Fair admission risk prediction with proportional multicalibration

William G. La Cava* WILLIAM.LACAVA@CHILDRENS.HARVARD.EDU
Elle Lett ELLE.LETT@CHILDRENS.HARVARD.EDU
Guangya Wan GWAN@HSPH.HARVARD.EDU
Computational Health Informatics Program, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Abstract

Fair calibration is a widely desirable fairness criteria in risk prediction contexts. One way to measure and achieve fair calibration is with multicalibration. Multicalibration constrains calibration error among flexibly-defined subpopulations while maintaining overall calibration. However, multicalibrated models can exhibit a higher percent calibration error among groups with lower base rates than groups with higher base rates. As a result, it is possible for a decision-maker to learn to trust or distrust model predictions for specific groups. To alleviate this, we propose *proportional multicalibration*, a criteria that constrains the percent calibration error among groups and within prediction bins. We prove that satisfying proportional multicalibration bounds a model's multicalibration as well its *differential calibration*, a fairness criteria that directly measures how closely a model approximates sufficiency. Therefore, proportionally calibrated models limit the ability of decision makers to distinguish between model performance on different patient groups, which may make the models more trustworthy in practice. We provide an efficient algorithm for post-processing risk prediction models for proportional multicalibration and evaluate it empirically. We conduct simulation studies and investigate a real-world application of PMC-postprocessing to prediction of emergency department patient admissions. We observe that proportional multicalibration is a promising criteria for controlling simultaneous measures of calibration fairness of a model over intersectional groups with virtually no cost in terms of classification performance.

Data and Code Availability This paper uses the MIMIC-IV-ED dataset (Johnson et al., 2021), which is available on the PhysioNet repository (Goldberger et al., 2000). Code for the experiments is available here: https://github.com/cavalab/proportional-multicalibration.

Institutional Review Board (IRB) This research does not require IRB approval.

1. Introduction

Today, machine learning (ML) models have an impact on outcome disparities across sectors due to their wide-spread use in decision-making. When applied in clinical decision support (CDS), ML models help care providers decide whom to prioritize to receive finite and time-sensitive resources among a population of potentially very ill patients. These resources include hospital beds (Barak-Corren et al., 2021a; Dinh and Berendsen Russell, 2021), organ transplants (Schnellinger et al., 2021), specialty treatment programs (Henry et al., 2015; Obermeyer et al., 2019), and, recently, ventilator and other breathing support tools to manage the COVID-19 pandemic (Riviello et al., 2022).

In scenarios like these, decision makers typically rely on risk prediction models to be *calibrated*. Calibration measures the extent to which a model's risk scores, R, match the observed probability of the event, $P(y)$ (Brier et al., 1950). Perfect calibration implies that $P(y|R = r) = r$, for all values of r. Calibration allows the risk scores to be used to rank patients in order of priority and informs care providers about the urgency of treatment. However, models that are not equally calibrated among subgroups defined by different sensitive attributes (race, ethnicity, gender, income, etc.) may lead to systematic denial of resources to marginalized groups (e.g. (Obermeyer et al., 2019; Ashana et al., 2021; Roberts, 2011; Zelnick et al., 2021; Ku et al., 2021)). For example, Obermeyer et al. (2019) analyzed a large health sys-

* Corresponding author. https://cavalab.org
Published version: https://proceedings.mlr.press/v209/la-cava23a.html

© 2023 W.G. La Cava, E. Lett & G. Wan.

tem algorithm used to enroll high-risk patients into care management programs and showed that, at a given risk score, Black patients exhibited significantly poorer health than white patients.

When evidence of algorithmic bias is observed, for example in estimates of kidney function (Diao et al., 2021), it is unclear how care should be adjusted for affected patient groups. Variance in clinical judgements that are not evidence-based are themselves a source of unfairness, referred to as unwarranted clinical variation (Harrison et al., 2019; Sutherland and Levesque, 2020; Newhouse and Garber, 2013). Ideally, patients with the same evidence, e.g. model risk prediction, would receive the same standard of care. In this work, we propose measures and methods designed to reduce unwarranted variation in care that may arise from biased prediction models.

To address equity in calibration, Hebert-Johnson et al. (2018) proposed a fairness measure called *multicalibration* (MC), which asks that calibration be satisfied simultaneously over many flexibly-defined subgroups. Remarkably, MC can be satisfied efficiently by post-processing risk scores without negatively impacting the generalization error of a model, unlike other fairness concepts like demographic parity (Foulds and Pan, 2020) and equalized odds (Hardt et al., 2016). This has motivated the use of MC in practical settings (e.g. Barda et al. (2021)) and has spurred several extensions (Kim et al., 2019; Jung et al., 2021; Gupta et al., 2021; Gopalan et al., 2022). If we bin our risk predictions, the MC criteria specifies that, for every group within each bin, the absolute difference between the mean observed outcome and the mean of the predictions should be small.

As Barocas et al. (2019) note, fair calibration reduces to a more general fairness notion dubbed *sufficiency*. Under sufficiency, the expected outcome should be independent of group membership, conditioned on the risk prediction. We see this criteria as highly desirable in CDS contexts. Sufficiency eliminates a source of uncertainty from decision-making: how to interpret a model recommendation from a model that has been observed to perform more or less well on certain subpopulations. Given a risk score from a model satisfying sufficiency, a decision maker cannot distinguish between patient outcomes on the basis of their group membership. Thus, they cannot justify variations from the model's recommendations on the basis of patient identity alone.

In this work, we start by assessing the conditions under which MC satisfies sufficiency. To do so, we derive a fairness criteria directly from sufficiency, *differential calibration* (DC). DC is an extension of differential fairness (Foulds et al., 2019b), both named due to their relation to differential privacy (Dwork and Roth, 2013). DC constrains ratios of population risk between groups within risk prediction bins. We show that DC measures the extent to which one satisfies sufficiency in an interpretable way. In short, among patients assigned the same risk score from a model satisfying ε-DC, the outcome is at most e^{ε} more likely in one group compared to any another. A low ε-DC thereby constrains the amount by which a decision maker may learn to unequally trust the model for different groups.

By relating sufficiency to MC, we describe a shortcoming of MC that can occur when the outcome probabilities are strongly tied to group membership. Under this condition, the amount of calibration error *relative to the expected outcome* can be unequal between groups. This inequality hampers the ability of MC to (approximately) guarantee sufficiency except by setting extremely low error thresholds, resulting in the need for large numbers of updates.

We propose a simple variant of MC called *proportional multicalibration* (PMC) that instead requires the proportion of calibration error within each bin and group to be small. We prove that PMC bounds both multicalibration and differential calibration. We show that PMC can be satisfied with an efficient post-processing method, similarly to MC (Pfisterer et al., 2021). Proportionally multicalibrated models thereby obtain robust fairness guarantees that are less dependent on population risk categories. It does so in fewer steps than MC by prioritizing updates to groups with high proportional calibration error.

Finally, we investigate the application of these methods to predicting patient admissions in the emergency department, a real-world resource allocation task that is targeted by current CDS models (Barak-Corren et al., 2021a). We create a benchmark dataset for this task using the recently released MIMIC-IV emergency department dataset Johnson et al. (2021), and benchmark PMC- and MC-based postprocessing approaches. We show that post-processing for PMC results in models that are accurate, multicalibrated, and differentially calibrated.

2. Reconciling Multicalibration and Sufficiency

2.1. Preliminaries

We consider the task of training a risk prediction model for a population of individuals with outcomes, $y \in \{0, 1\}$, and features, $x \in \mathcal{X}$. Let D be the joint distribution from which individual samples (y, x) are drawn. We assume the outcomes y are random samples from underlying independent Bernoulli distributions, denoted as $p^*(x) \in [0, 1]$. Individuals can be further grouped into *collections of subsets*, $\mathcal{C} \subseteq 2^{\mathcal{X}}$, such that $S \in \mathcal{C}$ is the subset of individuals belonging to S, and $x \in S$ indicates that individual x belongs to group S. We denote our risk prediction model as $R(x) : \mathcal{X} \to [0, 1]$.

In order to consider calibration in practice, the risk predictions are typically discretized into bins. We represent discretization by a parameter, λ, that specifies the width of the bins. As an example, $\lambda = 0.1$ corresponds to decile bins. For brevity, proofs and some formal definitions in the following sections are given in Appendices A.2 and A.3.2.

2.2. Multicalibration

Hebert-Johnson et al. (2018) define multicalibration by first defining calibration with respect to a subset (i.e., group) of individuals:

Definition 1 (α-calibration) *Let $S \subseteq \mathcal{X}$. For $\alpha \in [0, 1]$, R is* α-calibrated *with respect to S if there exists some $S' \subseteq S$ with $|S'| \geq (1 - \alpha)|S|$ such that for all $r \in [0, 1]$,*

$$\left| \mathop{\mathbb{E}}_{D}[y|R = r, x \in S'] - r \right| \leq \alpha.$$

MC then guarantees that α-calibration holds over every subset from a collection of subsets:

Definition 2 (α-Multicalibration) *Let $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be a collection of subsets of $\mathcal{X}$, $\alpha \in [0, 1]$. A predictor R is α-multicalibrated on $\mathcal{C}$ if for all $S \in \mathcal{C}$, R is α-calibrated with respect to S.*

We note that, according to Definition 1, a model need only be calibrated over a sufficiently large subset of each group (S') in order to satisfy the definition. This relaxation is used to maintain a satisfactory definition of MC when working with discretized predictions. For simplicity, we conduct most of our analysis using the continuous versions of fairness definitions like Definition 2 (see Appendix A.3.1 for an extended discussion).

MC is one of few approaches to achieving fairness that does not require a significant trade-off to be made between a model's generalization error and the improvement in fairness it provides. As Hébert-Johnson et al. (2018) show, this is because achieving multicalibration is not at odds with achieving accuracy in expectation for the population as a whole. This separates calibration fairness from other fairness constraints like demographic parity and equalized odds (Hardt et al., 2016), both of which may denigrate the performance of the model on specific groups (Chouldechova, 2017; Pleiss et al., 2017). In clinical settings, such trade-offs may be difficult or impossible to justify. In addition to its alignment with accuracy in expectation, Hébert-Johnson et al. (2018) propose an efficient post-processing algorithm for MC similar on boosting. We discuss additional extensions to MC in Appendix A.1.

2.3. Measuring Sufficiency via Differential Calibration

MC provides a sense of fairness by approximating *calibration by group*, which is perfectly satisfied when $P_D(y|R = r, x \in S) = r$ for all $S \in C$ and $r \in [0, 1]$. Calibration by group is closely related to the *sufficiency* fairness criterion (Barocas et al., 2019). Sufficiency states that the outcome probability is independent from $\mathcal{C}$, conditioned on the risk score. In the binary group setting ($\mathcal{C} = \{S_i, S_j\}$), we can express sufficiency as

$$P_D(y|R, x \in S_i) = P_D(y|R, x \in S_j) \tag{1}$$

or equivalently,

$$P_D(y|R, x \in S_i)/P_D(y|R, x \in S_j) = 1.$$

Unlike calibration by group, sufficiency does not stipulate that the risk scores be calibrated, yet from a fairness perspective, sufficiency and calibration by group are equivalent (Barocas et al., 2019). In both cases, the sense of *fairness* stems from the desire for R to capture everything about group membership that is relevant to predicting y.

Under sufficiency, the risk score is equally informative of the outcome, regardless of group membership. Because of this, given a risk score from a model satisfying sufficiency, a decision maker cannot distinguish

between patient outcomes on the basis of their group membership. In this sense, a model satisfying sufficiency provides an added level of trust in deployment: we know that a decision maker cannot justify different decisions on the basis of the patient's identity for the same risk prediction. If risk prediction models satisfied sufficiency, it would eliminate the need for group-specific decision protocol given recommendations from the same model. Consider, for example, the decision-making uncertainty related to estimates of kidney function (Diao et al., 2021).

Below, we define an approximate measure of sufficiency that constrains pairwise differentials between groups, and accomodates binned predictions:

Definition 3 (ε-Differential Calibration) *Let $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be a collection of subsets of $\mathcal{X}$. A model $R(x)$ is ε-differentially calibrated with respect to $\mathcal{C}$ if, for all pairs $(S_i, S_j) \in \mathcal{C} \times \mathcal{C}$ for which $P_D(S_i), P_D(S_j) > 0$, for any $r \in [0, 1]$,*

$$e^{-\varepsilon} \leq \frac{\mathbb{E}_D[y|R = r, x \in S_i]}{\mathbb{E}_D[y|R = r, x \in S_j]} \leq e^{\varepsilon}. \tag{2}$$

By inspection we see that ε in ε-DC measures the extent to which R satisifies sufficiency. That is, when $P(y|R = r, x \in S_i) \approx P(y|R = r, x \in S_j)$ for all pairs (S_i, S_j), $\varepsilon \approx 0$. ε-DC requires that, for any risk score, the outcome y is at most e^{ε} times more likely among one group than another, and a minimum of $e^{-\varepsilon}$ less likely.

Definition 3 fits the general definition of a *differential fairness* measure proposed by Foulds et al. (2019a) and previously used to study demographic parity criteria (Foulds and Pan, 2020). We describe the relation in more detail in Appendix A.1.1, including Eq. (2)'s connection to differential privacy (Dwork and Lei, 2009) and pufferfish privacy (Kifer and Machanavajjhala, 2014).

Taken alone, DC does not prevent a decision-maker from equally distrusting a model for all patients. That is because it does not guarantee the model is globally calibrated . Rather, it makes it harder to distinguish the calibration quality of the model between groups.

2.4. The differential calibration of multicalibrated models is limited by low-risk groups

At a basic level, the form of MC and sufficiency differ: MC constrains absolute differences between groups across prediction bins, whereas sufficiency constrains pairwise differentials between groups. To reconcile MC and DC/sufficiency more formally, we pose the following question: if a model satisfies α-MC, what, if anything does this imply about the ε-DC of the model? (In Appendix A.4,Theorem 22, we answer the inverse question). We now show that multicalibrated models have a bounded DC, but that this bound is limited by small values of R.

