agreed on a consistent and explicit expansion for the non-binary (i.e. *multi-group*) setting. For example, [6] considers absolute average disparity, [9] considers the maximum disparity, [2] uses pairwise comparisons between multiple groups, and [13] considers an mean squared disparity.

As with many applications, in healthcare, researchers often deal with multiple protected groups. In this work, we use *multi-group* to refer to fairness evaluations for which the demographic can be stratified into multiple groups. A clear way to expand binary formalizations of fairness to the multi-group setting is therefore necessary to assess and correct problematic biases. For example, while gender fairness is often evaluated using binary fairness definitions, ethnicity should be evaluated using multi-group fairness definitions.

While there are many existing expansions of binary definitions [9, 13], it is not clear for what situations different definitions would fit best. More importantly, under further analysis, it becomes evident that many of the commonly-used multi-group expansions can have unexpected edge cases whereby choosing a classifier which minimizes disparity per one expansion may inadvertently maximize disparity per another expansion.

In this work, we:

- Analyze different ways of expanding existing fairness definitions from its canonical binary form with two protected groups to the multi-group setting with many protected groups.
- Show that, for all studied expansions, there exists the edge case where choosing a classifier will actually *maximize unfairness* as measured by another fairness metric, when attempting to minimize unfairness by the expansion in question.
- Perform empirical experiments on MIMIC-III [11] for each definition and their associated edge cases to analyze likelihoods associated with each edge case.

Given the importance of analyzing fairness in deployed settings, and relevance of multi-group fairness dilemmas in many real-life applications (e.g. healthcare), we hope this work can serve as a guide and catalyst for relevant discussion.

2 Motivating Example

Accurate prediction of one-year mortality after a hospital visit can benefit patients by improving tailored end-of-life care, and reducing palliative care healthcare worker burnout by informing management decisions [1]. Consider an administrator in a palliative care department who must choose between two similarly performing algorithms (e.g., two classifier that share the same model, but different hyperparameters). If the administrator knows the *average disparity*, and given no statistically significant difference in the algorithms' performance, it is logical to choose the classifier with the lower average performance disparity. However, this could – and often does – result in choosing the classifier with the largest *maximum disparity* between any two protected groups (see Section 3.2).

In this work, we will consider group fairness for a classifier which predicts the label Y given input feature x from a dataset X. Each input $x \in X$ has associated with it a categorical identifier $z \in \mathcal{Z}$, which we will call the "protected group". When training a fair machine learning model, the goal is to learn some classifier $\tilde{Y} = f(x)$ such that f is *unbiased* with respect to $\mathcal{Z}$.

3 Multi-group Fairness Expansion

In this section, we explore common ways of expanding fairness definitions to the multi-group case.

In the binary case, the equality of odds for the positive group (i.e. equality of opportunity) can be computed by:

$$\forall \tilde{y} \in \tilde{Y}, P(\tilde{Y} = \tilde{y}|Y = 1) = \forall z \in \mathcal{Z}, P(\tilde{Y} = \tilde{y}|Y = 1, \mathcal{Z} = z)) \tag{1}$$

It follows that the equality gap for the positive group can be defined as:

$$TPR(z_1) - TPR(z_2)$$

It is not obvious how to best expand Equation 1 to adjust for multiple classification targets or multiple protected groups. We furthermore illuminate how such expansions can have unintended results.

Binary prediction tasks. For brevity, our analyses will consider the case of a binary *prediction task*, but multiple *protected groups*, a scenario which is common in the healthcare setting [2]. However, the analyses also hold for non-binary prediction problems.

When expanding beyond the binary case, it may be of interest to clinical researchers to measure both a performance gap for each protected group individually, as well as a single metric for the entire classifier. Below, we analyze the edge cases of methods for can be used to represent either a multi-group group-gap or a classifier's performance.

A case for evaluating positive equality gap in healthcare. In healthcare settings, we believe that the positive equality gap[1] is relevant, as cautionary examination is better than under-treating and risking death. Thus, for the each definition below, we will use lowercase define g_i as the gap for a particular ethnicity i, and uppercase G refers to the gap of the classifier when comparing all pairwise ethnicities. $\mathcal{Z}$ is the set of all ethnicities. TPR_i is the true positive rate of ethnicity i. TP_i and FN_i refers to the number of true positive and false negative examples for ethnicity i.

