The Roads Not Taken: Fairness and Predictive Multiplicity in Target Specification under Resource Constraints

JAMELLE WATSON-DANIELS, Microsoft Research, USA
SOLON BAROCAS, Microsoft Research, USA
JAKE HOFMAN, Microsoft Research, USA
ALEXANDRA CHOULDECHOVA, Microsoft Research, USA

When translating high-level goals into tractable predictive tasks, there are often many reasonable target variable options. Past work has argued that this flexibility in target choice can help address fairness concerns. Specifically, some targets may lead to less selection rate disparities across groups than others. Similarly, work on predictive multiplicity has explored minimizing selection rate disparities over a set of equally-good competing models. In this work we provide a novel mathematical and computational framework for exploring these intersecting phenomena: target choice, predictive multiplicity and fairness under resource constraints. To do so, we formalize measures of predictive multiplicity for resource-constrained settings, where there is a pre-specified budget for the number of cases that can receive a positive classification. We then provide mixed integer programs for computing notions of ambiguity both within and across different target variables, and apply the methodology to a large-scale healthcare dataset. Through these examples, we show that combining target variable choice and predictive multiplicity can give model developers varying degrees of latitude that they can exploit to address fairness concerns.

Additional Key Words and Phrases: target specification, fairness, multiplicity, predictive inconsistency

ACM Reference Format:
Jamelle Watson-Daniels, Solon Barocas, Jake Hofman, and Alexandra Chouldechova. 2023. The Roads Not Taken: Fairness and Predictive Multiplicity in Target Specification under Resource Constraints. In *FAccT 2024: ACM Conference on Fairness, Accountability and Transparency, June xx–yy, 2023, Chicago, IL.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Real-world problems rarely present themselves as machine learning tasks [36]. Instead, practitioners often need to engage in a good deal of work to figure out how predicting some outcome or quality of interest can actually help to solve these problems [17, 33]. It is far from obvious, for example, how employers should go about using machine learning methods in their hiring practices: if the goal is to hire the "best" people, what, exactly, should the model be predicting [2, 23, 33]? For a sales position, employers might choose to predict annual sales figures. But they could just as well choose to predict how well the applicant will work with others, whether customers will actually enjoy interacting with the applicant, etc. Likewise, the choice of prediction target is often unclear in domains such as criminal justice and human services. Many algorithmic tools in deployment today function by aggregating predictions of several different targets, ranging from different types of criminal justice system encounters, to mental and physical health outcomes, to measures of housing stability [24, 40]. Even in domains where it might seem more obvious what is the right target of prediction, there can still be a good deal of uncertainty. For example, while it might seem self-evident that creditors should be predicting default, what constitutes "default" is not a given. Creditors need to make an affirmative choice about the number of months of missed payments that ultimately count as "default" [18].

2023. Manuscript submitted to ACM

Much can go wrong in making these decisions. As Hand [17] has observed, statisticians sometimes develop models that end up answering questions that are subtly different than those of actual concern. This can happen when practitioners alter the problem to make it statistically tractable or opt for a proxy when the desired target is not directly observable. In either case, this can lead to a tenuous and therefore unreliable relationship to the outcome or quality of actual interest [17, 19]. These are pernicious problems because they can be easily overlooked: decision makers might not recognize the lack of alignment between their goals and the chosen target or the gap between the chosen proxy and the actual outcome or quality of interest. But this latitude can also be a virtue. Because there is rarely one correct target or proxy, practitioners will often have some flexibility in how they go about formulating the problem. In many cases, decision makers' goals can be served equally well by developing models that predict a range of possible targets or proxies [23]. They can take many different roads.