Theorem 4 *Let $R(x)$ be a model satisfying α-MC on a collection of subsets $\mathcal{C} \in 2^{\mathcal{X}}$. Let $r_{min} = \min_{S \in \mathcal{C}} \mathbb{E}_D[R|R = r, x \in S]$ be the minimum expected risk prediction among $S \in \mathcal{C}$ and $r \in [0, 1]$. Then $R(x)$ is $\left(\ln \frac{r_{min}+\alpha}{r_{min}-\alpha}\right)$-differentially calibrated.*

Theorem 4 illustrates the important point that, *in terms of percentage error*, *MC* does not provide equal protection to groups with different risk profiles. Imagine a model satisfying (0.05)-MC for groups $S \in \mathcal{C}$. Consider individuals receiving model predictions $R(x) = 0.9$. MC guarantees that, for any category $\{x : x \in S, R(x) = 0.9\}$, the expected outcome probability is at least $0.9 - \alpha = 0.85$ and at most $0.9 + \alpha = 0.95$. This bounds the percent error among groups with this prediction to about 6%. In contrast, consider individuals for whom $R(x) = 0.3$; each group may have a true outcome probability as low as 0.25, which is an error of 20% - about 3.4x higher than the percent error in the higher-risk group.

In Appendix Theorem 22, we show that differentially calibrated models can bound multicalibration only if they are also α-calibrated (Definition 1).

3. Proportional Multicalibration

We are motivated to define a measure that is efficiently learnable like MC (Definition 2) but better aligned with the multiplicative interpretation of sufficiency, like DC (Definition 3). To do so, we define PMC, a variant of MC that constrains the proportional calibration error of a model among subgroups and risk strata. In this section, we show that bounding a model's PMC is enough to meaningfully bound DC and MC. Furthermore, we provide an efficient algorithm for satisfying PMC based on a simple extension of MC/Multiaccuracy boosting (Kim et al., 2019). We begin by defining proportional calibration, which expresses calibration error as a percentage of the outcome probability among a group.

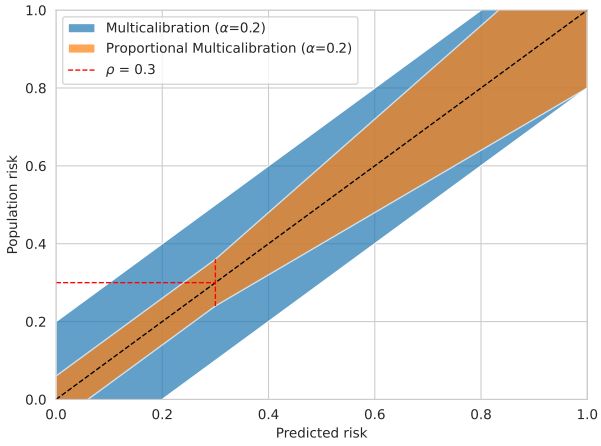

Figure 1: Visualization of the constraint differences between multicalibration and proportional multicalibration (PMC). The filled area represents the maximum the predicted risk can deviate from the fraction of positives in the population, for any group. Below ρ, the constraint is constant for PMC.

Definition 5 (α-Proportional Calibration) *Let $S \subseteq \mathcal{X}$. For $\alpha > 0$, R is* α-proportionally calibrated *with respect to S if there exists some $S' \subseteq S$ with $|S'| \geq (1-\alpha)|S|$ such that for all $r \in [0,1]$,*

$$\left| \mathop{\mathbb{E}}_{D}[y|R = r, x \in S'] - r \right| \leq \alpha \mathop{\mathbb{E}}_{D}[y|R = r, x \in S'].$$

Proportional multicalibration is then defined by requiring Definition 5 be satisified among a collection of groups:

Definition 6 (α-Proportional Multicalibration) *Let $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be a collection of subsets of $\mathcal{X}$, $\alpha > 0$. A predictor R is* α-proportionally multicalibrated *on $\mathcal{C}$ if for all $S \in \mathcal{C}$, R is α-proportionally calibrated with respect to S.*

We also define a discretized version of PMC in Appendix A.3.2 that is useful for implementing the measure in Algorithm 1 and measuring PMC in our experiments. In Appendix A.3.1, we show (α, λ)-PMC meaningfully bounds α-PMC under different discretizations, such that we can minimize (α, λ)-PMC to achieve low α-PMC.

In practice, we must ensure that the outcome probability for any (group, prediction bin) category is greater than zero for PMC to be meaningful. We later introduce a lower bound, ρ, to prevent the outcome probability from being too small. It is common in clinical settings to only show a risk prediction to a decision maker if it exceeds some threshold; in those settings, ρ can be set to match this threshold.

Comparison to Differential Calibration Rather than constraining the differentials of prediction- and group- specific outcomes among all pairs of subgroups in $\mathcal{C} \times \mathcal{C}$ as in DC (Definition 3), PMC constrains the relative error of each group in $\mathcal{C}$. In practical terms, this makes it more efficient to calculate PMC by a factor of $O(|\mathcal{C}|)$ steps compared to DC. In addition, PMC constrains each group's calibration with respect to ground truth, whereas DC only constrains the differentials *between* groups. We formalize the relationship between these two measures below.

Theorem 7 *Let $R(x)$ be a model satisfying (α)-PMC on a collection $\mathcal{C}$. Then $R(x)$ is $(\ln \frac{1+\alpha}{1-\alpha})$-differentially calibrated.*

Theorem 7 demonstrates that α-proportionally multicalibrated models satisfy a straightforward notion of differential fairness that depends monotonically only on α. Because PMC bounds DC, proportionally calibrated models inherit desirable properties of differentially calibrated models, especially the following:

Corollary 8 *Let $R(x)$ be a model satisfying (α)-PMC on a collection $\mathcal{C}$. Then $R(x)$ satisfies $(\ln \frac{1+\alpha}{1-\alpha})$-pufferfish privacy with respect to $\mathcal{C}$.*

Corollary 8 follows from the definitions of ε-DC and pufferfish privacy (Appendix Definition 13, Kifer and Machanavajjhala (2014)). This property is perhaps best understood as a guarantee of trust rather than privacy. Namely, it guarantees that the patient's group identity does not provide useful information to a decision-maker in deciding how much to trust model predictions. The patient's group identity is in this sense "private" from the model's calibration. Unlike many privacy-preserving algorithms, we do not need to add add any noise to the model to satisfy DC; conversely, it is achieved by making the model better calibrated for some groups.

Comparison to Multicalibration Rather than constraining the absolute difference between risk predictions and the outcome as in MC, PMC requires

that the calibration error be a small fraction of the expected risk in each category (S, I). In this sense, it provides a stronger protection than MC by requiring calibration error to be a small fraction regardless of the risk group. In many contexts, we would argue that this is also more aligned with the notion of fairness in risk prediction contexts. Under MC, the underlying probability of an outcome within a group affects the fairness protection that is received (i.e., the percentage error that Definition 2 allows). Because underlying probabilities of many clinically relevant outcomes vary significantly among subpopulations, multicalibrated models may systematically permit higher percentage error to specific groups. The difference in relative calibration error among populations with different risk profiles also translates in weaker sufficiency guarantees, as demonstrated in Theorem 4. In contrast, PMC provides a fairness guarantee that is less dependent on subpopulation risks.

In the following theorem, we show that MC is also constrained when a model satisfies PMC.

Theorem 9 *Let $R(x)$ be a model satisfying α-PMC on a collection $\mathcal{C}$. Then $R(x)$ is $(\frac{\alpha}{1-\alpha})$-multicalibrated on $\mathcal{C}$.*

The proof of Theorem 9 is given in Appendix A.2. This theorem implies that a proportionally calibrated model with sufficiently low α will satisfy a similarly low value of MC. We further discuss and illustrate the bounds given by Theorems 4, 7 and 9 in Appendix A.3.

Because PMC bounds a model's multicalibration, proportionally multicalibrated models also bound the generalization error of a predictor such that it is close to the "best in class" predictions.

Corollary 10 *Let $R(x)$ be a model satisfying (α)-PMC on a collection $\mathcal{C}$. Let $\mathcal{H}$ be a set of predictors, p^* the outcome-generating distribution, and $h^* = \arg\min_{h\in\mathcal{H}} ||h - p^*||^2$. Then*

$$||R - p^*||^2 - ||h^* - p^*||^2 < \frac{6\alpha}{1-\alpha}.$$

This guarantee follows directly from Theorem 5 in Hebert-Johnson et al. (2018). In a nutshell, it means that proportionally multicalibrated models are bound to be close to the best predictor of their hypothesis class $(\mathcal{H})$. This contrasts fair calibration with other approaches to fairness that typically form a trade-off with a model's overall error. We further illustrate relationship between PMC, DC, and MC in Appendix Fig. 6.

3.1. Learning proportionally multicalibrated predictors

In Algorithm 1, we propose an extension of MCBoost (Pfisterer et al., 2021) to efficiently update risk predictors to satisfy PMC. Algorithm 1 works by checking for calibration errors among groups and prediction intervals that violate the user threshold, and adjusting these predictions towards the target. PMCBoost differs in two main ways: first, it updates whenever calibration error is not within $\alpha\bar{y}$ for all categories, as opposed to simply within α. Second, it ignores updates for categories with low outcome probability (less than ρ). Next, we prove that PMCBoost learns an (α,λ)-PMC model in a polynomial number of steps.

Proposition 11 *Define $\alpha, \lambda, \gamma, \rho > 0$. Let $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be a collection of subsets of $\mathcal{X}$ such that, for all $S \in \mathcal{C}$, $P_D(S) > \gamma$. Let $R(x)$ be a risk prediction model to be post-processed. For all $(S, I) \in \mathcal{C} \times \Lambda_\lambda$, let $E[y|R \in I, x \in S] > \rho$. There exists an algorithm that satisfies (α, λ)-PMC with respect to $\mathcal{C}$ in $O(\frac{|C|}{\alpha^3\lambda^2\rho^2\gamma})$ steps.*

We analyze Algorithm 1 and show it satisfies Proposition 11 in Appendix A.2. This more stringent threshold requires an additional $O(\frac{1}{\rho^2})$ steps compared to the algorithm for MC, where $\rho > 0$ is a lower bound on the expected outcome within a category (S, I).

Achieving proportional multicalibration with MCBoost Another way to achieve α_P-PMC is to use the existing MCBoost algorithm, but setting $\alpha_M = \rho\alpha_P$. In other words, setting a very low value for α_M should also satisfy Theorem 6 because, if α_M can be made small enough, the calibration error on all categories will be small compared to the outcome prevalence, $\mathbb{E}_D[y|R \in I, x \in S]$. However, to achieve PMC guarantees by MCBoost requires a large number of unnecessary updates for high risk groups, since the DC and PMC of multicalibrated models are limited by low-risk groups (Theorem 4). Furthermore, the number of steps in MCBoost (and PMCBoost scales as an inverse high-order polynomial of α (cf. Thm. 2 (Hebert-Johnson et al., 2018)).

We can compare the algorithm complexity of MCBoost and PMCBoost to get a better comparison. Say we would like to satisfy α-PMC using MCBoost.

Hebert-Johnson et al. (2018) show that MCBoost takes $O(\frac{|C|}{\alpha^3\lambda^2\gamma})$ steps to complete. We must reduce α-MC by a factor of ρ to satisfy α-PMC; this increases the number of MCBoost steps by a factor of $(\frac{1}{\rho^3})$. For a reasonable value of ρ (say, $\rho = 0.1$), this takes 1000x more steps, and corresponds to 10x more steps than it would take to satisfy α-PMC with PMCBoost.

Now consider the reverse: satisfying α-MC using PMCBoost. Because of Theorem 9, we need only reduce α-PMC by a factor of $\frac{1}{1+\alpha\text{-MC}}$, which increases the number of PMCBoost steps by $(1+\alpha\text{-MC})^3$. For example, to achieve α-MC $= 0.1$, PMCBoost takes about 1.3x more steps than MCBoost.

Algorithm 1: PMCBoost

Input: predictor $R(x)$;
parameters $\alpha, \lambda, \gamma, \rho > 0$;
$\mathcal{C} \in 2^{\mathcal{X}}$ such that for all $S \in \mathcal{C}, P_D(S) \geq \gamma$;
$\mathcal{D} = \{(y, x)_i\}_{i=0}^{N} \sim D$
Output: Updated predictor $R(x)$
repeat
  $\{(y, x)\}$ sample $\mathcal{D}$
  for $S \in \mathcal{C}, I \in \Lambda_\lambda$ *such that* $P_D(R \in I, x \in S) \geq \alpha\lambda\gamma$ **do**
    $S_r \leftarrow S \cap \{x : R(x) \in I\}$
    $\bar{r} \leftarrow \frac{1}{|S_r|}\sum_{x \in S_r} R(x)$ ▷ average group prediction
    $\bar{y} \leftarrow \frac{1}{|S_r|}\sum_{x \in S_r} y(x)$ ▷ average subgroup risk
    $\Delta r \leftarrow \bar{y} - \bar{r}$
    if $\bar{y} \geq \rho$ **then**
      cutoff $\leftarrow \alpha\bar{y}$
    else ▷ limit low probability updates
      cutoff $\leftarrow \alpha\rho$
    end
    if $|\Delta r| \geq$ *cutoff* **then**
      $R(x) \leftarrow R(x) + \Delta r$ for all $x \in S_r$
      $R(x) \leftarrow \text{squash}(R(x), [0, 1])$ ▷ limit to $[0, 1]$
    end
  end
until *No updates to* $R(x)$

4. Experiments

In our first set of experiments (Section 4), we study MC and PMC in simulated population data to understand and validate the analysis in previous sections. In the second section, we compare the performance of varied model treatments on a real world hospital admission task, using an implementation of Algorithm 1. We make use of empirical versions of our fairness definitions which we refer to as *MC loss*, *PMC loss*, and *DC loss*. In short, these measures calculate the maximum (proportional) calibration error or pairwise calibration differential among subgroups and risk categories in the data sample. Due to space constraints the formal definitions are given in Appendix A.3.2 (Definitions 19, 20 and 21).