[1] In other words, the difference between true positive rates or recall

3.1 Expansion: Combining all other groups

3.1.1 Motivation

The following expansion can be used to measure the fairness of a classifier with respect to a single group. To evaluate if a classifier is performing significantly different for a given group, an intuitive option is to combine all other groups into a single group for comparison.

3.1.2 Formalization

We define the notation $\neg j : i \in j, i \neq j$ and the designation $TPR_{\neg j}$ to indicate the true positive rate across all groups in $\neg j$. We define g_j^{comb} as follows:

$$g_j^{comb} = TPR_j - TPR_{\neg j}$$

3.1.3 Edge Cases

Although simple and intuitive, this definition suffers from at least two degenerate edge cases. Firstly, such an expansion is not robust against group imbalances. Taking the example of racial protected groups, the differences between Hispanic and another population may be overshadowed by the lack of a difference between Hispanic and a majority group, illustrated in **Proof 1**.

Lemma 1 $g_j^{comb} \approx 0$ does not necessarily imply that the true positive rate (TPR) for group j is equal to the TPR of other groups.

Proof 1 Consider the case where:

$TP_1 = 60, FN_1 = 40$ and $TP_2 = 600, FN_2 = 400$ and $TP_3 = 5, FN_3 = 4$

$$g_1^{comb} = TPR_1 - TPR_{\neg 1}$$

$$g_1^{comb} = \frac{60}{100} - \frac{605}{1009}$$

$$g_1^{comb} \approx 0$$

We see that TPR_1 and TPR_3 do not match, with $TPR_1 = 0.6$ and $TPR_3 = 0.55$. In other words, using such an expansion may under-report differences between protected groups.

A second problem with this expansion is that differences in opposing directions can be lost when combined (see **Proof 2**).

Lemma 2 $g_j^{comb} = 0$ does not necessarily imply that the TPR for group j is equal to the TPR of other groups.

Proof 2 Consider the case where:

$TP_1 = 60, FN_1 = 40$ and $TP_2 = 70, FN_2 = 30$ and $TP_3 = 50, FN_3 = 50$

$$g_1^{comb} = TPR_1 - TPR_{\neg 1}$$

$$g_1^{comb} = \frac{60}{100} - \frac{120}{200}$$

$$g_1^{comb} = 0$$

However, *none* of the true positive rates for the individual groups match, with $TPR_1 = 0.6 \neq TPR_2 = 0.70 \neq TPR_3 = 0.5$.

3.2 Expansion: Absolute Average Disparity

3.2.1 Motivation

When comparing the unfairness of two classifiers, another intuitive expansion may be to examine the absolute average unfairness of the classifier.

3.2.2 Formalization

For an individual group:

$$g_j^{avg} = \frac{1}{|\mathcal{Z}|-1} \sum_{i \in \mathcal{Z}, i \neq j} |TPR_j - TPR_i|$$

For all groups:

$$G^{avg} = \frac{1}{(2\binom{|\mathcal{Z}|}{2}))} \sum_{i,j \in \mathcal{Z}, i \neq j} |TPR_j - TPR_i|$$

3.2.3 Edge Cases

Lemma 3 Choosing a classifier with lower average unfairness g_j^{avg} or G^{avg} may result in unintentionally maximizing the disparity g_j^{max} or G^{max} (further defined in Section 3.3.2) between two groups.

Proof 3a For the individual case, consider the performances of classifiers C_1 and C_2, where:

$C_1 : TPR_1 = 100, TPR_2 = 50, TPR_3 = 51, TPR_4 = 52$

$C_2 : TPR_1 = 100, TPR_2 = 25, TPR_3 = 98, TPR_4 = 99$

We observe that for C_1, $g_1^{avg}(C_1) = 49$ and $g_1^{max}(C_1) = 50$, while for C_2, $g_1^{avg}(C_2) = 26$ and $g_1^{max}(C_2) = 75$.

Thus, a hypothetical healthcare administrator selecting the classifier with the lowest average disparity for group 1 would use C_2; however, this actually maximizes disparity for this group and another single group. This can be particularly troubling if the other group is already uniquely historically disadvantaged.

Proof 3b This edge case also exists for the group case. Consider the performance of two classifiers:

$C_1 : TPR_1 = 100, TPR_2 = 49, TPR_3 = 98, TPR_4 = 99$

$C_2 : TPR_1 = 100, TPR_2 = 50, TPR_3 = 65, TPR_4 = 80$

We observe that for C_1, $G^{avg}(C_1) = 25.7$ and $G^{max}(C_1) = 51$, while for C_2, $G^{avg}(C_2) = 27.5$ and $G^{max}(C_2) = 50$.