A recent line of work has explored the implications of this flexibility for fairness. In particular, researchers have pointed out that different choices for the target of prediction can lead to more or less disparity in selection rates across groups [13, 21, 23, 27, 30–33]. One particularly well-known study by Obermeyer et al. [32] illustrates both the risks and benefits of target choice. The authors examined an algorithm developed by a healthcare system to predict which of its patients would benefit the most from a "high-risk care management" program. They found that the healthcare system's choice to adopt healthcare costs as the target of prediction led to notable racial disparities, with the benefits of enrollment for black patients being systematically underestimated in comparison to other patients, largely due to the fact that black patients of equivalently poor health do not receive equivalent treatment. The authors then showed that a good deal of the racial disparity could have been avoided had the healthcare system instead chosen to predict a more direct measure of health outcomes. The study has been received as a important lesson in the dangers of insufficiently careful target choice, with the racial disparity attributable to a gap between the chosen proxy and the actual outcome of interest. But it also highlights that practitioners can take advantage of the latitude afforded by target choice to reduce selection rate disparities.

In this paper, we seek to extend this prior work in three ways. First, we consider the possibility that practitioners might have multiple outcomes or qualities of interest, not just one, and may want some way to compare the fairness implications of choosing one of them rather than the others, or of combining them in some way. For example, it is entirely possible that the healthcare system studied by Obermeyer et al. [32] cared about both improving health outcomes and reducing healthcare costs—and could have had reasons for choosing either or both of them.

Second, we also explore the additional flexibility afforded by so-called "predictive multiplicity" [28], the fact that there often exist many models of near-equivalent performance that can nevertheless differ dramatically in their predictions on specific instances, sometimes called the "Rashomon Set" [6]. This phenomenon has been of particular interest to those working on issues of fairness because it suggests that there will often exist a model with comparable performance but with less disparity in selection rates across groups than some baseline model–and that there may be a way to avoid a trade-off between performance and fairness [1, 5, 9, 28, 42]. In other words, just like there is rarely one right choice of target, there is rarely one best model for any given task at practically achievable levels of performance [5, 37]. Thus, even if practitioners feel compelled to select the model that maximizes performance, they retain a good deal of freedom to choose from within the (potentially large) set of models that all exhibit a comparable level of performance. And they can make such choices with fairness concerns in mind explicitly.

Third, we grapple with the fact that practitioners often face constraints that limit the degree to which they can allocate a resource to the total number of people that any particular classifier assigns to the positive class. The algorithm investigated by Obermeyer et al. [32] was specifically developed to allocate a fixed amount of potential support, even though the healthcare system would have liked to make it available to a much larger number of people. Similarly, employers cannot offer jobs to *everyone* they predict will be a good employee, whatever target or set of targets they choose to predict to make such an assessment. Given their limited budgets, they are likely only able to offer jobs to a select few applicants. When it comes to the real-world implications of target choice and multiplicity, we argue that we should focus on the composition of the population that actually receives the desired resource.

In this paper, we bring all three of these threads together to investigate the relationship between target choice, multiplicity, and fairness under resource constraints. In particular, we study settings in which (1) there are many reasonable candidate targets and it might make sense to choose more than one, (2) there may be some degree of multiplicity in the models that we can train for the chosen target or combination of targets, and (3) there are real-world constraints on the total number of people to whom we can allocate some resource. To do this, we develop novel methods to answer the following questions:

(1) Given resource constraints, how do different target choices compare in terms of the *predictive multiplicity* and *selection rate disparities* exhibited by the resulting models?
(2) Given resource constraints, how much predictive multiplicity exists due to different ways of *combining targets*; and how does the combining approach affect *selection rate disparities* exhibited by the resulting models?