Simulation study We simulate data from α-multicalibrated models. For simplicity, we specify a data structure with a one-to-one correspondence between subset and model estimated risk, such that for all x in S, $R(x) = R(x|x \in S) = R(S)$. Therefore all information for predicting the outcome based on the features in x is contained in the attributes $\mathcal{A}$ that define subgroup S. Outcome probability is specified as $p_i^* = P_D(y|x \in S_i) = 0.2 + 0.01(i-1)$ and $i = 1, \cdots, N_s$, where N_s is the number of subsets S, defined by $\mathcal{A}$ and indexed by i with increasing p^*. For each group, $R_i = R(S_i) = R(x|x \in S_i) = p_i^* - \Delta_i$. We randomly select Δ_i for one group to be $\pm\alpha$ and for the remaining groups, $\Delta_i = \pm\delta$, where $\delta \sim$ Uniform(min $= 0$, max $= \alpha$). In all cases, the sign of Δ_i is determined by a random draw from a Bernoulli distribution. For these simulations we set $N_S = 61$ and $\alpha = 0.1$, such that $p_i^* \in [0.2, 0.8]$ and $R_i \in [0.1, 0.9]$. We generate $N_{sim} = 1000$ simulated datasets, with $n = 1000$ observations per group, and for each S_i, we calculate the ratio of the absolute mean error to p_i^*, i.e. the PMC loss function for this data generating mechanism.

We also simulate three specific scenarios where: 1) $|\Delta_i|$ is equivalent for all groups (Fixed); 2) $|\Delta_i|$ increases with increasing p_i^*; and 3) $|\Delta_i|$ decreases with increasing p_i^*, with $\alpha = 0.1$ in each case. These scenarios compare when α is determined by all groups, the group with the lowest outcome probability, and the group with the highest outcome probability, respectively.

Hospital admission Next, we test PMC alongside other methods in application to prediction of inpatient hospital admission for patients visiting the emergency department (ED). Overcrowding and long wait times in EDs have been shown to increase odds of inpatient death (5%, CI 2-8%), length of stay (0.8%, CI 0.5-1%), and costs per admission (1%, CI 0.7-2%) (Sun et al., 2013). The burden of overcrowding and long wait times in EDs is significantly higher among non-white, non-Hispanic patients and socioeconomically marginalized patients (James et al., 2005; McDonald et al., 2020).

Recent work has demonstrated risk prediction models that can expedite patient visits by predicting patient admission at an early stage of a visit with a high degree of certainty (AUC ≥ 0.9 across three large care centers) (Barak-Corren et al., 2017b,a, 2021b,a). Our goal is to ensure no group of patients will be over- or under-prioritized over another by these models,

Table 1: Admission prevalence (Admissions/Total (%)) among patients in the MIMIC-IV-ED data repository, stratified by the intersection of ethnoracial group and gender.

Gender Ethnoracial Group	F	M	Overall
American Indian/Alaska Native	70/257 (27%)	82/170 (48%)	152/427 (36%)
Asian	1043/3595 (29%)	1032/2384 (43%)	2075/5979 (35%)
Black/African American	3124/27486 (11%)	2603/14458 (18%)	5727/41944 (14%)
Hispanic/Latino	1063/10262 (10%)	1168/5795 (20%)	2231/16057 (14%)
Other	1232/5163 (24%)	1479/3849 (38%)	2711/9012 (30%)
Unknown/Unable to Obtain	1521/2156 (71%)	2074/2377 (87%)	3595/4533 (79%)
White	18147/50174 (36%)	18951/45435 (42%)	37098/95609 (39%)
Overall	26200/99093 (26%)	27389/74468 (37%)	53589/173561 (31%)

Table 2: Features used in the hospital admission task.

Description	Features
Vitals	temperature, heartrate, resprate, o2sat, systolic blood pressure, diastolic blood pressure
Triage Acuity	Emergency Severity Index (Tanabe et al., 2004)
Check-in Data	chief complaint, self-reported pain score
Health Record Data	no. previous visits, no. previous admissions
Demographic Data	ethnoracial group, gender, age, marital status, insurance, primary language

Table 3: Parameters for the hospital admission prediction experiment.

Parameter	Values
tolerance (α)	(0.001, 0.01, 0.1, 0.2)
min group probability (γ)	(0.05, 0.1)
binning (λ)	0.1
min outcome probability (ρ)	(0.01, 0.1)
Base Model	LR, RF, DNN
Groups	(race/ethnicity, gender), (race/ethnicity, gender, insurance product)

which could exacerbate the treatment and outcome disparities that currently exist.

We construct a prediction task similar to previous studies but using a new data resource: the MIMIC-IV-ED repository (Johnson et al., 2021). After data preparation (see Appendix A.5), the cohort consists of 173,561 visits with demographics and admission statistics given in Table 1. Table 2 shows the list of features used for prediction, modelled off prior work (Barak-Corren et al., 2021a). In Table 1 we observe stark differences in admission rates by demographic group and gender, suggesting that the use of a proportional measure of calibration could be appropriate for this task. We trained and evaluated ℓ_1-penalized logistic regression (LR), random forest (RF), and deep neural network (DNN) models of patient admission, with and without post-processing with MCBoost (Pfisterer et al., 2021) or PMCBoost. We varied α, γ, and ρ to characterize parameter sensitivity among the methods (Table 3). For each of the parameter settings, we conducted 100 repeat experiments with different shuffles of the data. Comparisons are reported on a test set of 20% of the data for each trial. Additional experiment details are available in Appendix A.5.

5. Results

Fig. 2 shows the results of our simulation study. The results indicate that, without the proportionality factor, α-multicalibrated models exhibit a dependence between the group prevalence and the amount of proportional calibration loss. The results demonstrate why α-MC alone is not sufficient to achieve sufficiency, particularly when outcome probabilities vary by group.

Results on the hospital admission prediction task are summarized in Figs. 3 to 5 and Table 4. As Table 4 shows, PMCBoost has a small effect on predictive performance (ΔAUROC $< 0.6\%$) while improving DC loss and PMC loss by 10-81% across LR, RF, and DNN models. Somewhat surprisingly, PMCBoost improves MC loss more effectively than MCBoost for both DNN (19% versus 11%) and RF models (18% versus 6%). We observe that PMCBoost and MCBoost do not improve MC loss of LR models, although PMCBoost improves other metrics of fairness for LR (PMC loss and DC loss) without changing AUROC. In Fig. 4, we illustrate the calibration improvement of RF models using PMCBoost for

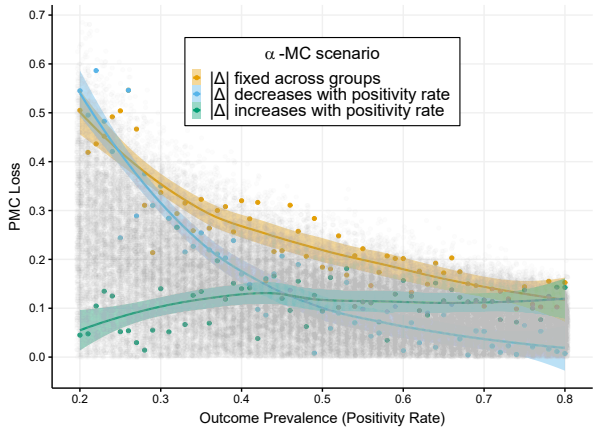

Figure 2: The relationship between MC, PMC, and outcome prevalence as illustrated via a simulation study in which the rates of the outcome are associated with group membership. Gray points denote the PMC loss of a (0.1)-MC model on 1000 simulated datasets, and colored lines denote three scenarios in which each group's calibration error ($|\Delta|$) varies. Although MC is identical in all scenarios, PMC loss is higher among groups with lower positivity rates in most scenarios unless the groupwise calibration error increases with positivity rate.

Table 4: Median percent change in loss from the base model using MC- and PMCBoost.

	ML	AUROC Δ%	MC Δ%	PMC Δ%	DC Δ%
MCBoost	DNN	-0.58	-10.83	-16.27	-2.41
	LR	-0.01	10.95	-9.50	-6.46
	RF	-0.09	-5.91	-55.44	-13.88
PMCBoost	DNN	-0.56	-19.49	-28.83	-9.91
	LR	0.01	14.18	-37.90	-14.90
	RF	-0.27	-17.53	-80.82	-23.52

Table 5: For MCBoost and PMCBoost, we compare the average number of updates and wall clock time (s) taken to train for the equivalent values of α-PMC.

ML	α_{PMCe}	Postprocessing	Iterations	Time (s)
LR	0.001	MC	2094.0	678.0
		PMC	263.0	320.0
	0.010	MC	631.0	223.0
		PMC	153.0	202.0
RF	0.001	MC	1996.0	556.0
		PMC	323.0	234.0
	0.010	MC	713.0	197.0
		PMC	256.0	167.0

three patient subgroups defined by the intersection of race/ethnicity, gender, and insurance type.

Computational Cost As mentioned in Section 3.1, MCBoost may be configured to satisfy α-PMC by setting α to a much smaller value, although this may take an excessive number of steps. To better understand this trade-off, we empirically compared MC- and PMCBoost by the number of steps required for each to reach their best performance in Fig. 5. On average, MCBoost requires approximately 5x more updates to achieve similar performance on PMC loss as PMCBoost, due to its dependence on very small values of α. Wall clock times scale similarly, as detailed in Table 5.

Sensitivity Analysis In Appendix A.6, we look at detailed performance comparisons of MCBoost and PMCBoost over values of α and group definitions in Figs. 8 to 10. We observe that, while low values of α for MCBoost improve its PMC loss performance, PMCBoost typically performs as well or better, particularly for larger values of α, and does so in fewer iterations.

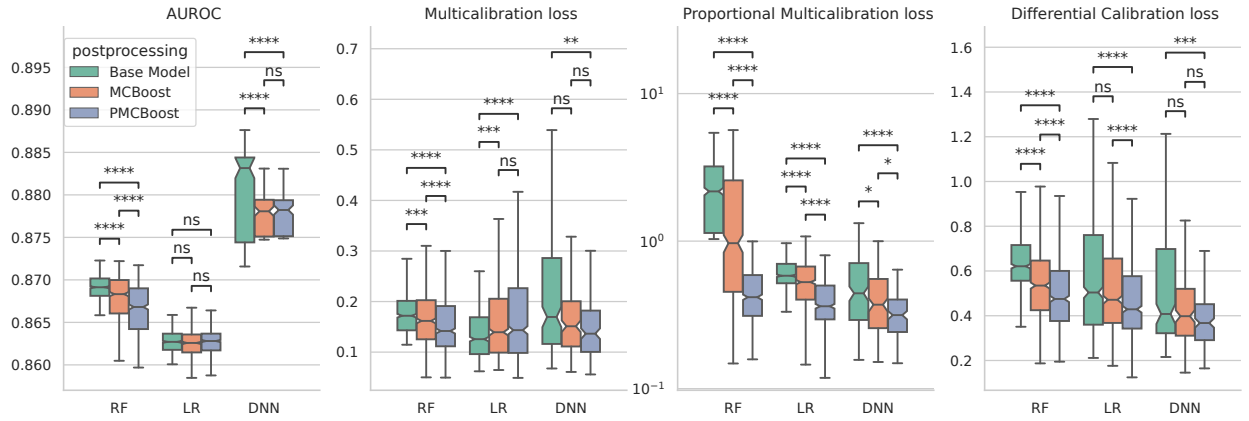

Figure 3: A comparison of LR, RF, and DNN models, with and without MCBoost and PMCBoost, on the hospital admission task. From left to right, trained models are compared in terms of test set AUROC, MC loss, PMC loss, and DC loss. Points represent the median performance over 100 shuffled train/test splits with bootstrapped 99% confidence intervals. We test for significant differences between post-processing methods using two-sided Wilcoxon rank-sum tests with Bonferroni correction. ns: $p <= 1$; **: 1e-03 $< p <=$ 1e-02; ***: 1e-04 $< p <=$ 1e-03; ****: $p <=$ 1e-04.

6. Discussion and Conclusion

In this paper we have analyzed multicalibration through the lens of sufficiency and differential calibration to reveal the sensitivity of this metric to correlations between outcome rates and group membership. We have proposed a measure, PMC, that alleviates this sensitivity and attempts to capture the "best of both worlds" of MC and sufficiency. PMC provides equivalent percentage calibration protections to groups regardless of their risk profiles, and in so doing, bounds a model's differential calibration. We provide an efficient algorithm for learning PMC predictors by postprocessing a given risk prediction model. On a real-world and clinically relevant task (admission prediction), we have shown that postprocessing three types of models with PMC leads to better performance across all three fairness metrics, with a small impact on predictive performance.

Our preliminary analysis suggests PMC can be a valuable metric for training fair algorithms in resource allocation contexts. Future work could extend this analysis on both the theoretical and practical side. On the theoretical side, the generalization properties of the PMC measure should be established and its sample complexity quantified, as Rose (2018) did with MC. Additional extensions of PMC could establish a bound on the accuracy of PMC-postprocessed models in a similar vein to work by Kim et al. (2019) and Hebert-Johnson et al. (2018). On the empirical side, future works should benchmark PMC on a larger set of real-world problems, and explore use cases in more depth. In addition, user studies could be employed to validate whether proportionally multicalibrated models do in fact instill more trust in decision-makers compared to baseline models.

Acknowledgments

E.L. was partially supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) grant 5 T32 HD40128-19. W.G.L. was partially supported by the National Institutes of Health (NIH) National Library of Medicine grant R00-LM012926.