We see that in some cases, selecting the classifier with the lower average disparity may actually maximize pairwise disparity.

3.3 Expansion: Maximum Disparity

3.3.1 Motivation

Given the possible grave impact and consequences of clinical and policy decisions in the healthcare setting, researchers often choose an aggressive approach when measuring unfairness, to ensure that no minority group is receiving unfair treatment. With this good intention, researchers may report the maximum pairwise fairness gap as the performance of the classifier [9, 2].

3.3.2 Formalization

Formalization for an individual group:

$$g_j^{max} = max(max(TPR_j, TPR_i) - min(TPR_j, TPR_i)), \forall i \in \mathcal{Z}$$

For all groups:

$$G^{max} = max(max(TPR_j, TPR_i) - min(TPR_j, TPR_i)), \forall i, j \in \mathcal{Z}$$

3.3.3 Edge Cases

Lemma 4 Selecting the classifier with the smaller maximum disparity may result in inadvertently selecting for the larger average disparity.

Proof 4a Consider the same two classifiers from above:

$C_1 : TPR_1 = 100, TPR_2 = 50, TPR_3 = 51, TPR_4 = 52$

$C_2 : TPR_1 = 100, TPR_2 = 25, TPR_3 = 98, TPR_4 = 99$

We observe that for C_1, $g_1^{max}(C_1) = 50$ and $g_1^{avg}(C_1) = 49$. For C_2 we see that $g_1^{max}(C_2) = 75$ and $g_1^{avg}(C_2) = 26$. Thus, a healthcare administrator taking a cautious approach to avoiding an unfair classifier using maximal pairwise difference may actually perpetrate greater average unfairness.

Proof 4b Consider the performance of the two classifiers with the following recalls (TPR) for each group:

$C_1 : TPR_1 = 100, TPR_2 = 49, TPR_3 = 98, TPR_4 = 99$

$C_2 : TPR_1 = 100, TPR_2 = 50, TPR_3 = 65, TPR_4 = 80$

For $C_1, G^{max}(C_1) = 51$ and $G^{avg}(C_1) = 25.7$. However for $C_2, G^{max}(C_2) = 50$ and $G^{avg}(C_2) = 27.5$ showing that the unexpected behavior also holds for

3.4 Expansion: Mean Squared Disparity

3.4.1 Motivation

In some cases, examining mean squared performance disparities may be a more reasonable choice than the average when one is particularly cautious about mitigating very large biases.

3.4.2 Formalization

Formalization for an individual group:

$$g_j^{mse} = \sqrt{\frac{1}{|\mathcal{Z}| - 1} \sum_{i \in \mathcal{Z}, i \neq j} (TPR_j - TPR_i)^2}$$

For all groups:

$$G^{mse} = \sqrt{\frac{1}{(2\binom{|\mathcal{Z}|}{2})} \sum_{i,j \in \mathcal{Z}, i \neq j} (TPR_j - TPR_i)^2}$$

3.4.3 Edge Cases

Lemma 5 Choosing a classifier with a smaller mean squared disparity gap may result in a larger maximum disparity gap between two groups.

Proof 5a For the individual case consider two classifiers:

$C_1 : TPR_1 = 100, TPR_2 = 50, TPR_3 = 60, TPR_4 = 70$

$C_2 : TPR_1 = 100, TPR_2 = 40, TPR_3 = 80, TPR_4 = 90$

We observe that for C_1, $g_1^{mse}(C_1) = 40.8$ and $g_1^{max}(C_1) = 50$, while for C_2, $g_1^{mse}(C_2) = 37.0$ and $g_1^{max}(C_2) = 60$.

Thus, a healthcare practitioner may choose C_2 to minimize mean squared disparity for group 1, and unintentionally maximizing the disparity for group 1 and another single group.

We can see that this edge case also exists for the group case.

Proof 5b Consider again the performance of two classifiers:

$C_1 : TPR_1 = 100, TPR_2 = 49, TPR_3 = 74, TPR_4 = 99$

$C_2 : TPR_1 = 100, TPR_2 = 50, TPR_3 = 98, TPR_4 = 99$

We observe that for C_1, $G^{mse}(C_1) = 34.2$ and $G^{max}(C_1) = 51$, while for C_2, $G^{mse}(C_2) = 34.7$ and $G^{max}(C_2) = 50$.