The paper proceeds as follows. We begin in §2 with an overview of related work, drawing connections to prior work on problem formulation and algorithmic fairness, multi-task learning, predictive multiplicity, and fairness in the presence of resource constraints. In §3 we formalize the resource-constrained predictive allocation problem in the single-target setting. While our ultimate focus is on multi-target multiplicity, the single-target setting allows us to introduce novel concepts and computational approaches in a more familiar context, and is also of interest in its own right. We introduce a measure of predictive multiplicity, top-κ ambiguity (§3.2), and then describe a mixed integer program (MIP) that allows us to calculate this ambiguity measure for linear models (§3.3). Unlike in prior work on predictive multiplicity, where calculating ambiguity typically involves solving a computationally expensive MIP for every data point, we establish bounds that enable us to run our MIP on a relatively small fraction of points (§3.4). Taken together, these contributions provide answers to question (1). To answer question (2), in §4 we introduce a framework for thinking about multiplicity for resource constrained predictive allocation problems with multiple target options. We formalize the notion of a *combining rule*, introduce a measure of ambiguity over combining rules (§4.1), and provide a MIP formulation for computing ambiguity for the family of *index model* combining rules (§4.2). To address the question of selection rate disparities, we present another MIP formulation that allows us to calculate best and worst− case selection rate disparities for the index model class of combining rules (§4.3). These methods enable us to determine (i) the degree to which selecting a particular target to the exclusion of others may result in disparate impact, and (ii) whether combining target choices can reduce disparities.

Our evaluation on the healthcare dataset released by Obermeyer et al. shows that both flexibility in target choice and predictive multiplicity within a given target can be effective levers for addressing fairness concerns. Furthermore, by comparing single and multi-target multiplicity measures we are able to show that the degree of multiplicity attributable to the presence of multiple targets can far exceed that arising from variation in near-optimal models of a single target variable. This provides further empirical evidence of the importance of target variable choice.

2 RELATED WORK

Problem formulation and fairness. Scholars have identified various reasons why the choice of target or proxy might raise concerns with fairness: some outcomes or qualities of interests might just be more evenly distributed across the population than others [23, 33]; certain outcomes or qualities of interests might be easier to predict with similar degrees of accuracy across the population than others [8]; some kind of selection bias might cause certain outcomes or qualities of interest to be observed more or less frequently in certain groups rather than others, even if they occur at similar rates in reality [26]; certain targets might suffer from more so-called "label bias" than others—that is, systematically less accurate observations of the true value of the target for members of some groups than others [9, 21, 22]. Indeed, one way to understand the Obermeyer et al. [32] study is as a form of label bias since healthcare costs acted as a systematically inaccurate measure of actual healthcare need. Our work departs from much of this literature by focusing on cases where there is no obviously right or clearly preferable choice of target or proxy and thus uncertainty about which to choose or whether to choose more than one.

Multi-task learning, multi-criteria decision-making, latent variable modeling, and fairness. While our use of the term "multi-target" might suggest a close connection to fairness considerations in multi-task learning (see, e.g., [41]), the problem we study is distinct. Whereas in multi-task learning the goal is to perform well on (and assess fairness for) *each* of K prediction tasks by borrowing strength across tasks, in our setting we are interested in arriving at a *single model*, which may not perform optimally on any individual task, but which successfully captures multiple desiderata. In this sense, our setting is more closely related to recent work on latent variable modeling in recommender systems that aims to optimize for a latent *value* using a combination of noisy observed measures, such as clicks, replies, reshares, and other observable forms of user engagement [25, 29]. A key difference is that we do not posit a specific notion of optimality, and instead explore the degree of multiplicity inherent in a class of learning procedures for forming a univariate prediction from multiple available targets. Lastly, our work connects to the extensive literature on multi-criteria decision-making (MCDM) in operations research. Indeed, the index model and index variable approaches we introduce in §4.1 parallel the classic *weighted sums* method of combining multiple criteria (e.g., loss or other objective functions) into a single objective [15]. However, whereas the focus of MCDM is in the values of the different objective functions, we examine multiplicity, which pertains to the variability in prediction decisions for *individual people or cases.*

Predictive multiplicity and fairness. There is also a growing literature that seeks to explore the normative implications of multiplicity. Scholars have investigated the degree to which multiplicity can be leveraged to improve interpretability [39] and explainability [11, 14, 34]. Others have examined the danger multiplicity poses for robustness [12] and non-arbitrariness [4, 5, 7, 20, 38]. Still others have focused on its implications for fairness [1, 5, 9, 28, 42]. Notably, some of this work has defined measures and developed methods for evaluating predictive multiplicity in binary classification [28] and probabilistic classification [20, 42], focusing on so called "ambiguity" in models' predictions (i.e., the amount of disagreement in models' predictions on different points). Our work is the first to extend the analysis of multiplicity to the problem of predictive allocation under resource constraints. We develop novel measures of ambiguity for both single-target and multi-target settings, and introduce efficient methods that, for a subset of points, can certifying whether those points do or do not contribute to the ambiguity measure.