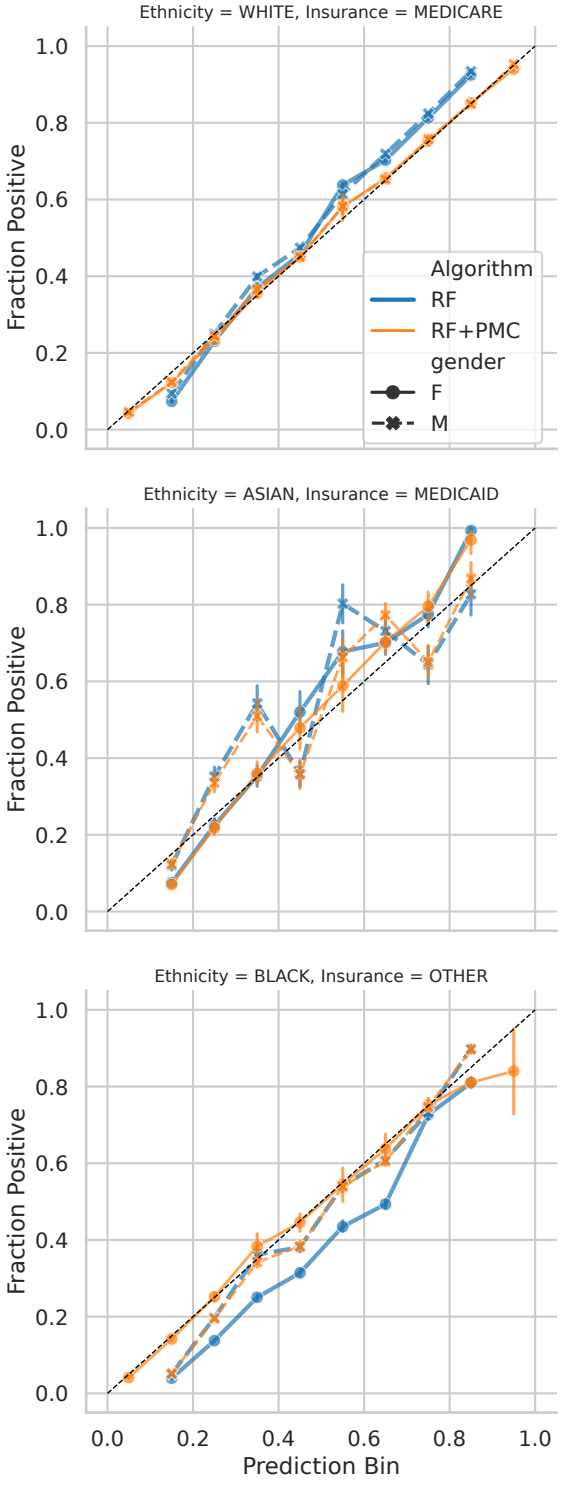

Figure 4: Calibration curves for RF models with and without PMCBoost postprocessing.

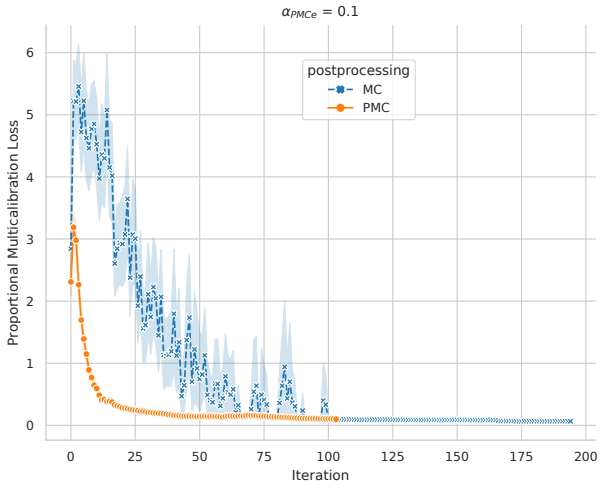

Figure 5: Number of iterations to convergence for PMCBoost and MCBoost when optimizing for the same value of α-PMC (α_{PMCe}).

References

Deepshikha Charan Ashana, George L Anesi, Vincent X Liu, Gabriel J Escobar, Christopher Chesley, Nwamaka D Eneanya, Gary E Weissman, William Dwight Miller, Michael O Harhay, and Scott D Halpern. Equitably allocating resources during crises: Racial differences in mortality prediction models. *American journal of respiratory and critical care medicine*, 204(2):178–186, 2021.

Nikhil Bansal and Anupam Gupta. Potential-function proofs for gradient methods. *Theory of Computing*, 15(1):1–32, 2019.

Yuval Barak-Corren, Andrew M. Fine, and Ben Y. Reis. Early Prediction Model of Patient Hospitalization From the Pediatric Emergency Department. *Pediatrics*, 139(5), May 2017a. ISSN 1098-4275. doi: 10.1542/peds.2016-2785.

Yuval Barak-Corren, Shlomo Hanan Israelit, and Ben Y Reis. Progressive prediction of hospitalisation in the emergency department: Uncovering hidden patterns to improve patient flow. *Emergency Medicine Journal*, 34(5):308–314, May 2017b. ISSN 1472-0205, 1472-0213. doi: 10.1136/emermed-2014-203819.

Yuval Barak-Corren, Isha Agarwal, Kenneth A Michelson, Todd W Lyons, Mark I Neuman, Susan C Lipsett, Amir A Kimia, Matthew A Eisenberg, Andrew J Capraro, Jason A Levy, Joel D Hudgins, Ben Y Reis, and Andrew M Fine. Prediction of patient disposition: Comparison of computer and human approaches and a proposed synthesis. *Journal of the American Medical Informatics Association*, 28(8):1736–1745, July 2021a. ISSN 1527-974X. doi: 10.1093/jamia/ocab076.

Yuval Barak-Corren, Pradip Chaudhari, Jessica Perniciaro, Mark Waltzman, Andrew M. Fine, and Ben Y. Reis. Prediction across healthcare settings: A case study in predicting emergency department disposition. *npj Digital Medicine*, 4(1): 1–7, December 2021b. ISSN 2398-6352. doi: 10.1038/s41746-021-00537-x.

Noam Barda, Gal Yona, Guy N. Rothblum, Philip Greenland, Morton Leibowitz, Ran Balicer, Eitan Bachmat, and Noa Dagan. Addressing bias in prediction models by improving subpopulation calibration. *Journal of the American Medical Informatics Association: JAMIA*, 28(3):549–558, March 2021. ISSN 1527-974X. doi: 10.1093/jamia/ocaa283.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, and Daniele Regoli. The zoo of Fairness metrics in Machine Learning. *arXiv:2106.00467 [cs, stat]*, June 2021.

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv:1703.00056 [cs, stat]*, February 2017.

Alexandra Chouldechova and Aaron Roth. The Frontiers of Fairness in Machine Learning. *arXiv:1810.08810 [cs, stat]*, October 2018.

Patricia Hill Collins. *Black Feminist Though: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge, 1 edition, September 1990.

Patricia Hill Collins and Sirma Bilge. *Intersectionality*. John Wiley & Sons, 2020. ISBN 1-5095-3969-7.

Kimberle Crenshaw. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, 1989(1):31, 1989.

Kimberle Crenshaw. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6):1241, July 1991. ISSN 00389765. doi: 10.2307/1229039. URL https://www.jstor.org/stable/1229039?origin=crossref.

James A Diao, Gloria J Wu, Herman A Taylor, John K Tucker, Neil R Powe, Isaac S Kohane, and Arjun K Manrai. Clinical implications of removing race from estimates of kidney function. *JAMA : the journal of the American Medical Association*, 325(2):184–186, 2021.

Michael M Dinh and Saartje Berendsen Russell. Overcrowding kills: How COVID-19 could reshape emergency department patient flow in the new normal. *Emergency Medicine Australasia*, 33(1): 175–177, 2021. ISSN 1742-6723. doi: 10.1111/1742-6723.13700.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, pages 371–380, 2009.

Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9 (3-4):211–407, 2013. ISSN 1551-305X, 1551-3068. doi: 10.1561/0400000042.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1115-1. doi: 10.1145/2090236.2090255.

Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Learning from Outcomes: Evidence-Based Rankings. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 106–125, November 2019. doi: 10.1109/FOCS.2019.00016.

James Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An Intersectional Definition of Fairness. *arXiv:1807.08362 [cs, stat]*, September 2019a.

James Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An Intersectional Definition of Fairness. *arXiv:1807.08362 [cs, stat]*, September 2019b. URL http://arxiv.org/abs/1807.08362. arXiv: 1807.08362.

James R. Foulds and Shimei Pan. Are Parity-Based Notions of AI Fairness Desirable? *Data Engineering*, page 51, 2020.

A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–220, June 2000. ISSN 1524-4539. doi: 10.1161/01.cir.101.23.e215.

Parikshit Gopalan, Michael P. Kim, Mihir Singhal, and Shengjia Zhao. Low-Degree Multicalibration. *arXiv:2203.01255 [cs]*, March 2022.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M. Pai, and Aaron Roth. Online Multivalid Learning: Means, Moments, and Prediction Intervals, January 2021.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 501–512, Barcelona Spain, January 2020. ACM. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372826.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3315–3323. Curran Associates, Inc., 2016.

Reema Harrison, Elizabeth Manias, Stephen Mears, David Heslop, Reece Hinchcliff, and Liz Hay. Addressing unwarranted clinical variation: A rapid review of current evidence. *Journal of Evaluation in Clinical Practice*, 25(1):53–65, 2019. ISSN 1365-2753. doi: 10.1111/jep.12930.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1939–1948. PMLR, July 2018.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Calibration for the (Computationally-Identifiable) Masses. *arXiv:1711.08513 [cs, stat]*, March 2018.

Katharine E. Henry, David N. Hager, Peter J. Pronovost, and Suchi Saria. A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine*, 7(299): 299ra122, August 2015. ISSN 1946-6242. doi: 10.1126/scitranslmed.aab3719.

Catherine A. James, Florence T. Bourgeois, and Michael W. Shannon. Association of Race/Ethnicity with Emergency Department Wait Times. *Pediatrics*, 115(3):e310–e315, March 2005. ISSN 0031-4005. doi: 10.1542/peds.2004-1541.

Heinrich Jiang and Ofir Nachum. Identifying and Correcting Label Bias in Machine Learning, January 2019.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. MIMIC-IV-ED, 2021.

Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Moment Multicalibration for Uncertainty Estimation. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 2634–2678. PMLR, July 2021.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 110, Barcelona, Spain, January 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3373154.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *arXiv:1711.05144 [cs]*, December 2018.

Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)*, 39(1):1–36, 2014.

Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Elaine Ku, Charles E McCulloch, Deborah B Adey, Libo Li, and Kirsten L Johansen. Racial disparities in eligibility for preemptive waitlisting for kidney transplantation and modification of eGFR thresholds to equalize waitlist time. *Journal of the American Society of Nephrology*, 32(3):677–685, 2021.

Erica J. McDonald, Matthew Quick, and Mark Oremus. Examining the Association between Community-Level Marginalization and Emergency Room Wait Time in Ontario, Canada. *Healthcare Policy*, 15(4):64–76, May 2020. ISSN 1715-6572. doi: 10.12927/hcpol.2020.26223.

Joseph P. Newhouse and Alan M. Garber. Geographic Variation in Medicare Services. *New England Journal of Medicine*, 368(16):1465–1468, April 2013. ISSN 0028-4793. doi: 10.1056/NEJMp1302981.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aax2342.

Florian Pfisterer, Christoph Kern, Susanne Dandl, Matthew Sun, Michael P. Kim, and Bernd Bischl. Mcboost: Multi-Calibration Boosting for R. *Journal of Open Source Software*, 6(64):3453, 2021.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On Fairness and Calibration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5680–5689. Curran Associates, Inc., 2017.

Elisabeth D. Riviello, Tenzin Dechen, Ashley L. O'Donoghue, Michael N. Cocchi, Margaret M. Hayes, Rose L. Molina, Nicole H. Moraco, Anne Mosenthal, Michael Rosenblatt, Noa Talmor, Daniel P. Walsh, David N. Sontag, and Jennifer P. Stevens. Assessment of a Crisis Standards of Care Scoring System for Resource Prioritization and Estimated Excess Mortality by Race, Ethnicity, and Socially Vulnerable Area During a Regional Surge in COVID-19. *JAMA Network Open*, 5 (3):e221744, March 2022. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2022.1744.

Dorothy Roberts. *Fatal invention: How science, politics, and big business re-create race in the twenty-first century.* New Press/ORIM, 2011. ISBN 1-59558-691-1.

Sherri Rose. Machine learning for prediction in electronic health data. *JAMA network open*, 1(4): e181404–e181404, 2018.

Erin M Schnellinger, Edward Cantu, Michael O Harhay, Douglas E Schaubel, Stephen E Kimmel, and Alisa J Stephens-Shields. Mitigating selection bias in organ allocation models. *BMC medical research methodology*, 21(1):1–9, 2021.

Benjamin C. Sun, Renee Y. Hsia, Robert E. Weiss, David Zingmond, Li-Jung Liang, Weijuan Han, Heather McCreath, and Steven M. Asch. Effect of Emergency Department Crowding on Outcomes of Admitted Patients. *Annals of Emergency Medicine*, 61(6):605–611.e6, June 2013. ISSN 01960644. doi: 10.1016/j.annemergmed.2012.10.026.

Kim Sutherland and Jean-Frederic Levesque. Unwarranted clinical variation in health care: Definitions and proposal of an analytic framework. *Journal of Evaluation in Clinical Practice*, 26(3):687–696, 2020. ISSN 1365-2753. doi: 10.1111/jep.13181.

Paula Tanabe, Rick Gimbel, Paul R. Yarnold, Demetrios N. Kyriacou, and James G. Adams. Reliability and validity of scores on The Emergency Severity Index version 3. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, 11(1):59–65, January 2004. ISSN 1069-6563. doi: 10.1197/j.aem.2003.06.013.

Leila R Zelnick, Nicolae Leca, Bessie Young, and Nisha Bansal. Association of the estimated glomerular filtration rate with vs without a coefficient for race with time to eligibility for kidney transplant. *JAMA network open*, 4(1):e2034004–e2034004, 2021.

Appendix A. Appendix

In this section, we include additional comparisons to related work, additional definitions, proofs to the theorems in the main text, and additional experimental details. The code to reproduce the figures and experiments is available here: https://github.com/cavalab/proportional-multicalibration.