Thus, we see that in some cases, a classifier with the lower mean squared disparity may have higher maximum pairwise disparity.

4 Empirical Evaluation

In this section, we evaluate the multi-group fairness of various classifiers trained on a variety of clinical binary classification tasks. We quantify the likelihood each edge case appearing when these fairness expansions are used to evaluate classifiers trained on real-world medical tasks.

4.1 Data and Methods

Data We use the Medical Information Mart for Intensive Care (MIMIC-III) database, which contains electronic medical records for 38,597 distinct adults admitted to the Beth Israel Deaconess Medical Center between 2001 and 2012 [11].

Tasks All tasks use clinical notes as features. We note that the purpose of this study is not to obtain state-of-the-art performance, but to examine the bias when *reasonable* classifiers are trained on these tasks. In total, 58 different tasks are tested. The tasks were defined as follows:

1. **Post-discharge Mortality.** Using only discharge summaries for patients who had survived during their stay, we predict whether a patient will die within 1 year and within 2 years after being discharged.
2. **Phenotyping using all notes.** We follow previous work for MIMIC cohort selection [8]. The following note types were selected: "Nursing", "Nursing/other", "Physician", "Discharge summary". All available notes written during a patient's stay were concatenated. The task is to classify if a patient belongs to one of 25 HCUP CCS groups, with label linking by the ICD-9 code in the dataset [8]. We consider three additional tasks: 1) acute phenotypes; 2) chronic phenotypes; and 3) all diseases. Therefore, this task actually consists of 28 separate binary classification problems.
3. **Phenotyping using first 48 hours.** Following the same cohort selection procedure, we drop discharge notes, and limit notes to those written during the first 48 hours of a patient's stay. We use the same 28 binary classification tasks defined previously.

Methods We examine predictive performance disparities with respect to ethnicity, a particularly relevant example in healthcare. Patients are stratified into six self-reported categories: White, Black, Hispanic/Latino, Asian, other, and unknown.

To obtain features for the notes, we use TF-IDF weighting on a bag-of-words representation of the 50,000 most frequent tokens in the corpus. This matrix is scaled to unit variance, and six different classifiers are trained: 1) logistic regression with L1 or L2 regularization (LR); 2) neural network with 1 hidden layer (NN); 3) random forest (RF); 4) AdaBoost (ADA) [10]; and 5) XGBoost (XGB) [3].

For each classifier, we conduct a grid search (see Appendix A), and select the classifier with the lowest log loss on the validation set. Performance is evaluated on a held-out test set. The metrics examined were: 1) predicted prevalence, 2) recall, 3) precision, 4) specificity.

Individual group edge case detection To detect edge cases for each expansion, for an individual, we define an edge case event as:

1.1. **Combining all other groups:**

$$\epsilon_1 = 0.01, \epsilon_2 = 0.05,\ \exists i, j\ \ |g_i^{comb}| < \epsilon_1, |TPR_i - TPR_j| > \epsilon_2$$

1.2. **Absolute Average disparity:**

$$\exists C_1, C_2, i\ \ g_i^{avg}(C_1) < g_i^{avg}(C_2), g_i^{max}(C_1) > g_i^{max}(C_2)$$

1.3. **Maximum disparity:**[2]

$$\exists C_1, C_2, i\ \ g_i^{max}(C_1) < g_i^{max}(C_2), g_i^{avg}(C_1) > g_i^{avg}(C_2)$$

[2] This edge case is equivalent to the previous edge case since C_1 and C_2 are interchangeable.

1.4. **Mean Squared disparity:**

$$\exists C_1, C_2, i \;\; g_i^{mse}(C_1) < g_i^{mse}(C_2), g_i^{max}(C_1) > g_i^{max}(C_2)$$

Classifer edge Case detection To detect edge cases for each expansion for a classifier we define the following as edge cases if:

2.1. **Absolute Average disparity:**

$$\exists C_1, C_2 \;\; G^{avg}(C_1) < G^{avg}(C_2), G^{max}(C_1) > G^{max}(C_2)$$

2.2. **Maximum disparity:** [3]

$$\exists C_1, C_2 \;\; G^{max}(C_1) < G^{max}(C_2), G^{avg}(C_1) > G^{avg}(C_2)$$

2.3. **Mean Squared disparity:**

$$\exists C_1, C_2 \;\; G^{mse}(C_1) < G^{mse}(C_2), G^{max}(C_1) > G^{max}(C_2)$$

4.2 Results

Different fairness gap definitions exhibit high correlation There is a high correlation between all pairings of expansions at the classifier level (i.e., G^{avg} with G^{max} and likewise with G^{mse}), with Spearman correlations greater than 0.99. The largest deviation occurs when a metric approaches a value of 1. Appendix B shows a pair plot of the three classifier-level multi-group expansions, plotted across the 58 tasks performed and 5 classifiers trained for each task.

Thus, in the large majority of cases, the three multi-group fairness definitions should be interchangeable. The caution arises when the classifiers exhibit extremely high bias.

Edge cases appear frequently at the classifier level. For each of the tasks trained, we compare the classifier level bias between all pairs of the five classifiers for each of the four metrics. We flag a task if at least one combination of classifiers exhibits an edge case (see Table 1). First, it can be seen that edge cases for the absolute average and maximum disparities occur more often than the mean squared disparity. Second, when comparing the performance between five different classifiers, the edge cases are actually fairly common - 34% of the tasks exhibited the absolute average edge case when recall is used as the metric.

Metric	# Edge Cases (2.1, 2.2)	# Edge Cases (2.3)	# Possibilities
Recall	20	13	58
Specificity	12	8	58
Precision	16	18	58
Predicted Prevalence	18	13	58

Table 1: Out of all 58 tasks, the number of which contained at least one edge case as defined in cases 2.1 to 2.3, when all $\binom{5}{2}$ pairs of classifiers are compared for each task.

Edge cases appear frequently at the individual group level. For case 1.1, looking at all the classifiers trained for each task, we evaluated the gap using the *combining all other groups* method for each of the six ethnicity groups. We flag a classifier if any combination of ethnicities satisfies the edge case, Table 2. We see that for these particular tasks, the edge case occurs quite frequently for the recall and precision gaps, up to 32% of the time. Moreover, Table 3 shows the average performance of each classifier on the 58 tasks. There does not seem to be a correlation between the performance of a model on a task, and how often the edge case 1.1 occurs when the bias of the classifier is evaluated between different ethnicities.

To evaluate cases 1.2 to 1.4, for each task, for each ethnicity, we evaluate all pairs of classifiers evaluated on that ethnicity, and see if any pair of classifiers satisfy the edge cases presented. A specific task, ethnicity combination is flagged if at least one pair of classifiers satisfy the edge case. Table 4 shows the number of such edge cases flagged. Similar to the group scenario, it is observed that the absolute average disparity edge case occurs more often than the mean squared disparity case. Also, these edge cases occur quite frequently - when comparing the individual group biases of five classifiers, the absolute average edge case occurs up to 50% of the time.

[3] This edge case is equivalent to the previous edge case since C_1 and C_2 are interchangeable.

	Recall	Specificity	Precision	Predicted Prevalence
# Edge Cases	90	15	94	22
# Possibilities	290	290	290	290
# from LR	14	2	22	7
# from NN	18	3	18	3
# from RF	17	2	12	2
# from ADA	26	4	23	4
# from XGB	15	4	19	6

Table 2: Out of all 290 possible combinations of tasks and classifiers, the number of which contained at least one edge case as defined in case 1.1, when all pairs of ethnicities are compared. Also shows the number of edge cases flagged for each classifier.

	LR	NN	RF	ADA	XGB
AUROC	0.862 (0.064)	0.772 (0.082)	0.854 (0.069)	0.849 (0.067)	0.882 (0.058)
AUPRC	0.639 (0.182)	0.480 (0.209)	0.629 (0.179)	0.626 (0.187)	0.682 (0.172)

Table 3: Average performance of each classifier across the 58 binary classification tasks on the test set. One standard deviation of the mean is shown in parentheses.

Metric	# Edge Cases (1.2, 1.3)	# Edge Cases (1.4)	# Possibilities
Recall	174	131	348
Specificity	112	71	348
Precision	159	107	348
Predicted Prevalence	161	114	348

Table 4: Out of all 348 possible combinations of tasks and ethnicity groups, the number of which contained at least one edge case as defined in cases 1.2 to 1.4, when all pairs of classifiers are compared.