A.1. Related Work

Definitions of Fairness There are myriad ways to measure fairness that are covered in more detail in other works (Barocas et al., 2019; Chouldechova and Roth, 2018; Castelnovo et al., 2021). We briefly review three notions here. The first, *demographic parity*, requires the model's predictions to be independent of patient demographics (A). Although a model satisfying demographic parity can be desirable when the outcome should be unrelated to sensitive attributes (Foulds and Pan, 2020), it can be unfair if important risk factors for the outcome are associated with those attributes (Hardt et al., 2016). For example, it may be more fair to admit socially marginalized patients to a hospital at a higher rate if they are assessed less able to manage their care at home. Furthermore, if the underlying rates of illness vary demographically, requiring demographic parity can result in a healthier patients from one group being admitted more often than patients who urgently need care.

When the base rates of admission are expected to differ demographically, we can instead ask that the model's errors be balanced across groups. One such notion is *equalized odds*, which states that for a given Y, the model's predictions should be independent of A. Satisfying equalized odds is equivalent to having equal FPR and FNR for every group in A.

When the model is used for patient risk stratification, as in the target use case in this paper, it is important to consider a model's calibration for each demographic group in the data. Because risk prediction models influence who is prioritized for care, an unfairly calibrated model can systematically underpredict risk for certain demographic groups and result in under-allocation of patient care to those groups. Thus, guaranteeing group-wise calibration via an approach such as multicalibration also guarantees fair patient prioritization for health care provision. In some contexts, risk predictions are not directly interpreted, but only used to *rank* patients, which in some contexts is sufficient for resource allocation. Authors have proposed various ways of measuring the fairness of model rankings, for example by comparing AUROC between groups (Kallus et al., 2020).

Approaches to Fairness Many approaches to achieving fairness guarantees according to demographic parity, equalized odds and its relaxations have been proposed (Dwork et al., 2012; Hardt et al., 2016; Berk et al., 2017; Jiang and Nachum, 2019; Kearns et al., 2018). When choosing an approach, is important to carefully weigh the relative impact of false positives, false negatives, and miscalibration on patient outcomes, which differ by use case. When group base rates differ (i.e., group-specific positivity rates), *equalized odds and calibration by group cannot both be satisfied* (Kleinberg et al., 2016). Instead, one can often equalized multicalibration while satisfying relaxations of equalized odds such as *equalized accuracy*, where $Accuracy = \mu TPR + (1-\mu)(1-FPR)$ for a group with base rate μ. However, to do so may require denigrating the performance of the model on specific groups (Chouldechova, 2017; Pleiss et al., 2017), which is unethical in our context.

As mentioned in the introduction, we are also motivated to utilize approaches to fairness that 1) dovetail well with intersectionality theory, and 2) provide privacy guarantees. Most work in the computer science/ machine learning space does not engage with

the broader literature on socio-cultural concepts like intersectionality, which we see as a gap that makes adoption in real-world settings difficult (Hanna et al., 2020). One exception to this statement is differential fairness (Foulds et al., 2019a), a measure designed with intersectionality in mind. In addition to being a definition of fairness that provides equal protection to groups defined by intersections of protected attributes, models satisfying ε-differential fairness also satisfy ε-pufferfish privacy. This privacy guarantee is very desirable in risk prediction contexts, because it limits the extent to which the model reveals sensitive information to a decision maker that has the potential to influence their interpretation of the model's recommendation. However, prior work on differential fairness has been limited to using it to control for demographic parity, which is not an appropriate fairness measure for our use case (Foulds and Pan, 2020).

Multicalibration has inspired several extensions, including relaxations such as multiaccuracy (Kim et al., 2019), low-degree multicalibration (Gopalan et al., 2022), and extensions to conformal prediction and online learning (Jung et al., 2021; Gupta et al., 2021). Noting that multicalibration is a guarantee over mean predictions on a collection of groups $\mathcal{C}$, Jung et al. (2021) propose to extend multicalibration to higher-order moments (e.g., variances), which allows one to estimate a confidence interval for the calibration error for each category. Gupta et al. (2021) extend this idea and generalize it to the online learning context, in which an adversary chooses a sequence of examples for which one wishes to quantify the uncertainty of different statistics of the predictions. Recent work has also utilized higher order moments to "interpolate" between the guarantees provided by multiaccuracy, which only requires accuracy in expectation for groups in $\mathcal{C}$, and multicalibration, which requires accuracy in expectation at each prediction interval (Kim et al., 2019). Like proportional multicalibration (Definition 6), definitions of multicalibration for higher order moments provide additional criteria for quantifying model performance over many groups; in general, however, much of the focus in other work is on statistics for uncertainty estimation. Like these works, one may view our proposal for proportional multicalibration as alternative definition of what it means to be multicalibrated. The key difference is that proportional multicalibration measures the degree to which multicalibration depends on differences in outcome prevalence between groups, and in doing so provides guarantees of pufferfish privacy and differential calibration.

Dwork et al. (2019) study the relation of fair rankings to multicalibration, and, in a similar vein to differential fairness measures, formulate a fairness measure for group rankings using the relations between pairs of groups. However, these definitions are specific to the ranking relation between the groups, whereas differential calibration cares only about the outcome differential (conditioned on model predictions) between pairs of groups.

A.1.1. Differential Fairness

DF was explicitly defined to be consistent with the social theoretical framework of *intersectionality*. This framework dates back as early as the social movements of the '60s and '70s (Collins and Bilge, 2020) and was brought into the academic mainstream by pioneering work from legal scholar Kimberlé Crenshaw (Crenshaw, 1989, 1991) and sociologist Patricia Hill Collins (Collins, 1990). Central to intersectionality is that hierarchies of power and oppression are structural elements that are fundamental to our society. Through an intersectional lens, these power structures are viewed as interacting and co-constituted, inextricably related to one another. To capture this viewpoint, DF (Foulds et al., 2019a) constrains the differential of a general data mechanism among all pairs of groups, where groups are explicitly defined as the intersections of protected attributes in $\mathcal{A}$.

Definition 12 (ε-differential fairness) *(Foulds et al., 2019a) Let Θ denote a set of distributions and let $x \sim \theta$ for $\theta \in \Theta$. A mechanism $M(x)$ is ε-differentially fair with respect to ($\mathcal{C}$,Θ) for all $\theta \in \Theta$ with $x \sim \theta$, and $m \in Range(M)$ if, for all $(S_i, S_j) \in \mathcal{C} \times \mathcal{C}$ where $P(S_i|\theta) > 0$, $P(S_j|\theta) > 0$,*

$$e^{-\varepsilon} \leq \frac{P_{M,\theta}(M(x) = m|S_i, \theta)}{P_{M,\theta}(M(x) = m|S_j, \theta)} \leq e^{\varepsilon} \tag{3}$$

Definition 13 (Pufferfish Privacy) *Let the collection of subsets $\mathcal{C}$ represent sets of secrets. A mechanism $M(x)$ is* ε-pufferfish private *(Kifer and Machanavajjhala, 2014) with respect to $(\mathcal{C}, \Theta)$ if for all $\theta \in \Theta$ with $x \sim \theta$, for all secret pairs $(S_i, S_j) \in \mathcal{C} \times \mathcal{C}$ and $y \in Range(M)$,*

$$e^{-\varepsilon} \leq \frac{P_{M,\theta}(M(x) = y|S_i, \theta)}{P_{M,\theta}(M(x) = y|S_j, \theta)} \leq e^{\varepsilon} \ , \tag{4}$$

when S_i and S_j are such that $P(S_i|\theta) > 0$, $P(S_j|\theta) > 0$.

Note on pufferfish and differential privacy Although Eq. (3) is notable in its similarity to differential privacy (Dwork and Lei, 2009), they differ in important ways. Differential privacy aims to limit the amount of information learned about any one individual in a database by computations performed on the data (e.g. $M(x)$). Pufferfish privacy only limits information learned about the group membership of individuals as defined by $\mathcal{C}$. Kifer and Machanavajjhala (2014) describe in detail the conditions under which these privacy frameworks are equivalent.

Efficiency Property Foulds et al. (2019a) also define an interesting property of ε-differential fairness that allows guarantees of higher order (i.e., marginal) groups to be met for free; the property is given in Appendix A.3.2.

Definition 14 (Efficiency Property) *(Foulds et al., 2019a) Let $M(x)$ be an ε-differentially fair mechanism with respect to $(\mathcal{C}, \Theta)$. Let the collection of subsets $\mathcal{C}$ group individuals according to the Cartesian product of attributes $A \subseteq \mathcal{A}$. Let $\mathcal{G}$ be any collection of subsets that groups individuals by the Cartesian product of attributes in A', where $A' \subset A$ and $A' \neq \emptyset$. Then $M(x)$ is ε-differentially fair in $(\mathcal{G}, \Theta)$.*

The authors call this the "intersectionality property", although in practice it guarantees the reverse: if a model satisfies ε-DF for the low level (i.e. intersectional) groups in $\mathcal{C}$, then it satisfies ε-DF for every higher-level (i.e. marginal) group. For example, if a model is (ε)-differentially fair for intersectional groupings of individuals by race and sex, then it is ε-DF for the higher-level race and sex groupings as well. Whereas the number of intersections grows exponentially as additional attributes are protected (Kearns et al., 2018), the number of total possible subgroupings grows at a larger combinatorial rate: for p protected attributes, we have $\sum_{k=1}^{p} \binom{p}{k} m_a^k$ groups, where m_a is the number of levels of attribute a.

Limitations To date, analysis of DF for predictive modeling has been limited to defining $R(x)$ as the mechanism, which is akin to asking for *demographic parity*. Under demographic parity, one requires that model predictions be independent from group membership entirely, and this limits the utility of it as a fairness notion. Although a model satisfying demographic parity can be desirable when the outcome should be unrelated to $\mathcal{C}$ (Foulds and Pan, 2020), it can be unfair if important risk factors for the outcome are associated with demographics (Hardt et al., 2016). For example, if the underlying rates of an illness vary demographically, requiring demographic parity can result in a healthier patients from one group being admitted more often than patients who urgently need care.

A.2. Proofs for Theorems in the Main Text

Theorem 4 *Let $R(x)$ be a model satisfying α-MC on a collection of subsets $\mathcal{C} \in 2^{\mathcal{X}}$. Let $r_{min} = \min_{S \in \mathcal{C}} \mathbb{E}_D[R|R = r, x \in S]$ be the minimum expected risk prediction among $S \in \mathcal{C}$ and $r \in [0, 1]$. Then $R(x)$ is $\left(\ln \frac{r_{min}+\alpha}{r_{min}-\alpha}\right)$-differentially calibrated.*

Proof Let $r = \mathbb{E}_D[R|R = r, x \in S]$ and $p^* = \mathbb{E}_D[y|R = r, x \in S]$. α-MC guarantees that $r - \alpha \leq p^* \leq r + \alpha$ for all groups $S \in \mathcal{C}$ and predictions $r \in [0, 1]$. Plugging these lower and upper bounds into Eq. (2), we observe that the lower bound on ε-DC for $R(x)$ is given by $\frac{r+\alpha}{r-\alpha} \leq e^{\varepsilon}$. The maximum of the left-hand side for a fixed α occurs at the smallest value of r; therefore $R(x)$ satisfies $\ln \frac{r_{min}+\alpha}{r_{min}-\alpha} \leq \varepsilon$. By switching the numeratator and denominator we obtain the minimum differential and the left-hand side constraint from Definition 3, i.e. $e^{-\varepsilon} \leq \frac{r_{min}-\alpha}{r_{min}+\alpha}$. Thus $R(x)$ is $\left(\ln \frac{r_{min}+\alpha}{r_{min}-\alpha}\right)$-differentially calibrated. ■

Theorem 7 *Let $R(x)$ be a model satisfying (α)-PMC on a collection $\mathcal{C}$. Then $R(x)$ is $(\ln \frac{1+\alpha}{1-\alpha})$-differentially calibrated.*

Proof Let $r = \mathbb{E}_D[R|R = r, x \in S]$ and $p^* = \mathbb{E}_D[y|R = r, x \in S]$. If $R(x)$ satisfies α-PMC (Definition 6), then $r/(1 + \alpha) \leq p^* \leq r/(1 - \alpha)$. Solving for the upper bound on ε-DC, we immediately have $\varepsilon \geq \ln \frac{r(1+\alpha)}{r(1-\alpha)} \geq \ln \frac{1+\alpha}{1-\alpha}$. ■

Theorem 9 *Let $R(x)$ be a model satisfying α-PMC on a collection $\mathcal{C}$. Then $R(x)$ is $(\frac{\alpha}{1-\alpha})$-multicalibrated on $\mathcal{C}$.*

Proof To distinguish the parameters, let $R(x)$ be a model satisfying δ-PMC. Let $r = \mathbb{E}_D[R|R = r, x \in S]$ and $p^* = \mathbb{E}_D[y|R = r, x \in S]$. Then $r/(1 + \delta) \leq p^* \leq r/(1 - \delta)$. We solve for the upper bound on α-MC from Definition 2 for the case when $p^* > r$. This

yields

$$\begin{aligned}\alpha &\leq p^* - r \\ &\leq \frac{r}{1-\delta} - r \\ &= r\frac{\delta}{1-\delta} \\ &\leq \frac{\delta}{1-\delta}.\end{aligned}$$

We can also solve for the lower bound on α-MC from Definition 2 for the case when $p^* < r$. This yields

$$\begin{aligned}\alpha &\leq r - p^* \\ &\leq r - \frac{r}{1+\delta} \\ &= r\frac{\delta}{1+\delta} \\ &\leq \frac{\delta}{1+\delta}.\end{aligned}$$

For any $\delta > 0$, $\frac{\delta}{1-\delta} > \frac{\delta}{1+\delta}$. Therefore the first case ($p^* > r$) limits the multicalibration of $R(x)$. ■

Proposition 11 *Define $\alpha, \lambda, \gamma, \rho > 0$. Let $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be a collection of subsets of $\mathcal{X}$ such that, for all $S \in \mathcal{C}$, $P_D(S) > \gamma$. Let $R(x)$ be a risk prediction model to be post-processed. For all $(S, I) \in \mathcal{C} \times \Lambda_\lambda$, let $E[y|R \in I, x \in S] > \rho$. There exists an algorithm that satisfies (α, λ)-PMC with respect to $\mathcal{C}$ in $O(\frac{|C|}{\alpha^3\lambda^2\rho^2\gamma})$ steps.*