5 Discussion and Conclusion

In this work we have examined the effects of different multi-group expansions of fairness definitions in a rigorous and robust manner. To our knowledge, we are the first to highlight edge-cases where choosing between classifiers based on an expansion may result in sub-optimal results per some other expansion. Our novel analysis demonstrates the need for further research regarding multi-group fairness definitions.

The purpose of our work is not to advocate for nor dissuade from the usage of any specific multi-group expansion for fairness definitions. Our aim is to simply highlight that the expansions used in the literature have edge cases whereby minimizing one unfairness metric can maximize another. We believe that researchers working to improve the fairness of their algorithms should be aware of this fact, and utilize more than one expansion when comparing classifiers as to avoid inadvertently making decisions with unintended outcomes. Furthermore, more justification should be used when choosing one expansion over another.

Our work is limited by the expansions we chose to compare. By looking at only four expansions, we may have missed other important edge cases. Our expansions are also not compatible with definitions of fairness that do not explicitly consider protected groups (e.g., individual fairness).

Our rigorous empirical analysis demonstrates how often such edge cases manifest. The results of our analysis is limited to our specific dataset and the methods utilized. It may be the case that other datasets or methods will result in a different prevalence of edge cases. It is therefore incumbent upon developers and researchers to ensure that their decisions between classifiers take such edge cases into account.

References

[1] Anand Avati, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H Shah. 2018. Improving palliative care with deep learning. *BMC medical informatics and decision making* 18, 4 (2018), 122.

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

[3] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. *CoRR* abs/1603.02754 (2016). arXiv:1603.02754 `http://arxiv.org/abs/1603.02754`

[4] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.

[5] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).

[6] Josh Gardner, Christopher Brooks, and Ryan Baker. 2019. Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. ACM, 225–234.

[7] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[8] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data* 6, 1 (2019), 96.

[9] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *International Conference on Machine Learning*. 1934–1943.

[10] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. 2009. Multi-class adaboost. *Statistics and its Interface* 2, 3 (2009), 349–360.

[11] 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 (2016), 160035.

[12] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (10 2016), 14–19. `https://doi.org/10.1111/j.1740-9713.2016.00960.x`

[13] Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. *arXiv preprint arXiv:1904.05233* (2019).

[14] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

A Gridsearch Hyperparameters

Classifier	Hyperparameter	Values
Logistic Regression	Penalty	$\{L_1, L_2\}$
	C	{1e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 1e-1, 5e-1, 1e0, 1e1}
Random Forest	Max depth	{2, 3, 5, 7, 10, None}
AdaBoost	# Estimators	{100, 200, 300, 500}
NN with 1 hidden layer	# neurons in hidden layer	{200, 400}
	L2 regularization parameter	{1e-4, 1e-3}
XGBoost	Max depth	{2, 3, 5, 7, 10, None}

Table 5: Hyperparameter grids used for the grid search when optimizing each classifier.