Proof

We show that Algorithm 1 converges using a potential function argument (Bansal and Gupta, 2019), similar to the proof techniques for the MC boosting algorithms in Hebert-Johnson et al. (2018); Kim et al. (2019). Let p_i^* be the underlying risk, R_i be our initial model, and R_i' be our updated prediction model for individual $i \in S_r$, where $S_r = \{x | x \in S, R(x) \in I\}$ and $(S, I) \in \mathcal{C} \times \Lambda_\lambda$. We use p^*, R, and R' without subscripts to denote these values over S_r. We cannot easily construct a potential argument using progress towards (α,λ)-PMC, since its derivative is undefined at $\mathbb{E}_D[y|R \in I, x \in S]$=0. Instead, we analyze progress towards the difference in the ℓ_2 norm at each step.

$$\begin{aligned}&||p^* - R|| - ||p^* - R'|| \\ &= \sum_{i \in S_r} (p_i^* - R_i)^2 - \sum_{i \in S_r} (p_i^* - \text{squash}(R_i + \Delta r))^2 \\ &\geq \sum_{i \in S_r} \left((p_i^* - R)^2 - (p_i^* - (R_i + \Delta r))^2\right) \\ &= \sum_{i \in S_r} \left(2p_i^*\Delta r - 2R_i\Delta r - \Delta r^2\right) \\ &= 2\Delta r \sum_{i \in S_r} (p_i^* - R_i) - |S_r|\Delta r^2 \qquad (5)\end{aligned}$$

From Algorithm 1 we have

$$\Delta r = \frac{1}{|S_r|} \sum_{i \in S_r} (p_i^* - R_i)$$

Substituting into Eq. (5) gives

$$||p^* - R|| - ||p^* - R'|| \geq |S_r|\Delta r^2$$

We know that $|S_r| \geq \alpha\lambda\gamma N$, and that the smallest update Δr is $\alpha\rho$. Thus,

$$||p^* - R|| - ||p^* - R'|| \geq \alpha^3\rho^2\lambda\gamma N$$

Since our initial loss, $||p^* - R||$, is at most N, Algorithm 1 converges in at most $O(\frac{1}{\alpha^3\rho^2\lambda\gamma})$ updates for category S_r.

To understand the total number of steps, including those without updates, we consider the worst case, in which only a single category S_r is updated in a cycle of the for loop (if no updates are made, the algorithm exits). Since each repeat consists of at most $|C|/\lambda$ loop iterations, this results in $O(\frac{|C|}{\alpha^3\lambda^2\rho^2\gamma})$ total steps. ■

A.3. Extended Theoretical Analysis

Illustrating Relationships between Definitions Fig. 6 shows how the definitions of MC, DC, and PMC are related. In each subplot, the x and y coordinates map the guarantee from one metric (x axis) to the implied guarantee in the other metric (y axis).

The right panel of Fig. 6 illustrates this relation in comparison to the DC-MC relationship described in Appendix A.4, Theorem 22. At small values of ε and α and when the model is perfectly calibrated

overall, α-PMC and ε-DC behave similarly. However, given $\delta > 0$, ε-differentially calibrated models suffer from higher MC error than proportionally calibrated models when α-PMC < 0.3. The right graph also illustrates the feasible range of α for α-PMC is $0 < \alpha < 0.5$, past which it does not provide meaningful α-MC. The steeper relation between α-PMC and MC may have advantages or disadvantages, depending on context. It suggests that, by optimizing for α-PMC, small improvements to this measure can result in relatively large improvements to MC; conversely, ε-DC models that are well calibrated may satisfy a lower value of α-MC over a larger range of ε.

A.3.1. Discretization

To clarify and simplify our analysis, we work mainly with the continuous versions of multicalibration and proportional multicalibration, under the assumption that minimizing the discretized versions (i.e., binning $R(x)$) will translate to low values of the continous version. In this section we provide detailed bounds on the continuous versions of PMC and DC that are implied by the discretized versions.

First, we will formally define two different discretization schemes. The first, λ-discretization, defines equally spaced bins on the interval $[0, 1]$, as follows.

For ensuring multiplicative closeness under PMC, it can be useful to instead discretize the prediction bins so that the bins are equally spaced on a log scale. We define such a discretization below.

Definition 15 ((λ, ρ)-geometric discretization.) *Let $\lambda \in [0, 1], \rho \in [0, 1]$. The (λ, ρ)-*geometric discretization *of $[0, 1]$ is denoted by a set of intervals, $\Lambda_\lambda^\rho = \left\{\{I_j\}_{j=0}^{1/\lambda-1}\right\}$, where $I_j = [\rho^{(1-j\lambda)}, \rho^{(1-j\lambda-\lambda)})$.*

(Hebert-Johnson et al., 2018) define a discretized version of MC in which $R(x)$ is binned according to a discretization parameter, λ:

Definition 16 ((α, λ)-multicalibration) *Let $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be a collection of subsets of $\mathcal{X}$. For any $\alpha, \lambda > 0$, a predictor R is (α, λ)-*multicalibrated *on $\mathcal{C}$ if, for all $I \in \Lambda_\lambda$ and $S \in \mathcal{C}$ where $P_D(R \in I | x \in S) \geq \alpha\lambda$,*

$$\left|\mathop{\mathbb{E}}_D[y | R \in I, x \in S] - \mathop{\mathbb{E}}_D[R | R \in I, x \in S]\right| \leq \alpha.$$

(Hébert-Johnson et al., 2018) establish that (α, λ)-multicalibrated models are at most $(\alpha + \lambda)$-multicalibrated. In an analagous fashion, we show below that (α, λ)-PMC implies $(\alpha + \lambda/\rho)$-PMC for bins defined by a λ-discretization. When using a (λ, ρ)-geometric discretization, (α, λ)-PMC implies $(\alpha\rho^{-\lambda} + \rho^{-\lambda} - 1)$-PMC, which can be a tighter bound than the former.

Claim 1 *Define $\rho, \alpha, \lambda > 0$ and let $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be a collection of subsets of $\mathcal{X}$. Let $\mathbb{E}_D[y | R \in I, x \in S] \geq \rho$ for all $S \in \mathcal{C}$ and $I \in \Lambda_\lambda$. Let $R(x)$ be a model satisfying (α, λ)-proportional multicalibration. Then $R(x)$ is at most $(\alpha + \frac{\lambda}{\rho})$-proportionally multicalibrated.*

Proof

By Definition 18, $R(x)$ satisfies

$$\frac{|\mathbb{E}_D[y | R \in I, x \in S] - \mathbb{E}_D[R | R \in I, x \in S]|}{\mathbb{E}_D[y | R \in I, x \in S]} \leq \alpha$$

for categories $(S, I) \in \mathcal{C} \times \Lambda_\lambda$ satisfying $P_D(R(x) \in I | x \in S) \geq \alpha\lambda$. Given $1/\lambda$ bins, the subset where $P_D(R(x) \in I | x \in S) < \alpha\lambda$ has a size of at most $\alpha|S|$. Therefore there is a subset $|S'| \geq (1 - \alpha)|S|$ where for all $r \in \Lambda_\lambda$, α-PMC (Definition 6) is satisfied.

Let δ be the constaint on δ-PMC. Let $p^* = \mathbb{E}_D[y | R = r, x \in S]$ and $r = \mathbb{E}_D[R | R = r, x \in S]$. Consider the case $r > p^*$ and let $\alpha = (r - p^*)/p^*$. λ-discretization shifts r by at most λ. Let

$$\delta \leq (r + \lambda - p^*)/p^*$$

Substituting $r \leq \alpha p^* + p^*$ yields

$$\delta \leq \alpha + \frac{\lambda}{p^*}$$

Plugging in ρ as the minimum of p^*, we complete the proof. ■

The term $\frac{\lambda}{\rho}$ can be potentially large when $\rho < \lambda$. One way to avoid this issue is to make the change in $R(x)$ between bins scale with $R(x)$ using Definition 15. What makes Definition 15 different from λ-discretization is that the intervals are a multiplicative, rather than additive, distance apart. Hence, for a given $r \in [0, 1]$, a model satisfying (α, λ)-PMC can have its prediction shift by at most a factor of $\rho^{-\lambda}$. This leads us to the following claim.

Claim 2 *Define $\rho, \alpha, \lambda > 0$ and let $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be a collection of subsets of $\mathcal{X}$. Let $\mathbb{E}_D[y | R \in I, x \in S] \geq \rho$ for all $S \in \mathcal{C}$ and $I \in \Lambda_\lambda$. Let $R(x)$ be a model satisfying (α, λ)-proportional multicalibration. Given a (λ, ρ)-geometric discretization, $R(x)$ is at most $(\alpha\rho^{-\lambda} + \rho^{-\lambda} - 1)$-proportionally multicalibrated.*

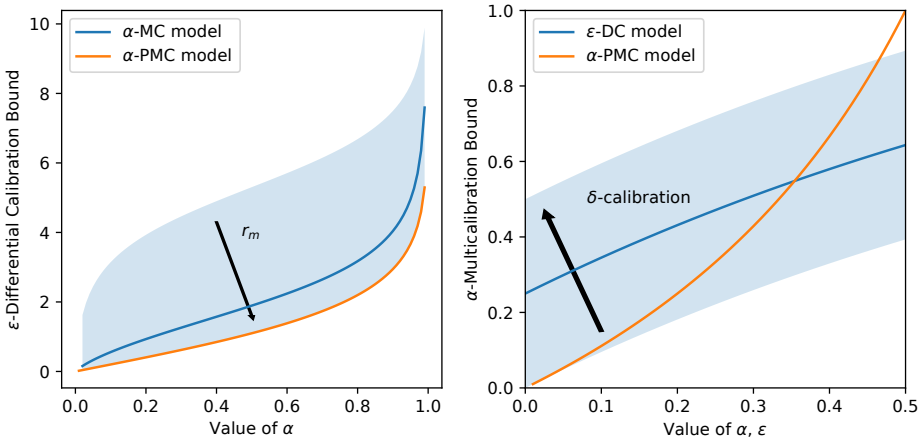

Figure 6: A comparison of ε-DC, α-MC, and α-PMC in terms of their parameters α and ε. In both panes, the x value is a given value of one metric for a model, and the y axis is the implied value of the other metric, according to Theorem 22-Theorem 9. The left filled area denotes the dependence of the privacy/DC of α-multicalibrated models on the minimum risk interval, $r_{min} \in [0.01, 1.0]$. The right filled area denotes the dependence of the MC of ε-differentially calibrated models on their overall calibration, $\delta \in [0.0, 0.5]$. α-PMC does not have these sensitivities.

Proof

By Definition 18, $R(x)$ satisfies

$$\frac{|\mathbb{E}_D[y|R \in I, x \in S] - \mathbb{E}_D[R|R \in I, x \in S]|}{\mathbb{E}_D[y|R \in I, x \in S]} \leq \alpha$$

for categories $(S, I) \in \mathcal{C} \times \Lambda_\lambda$ satisfying $P_D(R(x) \in I | x \in S) \geq \alpha\lambda$. Given $1/\lambda$ bins, the subset where $P_D(R(x) \in I | x \in S) < \alpha\lambda$ has a size of at most $\alpha|S|$. Therefore there is a subset $|S'| \geq (1-\alpha)|S|$ where for all $r \in \Lambda_\lambda$, α-PMC (Definition 6) is satisfied.

Let δ be the constaint on δ-PMC. Let $p^* = \mathbb{E}_D[y|R = r, x \in S]$ and $r = \mathbb{E}_D[R|R = r, x \in S]$. Consider the case $r > p^*$ and let $\alpha = (r - p^*)/p^*$, i.e. the tight bound. (λ, ρ)-geometric discretization shifts r by at most a factor of $\rho^{-\lambda}$. This implies

$$\delta \leq (r\rho^{-\lambda} - p^*)/p^*$$

Substituting $r = \alpha p^* + p^*$ yields

$$\delta \leq \alpha\rho^{-\lambda} + \rho^{-\lambda} - 1.$$

■

We illustrate the relationship between (α, λ)-PMC and α-PMC given a geometric discretization in Fig. 7, which quantifies the relationship for different settings of λ and ρ.

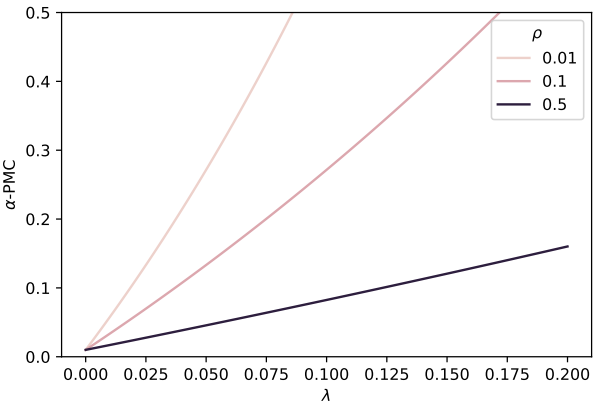

Figure 7: Relationship between (α, λ)-PMC and α-PMC given a geometric discretization. Illustrated for (α, λ)-PMC=0.1, for various values of ρ and λ.

A.3.2. Additional Definitions

Definition 17 (λ-discretization.) *Given $\lambda \in [0,1]$, the* λ-discretization *of $[0,1]$ is denoted by a set of intervals, $\Lambda_\lambda = \left\{\{I_j\}_{j=0}^{1/\lambda-1}\right\}$, where $I_j = [j\lambda, (j+1)\lambda)]$.*

Definition 18 ((α,λ)-PMC) *A model $R(x)$ is* (α,λ)-proportionally multicalibrated *with respect to a collection of subsets $\mathcal{C}$ if, for all $S \in \mathcal{C}$ and $I \in \Lambda_\lambda$ satisfying $P_D(R \in I | x \in S) \geq \alpha\lambda$,*

$$\frac{|\mathbb{E}_D[y|R \in I, x \in S] - \mathbb{E}_D[R|R \in I, x \in S]|}{\mathbb{E}_D[y|R \in I, x \in S]} \leq \alpha. \tag{6}$$

The following loss functions are empirical analogs of the definitions of MC, PMC, and DC, and are used in the experiment section to measure performance.

Definition 19 (MC loss) *Let $\mathcal{D} = \{(y,x)_i\}_{i=0}^N \sim D$, and let $\alpha, \lambda, \gamma > 0$. Define a collection of subsets $\mathcal{C} \in 2^{\mathcal{X}}$ such that for all $S \in \mathcal{C}, |S| \geq \gamma N$. Let $S_I = \{x : R(x) \in I, x \in S\}$ for $(S,I) \in \mathcal{C} \times \Lambda_\lambda$. Define the collection $\mathcal{S}$ containing all S_I satisfying $S_I \geq \alpha\lambda N$. The MC loss of a model $R(x)$ on $\mathcal{D}$ is*

$$\max_{S_I \in \mathcal{S}} \frac{1}{|S_I|} \left| \sum_{i \in S_I} y_i - \sum_{i \in S_I} R_i \right|$$

Definition 20 (PMC loss) *Let $\mathcal{D} = \{(y,x)_i\}_{i=0}^N \sim D$, and let $\alpha, \lambda, \gamma, \rho > 0$. Define a collection of subsets $\mathcal{C} \in 2^{\mathcal{X}}$ such that for all $S \in \mathcal{C}, |S| \geq \gamma N$. Let $S_I = \{x : R(x) \in I, x \in S\}$ for $(S,I) \in \mathcal{C} \times \Lambda_\lambda$. Define the collection $\mathcal{S}$ containing all S_I satisfying $S_I \geq \alpha\lambda N$. Let $\frac{1}{|S_I|}\sum_{i \in S_I} y_i \geq \rho$. The PMC loss of a model $R(x)$ on $\mathcal{D}$ is*

$$\max_{S_I \in \mathcal{S}} \frac{\left|\sum_{i \in S_I} y_i - \sum_{i \in S_I} R_i\right|}{\sum_{i \in S_I} y_i}$$

Definition 21 (DC loss) *Let $\mathcal{D} = \{(y,x)_i\}_{i=0}^N \sim D$, and let $\alpha, \lambda, \gamma > 0$. Define a collection of subsets $\mathcal{C} \in 2^{\mathcal{X}}$ such that for all $S \in \mathcal{C}, |S| \geq \gamma N$. Given a risk model $R(x)$ and prediction intervals I, Let $S_I = \{x : R(x) \in I, x \in S\}$ for $(S,I) \in \mathcal{C} \times \Lambda_\lambda$. Define the collection $\mathcal{S}$ containing all S_I satisfying $S_I \geq \alpha\lambda N$. The DC loss of a model $R(x)$ on $\mathcal{D}$ is*

$$\max_{(S_I^a, S_I^b) \in \mathcal{S} \times \mathcal{S}} \log \left| \frac{1}{|S_I^a|} \sum_{i \in S_I^a} y_i - \frac{1}{|S_I^b|} \sum_{j \in S_I^b} y_j \right|$$

A.4. Additional Theorems

A.4.1. Differentially calibrated models with global calibration are multicalibrated

Here we show that, under the assumption that a model is globally calibrated (satisfies δ-calibration), models satisfying ε-DC are also multicalibrated.

Theorem 22 *Let R(x) be a model satisfying (ε,λ)-DC and δ-calibration. Then $R(x)$ is ($1 - e^{-\varepsilon} + \delta$, λ)-multicalibrated.*

Proof

From Eq. (2) we observe that ε is bounded by the two groups with the largest and smallest group- and prediction- specific probabilities of the outcome. Let I_M be the risk stratum maximizing (ε,λ)-DC, and let $p_n = \max_{S \in \mathcal{C}} P_D(y|R \in I_M, x \in S)$ and $p_d = \min_{S \in \mathcal{C}} P_D(y|R \in I_M, x \in S)$. These groups determine the upper and lower bounds of ε as $e^{-\varepsilon} \leq p_d/p_n$ and $p_n/p_d \leq e^{\varepsilon}$.

We note that $p_d \leq P_D(y|R \in I_M) \leq p_n$, since $P(y|R \in I_M) = \frac{1}{N}\sum_{S \in \mathcal{C}} |S| P_D(y|R \in I_M, x \in S)$, and p_n and p_d are the extreme values of $P(y|R \in I_M, x \in S)$ among S. So, α-MC is bound by the group outcome that most deviates from the predicted value, which is either p_n or p_d. Let $r = P_D(R|R \in I_M)$. There are then two scenarios to consider:

1. $\alpha \leq |p_n - r| = p_n - r$ when $r \leq \frac{1}{2}(p_n + p_d)$; and
2. $\alpha \leq |p_d - r| = r - p_d$ when $r \geq \frac{1}{2}(p_n + p_d)$.

We will look at the first case. Let $p_r^* = P_D(y|R \in I_M)$. Due to δ-calibration, $p_r^* - \delta \leq r \leq p_r^* + \delta$. Then

$$\begin{aligned} \alpha &\leq p_n - r \\ &\leq p_n - (p_r^* - \delta) \\ &\leq p_n - p_d + \delta \\ &= p_n(1 - e^{-\varepsilon}) + \delta \\ \alpha &\leq 1 - e^{-\varepsilon} + \delta. \end{aligned}$$

Above we have used the facts that $r \leq p_r^* - \delta$, $p_r^* \geq p_d$, $p_d \leq e^{-\varepsilon} p_n$, and $p_n \leq 1$. The second scenario is complementary and produces the identical bound. ■

Theorem 22 formally describes how δ-calibration controls the baseline calibration error contribution to α-MC, while ε-DC limits the deviation around this value by constraining the (log) maximum and minimum risk within each category.

A.4.2. Multicalibrated models satisfy intersectional guarantees

In contrast to DF, MC (Hebert-Johnson et al., 2018) was not designed to explicitly incorporate the principles of intersectionality. However, we show that it provides an identical efficiency property to DF in the theorem below. Given an individual's attributes $x = (x_1, \ldots, x_d)$, it will be useful to refer to subsets we wish to protect, e.g. demographic identifiers. To do so, we define $\mathcal{A} = \{A_1, \ldots, A_p\}$, $p \leq d$, such that A_1 is the set of values taken by attribute x_1.

Theorem 23 *Let the collection of subsets $\mathcal{C} \subseteq 2^{\mathcal{X}}$ define groups of individuals according to the Cartesian product of attributes $A \subseteq \mathcal{A}$. Let $\mathcal{G} \in 2^{\mathcal{X}}$ be any collection of subsets that groups individuals by the Cartesian product of attributes in A', where $A' \subset A$ and $A' \neq \emptyset$. If $R(x)$ satisfies α-MC on $\mathcal{C}$, then $R(x)$ is α-multicalibrated on $\mathcal{G}$.*

In proving Theorem 23, we will make use of the following lemma.

Lemma 24 *The α-MC criteria can be rewritten as: for a collection of subsets $\mathcal{C} \subseteq \mathcal{X}$, $\alpha \in [0, 1]$, and $r \in [0, 1]$,*

$$\max_{c \in \mathcal{C}} \mathop{\mathbb{E}}_{D}[y|R(x) = r, x \in c] \leq r + \alpha$$

and

$$\min_{c \in \mathcal{C}} \mathop{\mathbb{E}}_{D}[y|R(x) = r, x \in c] \geq r - \alpha$$

Proof The lemma follows from Definition 2, and simply restates it as a constraint on the maximum and minimum expected risk among groups at each prediction level. ■

Proof [Proof of Theorem 23] We use the same argument as Foulds et al. (2019a) in proving this property for DF. Define Q as the Cartesian product of the protected attributes included in $\mathcal{A}$, but not $\mathcal{A}'$. Then for any $(y, x) \sim D$,

$$\max_{g \in \mathcal{G}} \mathop{\mathbb{E}}_{D}[y|R(x) = r, x \in g]$$

$$= \max_{g \in \mathcal{G}} \sum_{q \in Q} \mathop{\mathbb{E}}_{D}[y|R(x) = r, x \in g \cap q] P[x \in q | x \in g] \tag{7}$$

$$\leq \max_{g \in \mathcal{G}} \sum_{q \in Q} \max_{q' \in Q} \mathop{\mathbb{E}}_{D}[y|R(x) = r, x \in g \cap q'] P[x \in q | x \in g] \tag{8}$$

$$= \max_{g \in \mathcal{G}} \max_{q' \in Q} \mathop{\mathbb{E}}_{D}[y|R(x) = r, x \in g \cap q'] \tag{9}$$

$$= \max_{c \in \mathcal{C}} \mathop{\mathbb{E}}_{D}[y|R(x) = r, x \in c]. \tag{10}$$

Moving from Eq. (7) to Eq. (8) follows from substituting the maximum value of $\mathbb{E}_D[y|R(x) = r, x]$ for observations in the intersection of subsets in $\mathcal{G}$ and Q which is the upper limit of the expression in Eq. (7). Moving from Eq. (8) to Eq. (9) follows from recognizing that the sum $P[x \in q | x \in g]$ for all subsets in $\mathcal{Q}$ is 1. Finally, moving from Eq. (9) to Eq. (10) follows from recognizing that the intersections of subsets in $\mathcal{G}$ and $\mathcal{Q}$ that satisfy Eq. (9), must define a subset of $\mathcal{C}$. Applying the same argument, we can show that

$$\min_{g \in \mathcal{G}} \mathop{\mathbb{E}}_{D}[y|R(x) = r, x \in g] \geq \min_{c \in \mathcal{C}} \mathop{\mathbb{E}}_{D}[y|R(x) = r, x \in c].$$

Substituting into Theorem 24,

$$\max_{g \in \mathcal{G}} \mathop{\mathbb{E}}_{D}[y|R(x) = r, x \in g] \leq \alpha + r$$

and

$$\min_{g \in \mathcal{G}} \mathop{\mathbb{E}}_{D}[y|R(x) = r, x \in g] \geq r - \alpha$$

or

$$\left| \mathop{\mathbb{E}}_{D}[y|R(x) = r, x \in g] - r \right| \leq \alpha$$

for all $g \in \mathcal{G}$. Therefore $R(x)$ is α-multicalibrated with respect to $\mathcal{G}$. ■

As a concrete example, imagine we have the protected attributes $A = \{\text{race} \in \{B, W\}, \text{gender} \in \{M, F\}\}$. According to Theorem 23, $\mathcal{C}$ would contain four sets: $\{(B, M), (B, F), (W, M), (W, F)\}$. In contrast, there are eight possible sets in $\mathcal{G}$: $\{(B, M), (B, F), (W, M), (W, F), (B, *), (W, *), (*, M), (*, F)\}$, where the wildcard indicates a match to

either attribute. As noted in Appendix A.1.1, the efficiency property is useful because the number of possible sets in $\mathcal{G}$ grows at a large combinatorial rate, rate as additional attributes are added; meanwhile $\mathcal{C}$ grows at a slower, yet exponential, rate. For an intuition for why this property holds, consider that the maximum calibration error of two subgroups is at least as large as the maximum expected error of those groups combined; e.g., the maximum calibration error in a higher order groups such as $(B, *)$ will be covered by the maximum calibration error in either (B, M) or (B, F).

A.5. Additional Experiment Details

Models The deep neural network (DNN) was a five layer feed-forward NN with 100 units per layer and ReLU activations. We trained using an adam solver with a learning rate of 0.001, a batch size of 200 and used early stopping to terminate when a 10% validation set did not improve for 10 epochs.

Both LR and DNN models used median imputation and feature normalization as pre-processing steps. For the RF, we used the XGBoost implementation Chen and Guestrin (2016) which handles missing data natively.

Training Models were trained on a heterogenous computing cluster. Each training instance was limited to a single core and 4 GB of RAM. We conducted a full parameter sweep of the parameters specified in Table 3. A single trial consisted of a method, a parameter setting from Table 3, and a random seed. Over 100 random seeds, the data was shuffled and split 75%/25% into train/test sets. Results in the manuscript are summarized over these test sets.

Code Code for the experiments is available here: https://github.com/cavalab/proportional-multicalibration. Code is licensed under GNU Public License v3.0.

Data We make use of data from the MIMIC-IV-ED repository, version 1.0, to train admission risk prediction models (Johnson et al., 2021). This resource contains more than 440,000 ED admissions from Beth Isreal Deaconness Medical Center between 2011 and 2019. We preprocessed these data to construct an admission prediction task in which our model delivers a risk of admission estimate for each ED visitor after their first visit to triage, during which vitals are taken. Additional historical data for the patient was also included (e.g., number of previous visits and admissions). A list of features is given in Table 2.

A.6. Additional Experimental Results

Table 3 lists a few parameters that may affect the performance of post-processing for both MC and PMC. Of particular interest when comparing MC versus PMC post-processing is the parameter α, which controls how stringent the calibration error must be across categories to terminate, and the group definition (A), which selects which features of the data will be used to asses and optimize fairness. We look at the performance of MC and PMC postprocessing over values of α and group definitions in Figs. 8 to 10. Finally, we empirically compare MC- and PMC-postprocessing by the number of steps required for each to reach their best performance in Fig. 11 and Table 5.

From Fig. 8, it is clear that post-processing has a minimal effect on AUROC in all cases; note the differences disappear if we round to two decimal places. When post processing with RF, we do note a relationship between lower values of α and a very slight decrease in performance, particularly for MC-postprocessing.

Figs. 9 and 10 show performance between methods on MC loss and PMC loss, respectively. In terms of MC loss, PMC-postprocessing tends to produce models with the lowest loss, at α values greater than 0.01. Lower values of α do not help MC-postprocessing in most cases, suggesting that these smaller updates may be overfitting to the post-processing data. In terms of PMC loss (Fig. 10), we observe that performance by MC-postprocessing is highly sensitive to the value of α. For smaller values of α, MC-postprocessing is able to achieve decent performance by these metrics, although in all cases, PMC-postprocessing generates a model with a better median loss value at some configuration of α.