B Pairwise Correlation Analysis

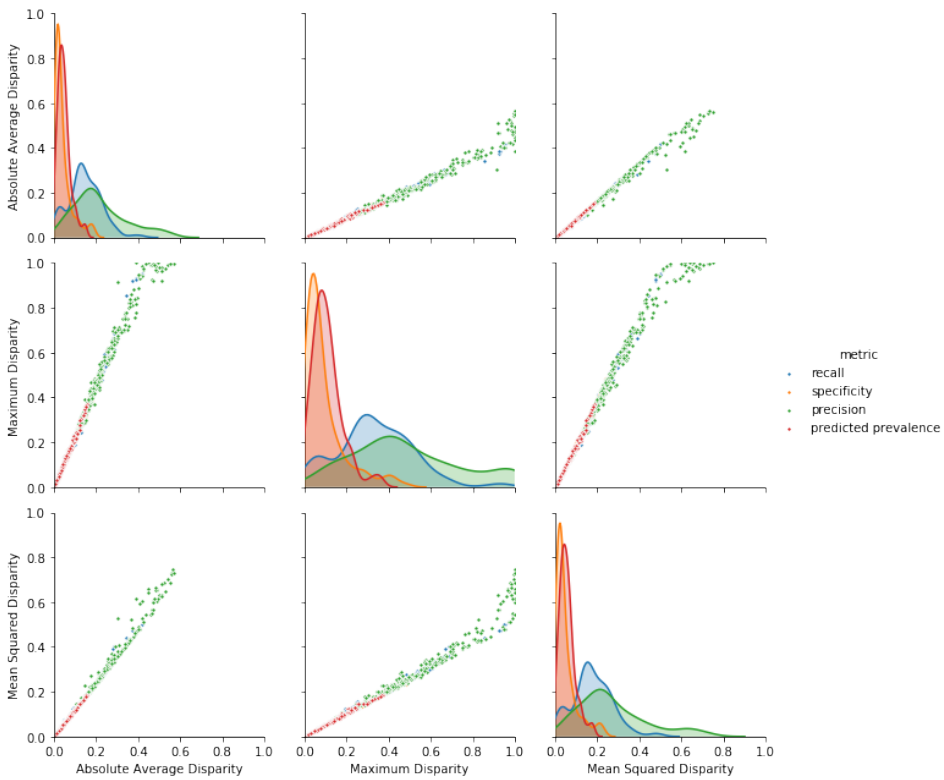

Figure 1: We use ethnicity as an example where the multi-group fairness expansion is particularly relevant in healthcare. For each expansion, we demonstrate a pair plot presenting the correlation between different classifier performance metrics, and the distribution of performance metrics. Measures are aggregated across the five different classifiers and the 58 tasks performed.

C Desirable Properties

In this section, we list desirable properties for multi-group expansions. First we define the following:

$$\text{Let } h_1, h_2 \text{ be two classifiers}$$
$$\text{Let } \mathcal{Z} \text{ be the set of all protected groups}$$
$$\text{Let } P_{i,j}(h_1) \text{ be the set relevant performance gap for protected groups } i, j \in \mathcal{Z} \text{ given classifier } h_1$$
$$\text{Let } G(h_1) \text{ be an expansion function which takes in a classifier } h_1 \text{ and returns a single metric } G(h_1)$$

Desirable Property #1: Perfect Fairness

$$G(h_1) = 0 \iff \forall i, j \in \mathcal{Z}, P_{i,j}(h_1) = 0 \tag{2}$$

The expansion metric of a classifier is only 0 if there is no disparity between any two groups as measured by the fairness metric.

Desirable Property #2: Maximum Fairness

$$G(h_1) < G(h_2) \iff \max_{i,j} P_{i,j}(h_1) < \max_{i,j} P_{i,j}(h_2) \tag{3}$$

The expansion metric of a classifier is only less the same expansion of another classifier if the largest disparity between any two groups for the first classifier is less than the largest disparity between any two groups for the second classifier.

Desirable Property #3: Monotonicity

$$\text{Define } h_1 \text{ as a classifier.}$$
$$\text{Define } h_2 \text{ as a classifier with performance equivalent to } h_1$$
$$\text{Modify } h_2 \text{ only so that } P_{i,j}(h_1) < P_{i,j}(h_2)$$
$$\text{Then an expansion satisfying this property would have } G(h_1) < G(h_2)$$

If two classifiers have the exact same gap between all groups except for one pairing where the second classifier has a larger gap than the first classifier then the expansion metric of the second classifier should be larger than the expansion metric of the first classifier.

D Thoughts

Research Question: How can we effectively choose between two (or more) multi-group classifiers with respect to fairness?

Group Fairness Group Fairness compares performance disparities in a pairwise fashion between groups. Show that comparing two classifier using the same fairness metric, the classifier which minimizes one aggregation methods can result maximize another aggregation method. Perform empirical analysis to see how often this happens

Individual Fairness By disregarding group labels we no longer have to worry about different aggregation functions, and are more easily able to obtain a single number for our classifier. However, ignoring groups does not solve the issue. Show that it is possible that two classifiers have the same unfairness metric but one affects all/both groups equally, and one disproportional affects one group, by comparing classifiers with the same individual fairness metric.

Subgroup Fairness A middle ground of the previous two approaches is to consider all subgroup fairness groupings. This prevents fairness gerrymandering, but, *I think*, it suffers from the same problems as Group Fairness. TODO: Show that two classifiers with the same subgroup fairness metrics still hurt one larger group over another.

Unified Approach Fairness By defining the fairness of classifiers as the sum of individual, between-group, and inter-group unfairness, the metric tries to better capture the trends here. However, as the unfairness is composed of three terms, it would probably be possible to conceive a setting where one classifier has large unfairness in a single measure, while another classifier has it evenly spread out. Should that one measure be the between class unfairness, then we'd have the same issues as with the above measurements.