We assess how many steps/updates MC and PMC take for different values of α in Fig. 11, and summarize empirical measures of running time in Table 5. On the figure, we annotate the point for which each post-processing algorithm achieves the lowest median value of PMC loss across trials. Fig. 11 validates that PMC-postprocessing is more efficient than MC-postprocessing at producing models with low PMC loss, on average requiring 4.0x fewer updates to achieve its lowest loss on test. From Table 5 we observe that PMC typically requires a larger

Table 6: The number of times each postprocessing method achieved the best score among all methods, out of 100 trials.

postprocessing metric	Base Model	MC	PMC
AUROC	5	88	6
MC loss	8	21	70
PMC loss	0	27	72
DC loss	0	36	63

number of updates to achieve its best performance on MC loss (about 2x wall clock time and number of updates), whereas MC-postprocessing requires a larger number of updates to achieves its best performance on PMC loss and DC loss, due to its dependence on very small values of α. We accompany these results with the caveat that they are based on performance on one real-world task, and wall clock time measurements are influenced by the heterogenous cluster environment; future work could focus on a larger empirical comparison.

We quantify how often the use of each postprocessing algorithm gives the best loss for each metric and trial in Table 6. PMCBoost (Algorithm 1) achieves the best fairness the highest percent of the time, according to DC loss (63%), MC loss (70%), and PMC loss (72%), while MC-postprocessed models achieve the best AUROC in 88% of cases. This provides strong evidence that, over a large range of α values, PMCBoost is beneficial compared to MCBoost.

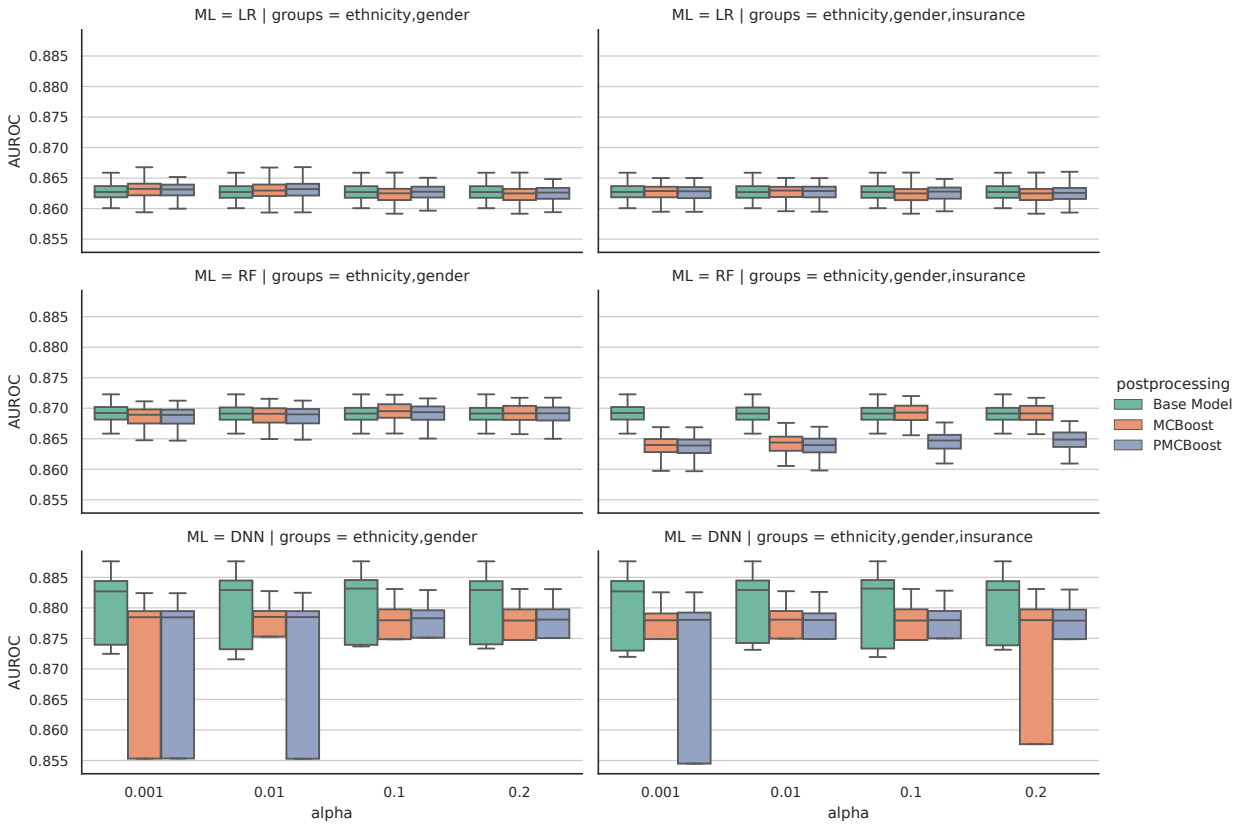

Figure 8: AUROC test performance versus α across experiment settings. Rows are different ML base models, and columns are different attributes used to define $\mathcal{C}$. The color denotes the post-processing method.

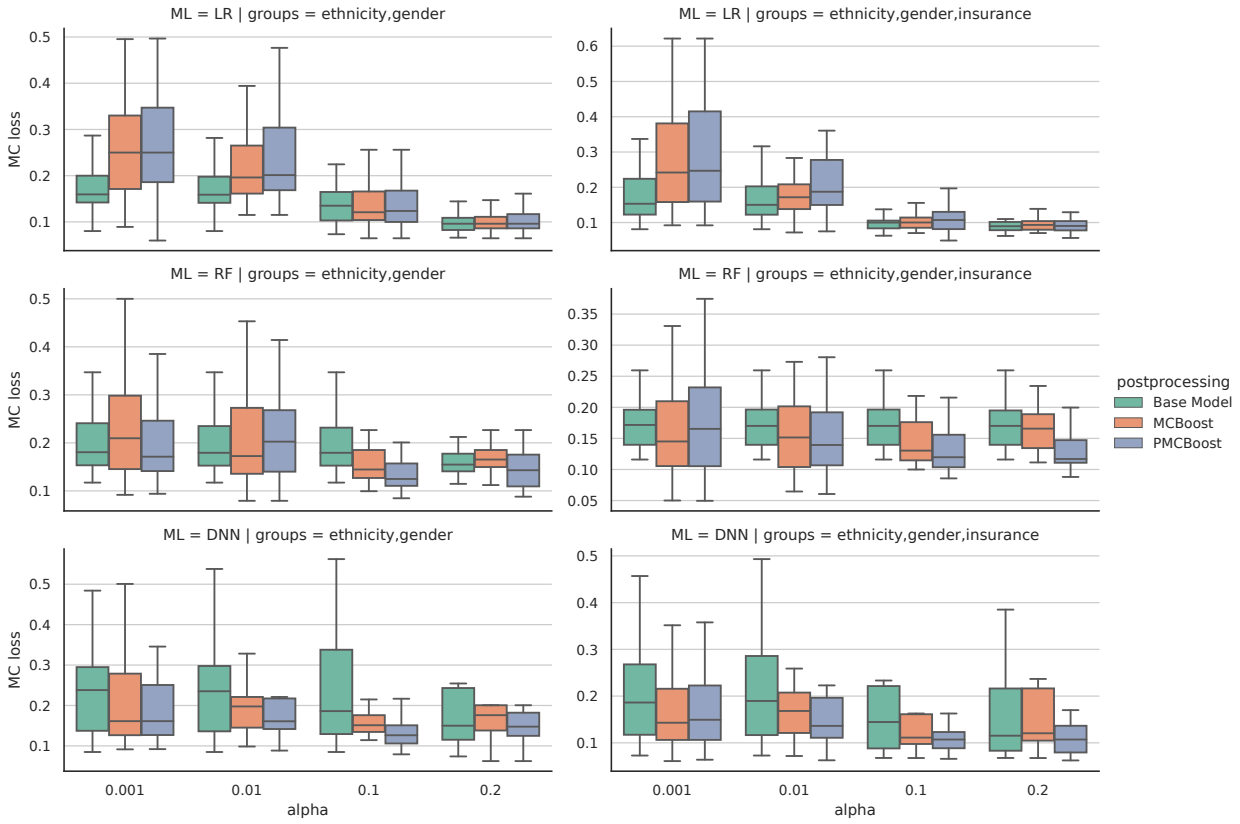

Figure 9: MC loss test performance versus α across experiment settings. Rows are different ML base models, and columns are different attributes used to define $\mathcal{C}$. The color denotes the post-processing method.

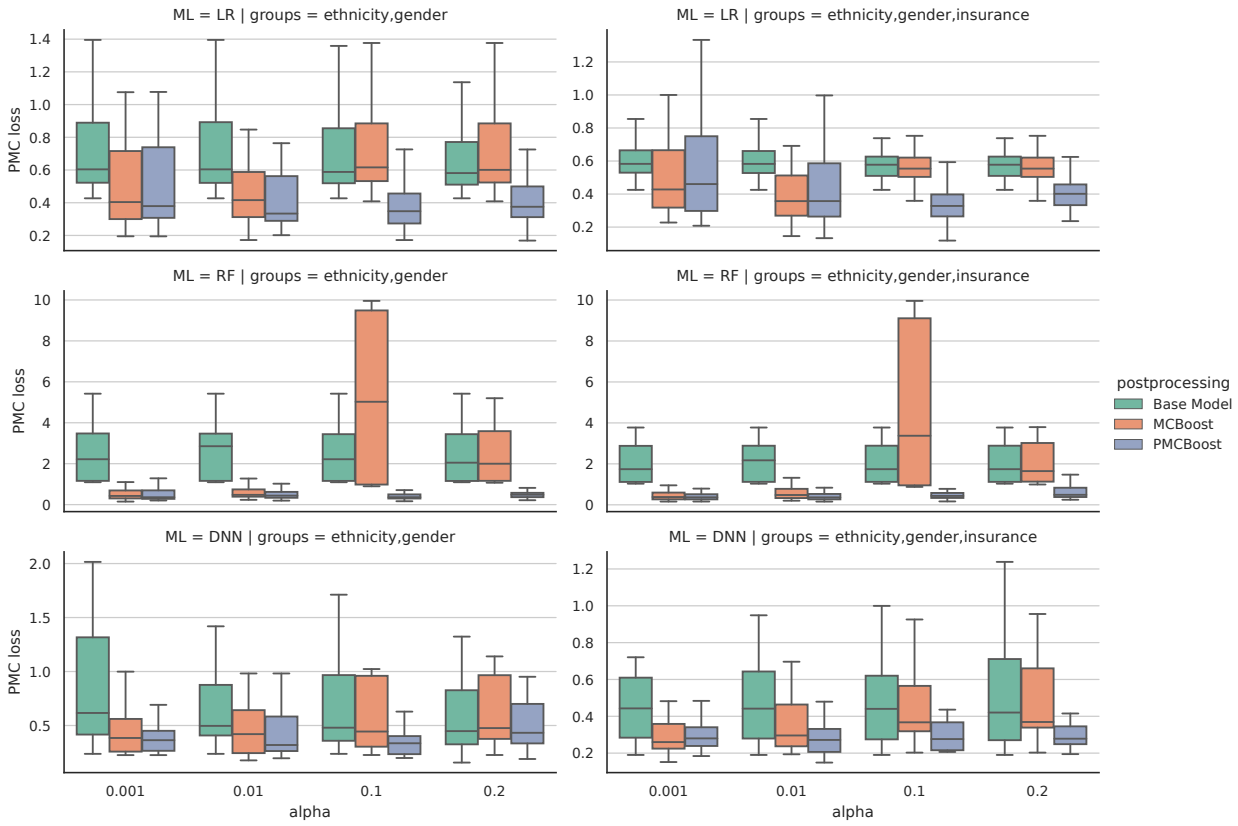

Figure 10: PMC loss test performance versus α across experiment settings. Rows are different ML base models, and columns are different attributes used to define $\mathcal{C}$. The color denotes the post-processing method.

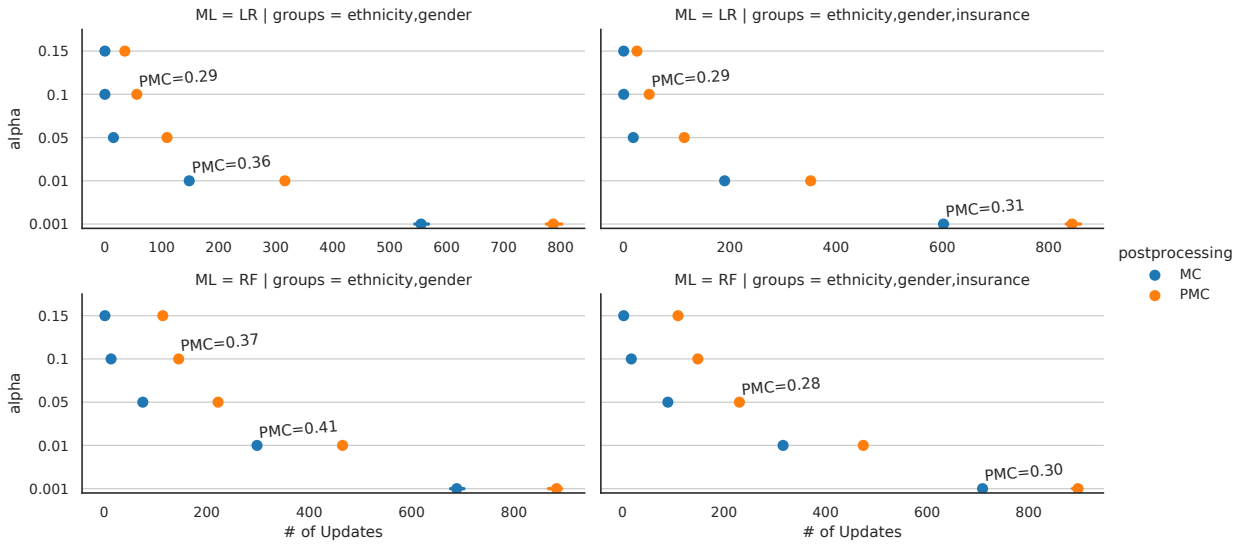

Figure 11: Number of post-processing updates by MC and PMC versus α across experiment settings. Rows are different ML base models, and columns are different attributes used to define $\mathcal{C}$. The color denotes the post-processing method. Each result is annotated with the median PMC loss for that method and parameter combination.