Resource constraints and fairness. Recent work on algorithmic fairness has noted the importance of considering resource constraints. For instance, Black et al. [3] discuss how the increased cost of auditing more complex tax filings can lead to prediction-based auditing strategies that disproportionately focus on lower income earners. Other work has emphasized the importance of considering resource constraints in the context of algorithmic fairness in healthcare [35] and business analytics [10]. Our work provides a conceptual and computational framework for reasoning about fairness in the presence of resource constraints.

3 PREDICTIVE MULTIPLICITY WITH RESOURCE CONSTRAINTS

Before introducing the multi-target multiplicity framework we first examine predictive multiplicity in the presence of resource constraints. This setting is of interest both in its own right, and also as a stepping stone towards the computational techniques we will rely on in the multi-target setting. We introduce here a resource-constrained ambiguity measure along with a computational framework based on mixed-integer programming for computing ambiguity for linear models.

3.1 Preliminaries

We consider a dataset, $\mathcal{D} = \{(x_i, a_i, \tilde{y}_i^{(k)})\}_{i=1}^n$, consisting of n cases where $x_i = [1, x_{i1}, \dots, x_{id}] \in \mathcal{X} \subseteq \mathbb{R}^{d+1}$ is a feature vector, $y_i \in \mathbb{R}$ is an outcome of interest (potentially binary), and $a_i \in A$ is a protected attribute. We operate within the prediction-based allocation setting where a limited resource is to be allocated to instances in descending order of the predicted value $\hat{y}_i = \hat{y}(x_i)$. If case i is selected, it is allocated r_i resources. Let κ denote the resource cap, and let $i_{(j)} = i_{(j)}(\hat{y})$ denote the instance with the jth largest value of $\hat{y}_i$ (so that $i_{(1)}$ is the index with the largest predicted value). Let $\tau_i = \tau_i(\hat{y})$ denote the rank of instance i in *descending* order. By definition, $\tau_{i_{(j)}} = j$. We assume that resources get allocated to instances $i_{(1)}, \dots, i_{(J)}$, where J is the largest value such that $\sum_{j=1}^J r_{i_{(j)}} \leq \kappa$. The most common prediction-based allocation setting in practice is where there is simply a limit to the number of cases that can be selected; i.e., $r_i = 1 \; \forall i$, in which case $J = \kappa$. While we restrict our attention to this setting, all metrics and computational methods can be generalized to general $r_i \in \mathbb{R}_{>0}$.

3.2 Measuring predictive multiplicity under resource constraints

In previous work, Marx et al. [28] (binary predictions) and Watson-Daniels et al. [42] (probabilistic classification) provide analogous definitions. As in the standard predictive multiplicity setting [28], we begin with a *baseline model* h_0 that is the solution to an empirical risk minimization (ERM) problem of the form $\min_{h \in \mathcal{H}} L(h; \mathcal{D})$, over a hypothesis class, $\mathcal{H}$, with loss $L(\,\cdot\,; \mathcal{D})$.

Definition 3.1 (ϵ-Rashomon set). For a baseline model h_0 and error tolerance $\epsilon > 0$, the ϵ-Rashomon set of competing models is:

$$\mathcal{R}_\epsilon(h_0) := \{h \in \mathcal{H} : L(h) \leq L(h_0) + \epsilon\}.$$

In [28], $\mathcal{H}$ is assumed to be a class of binary classifiers, and the ambiguity of a prediction problem is defined as,

$$\alpha_\epsilon(h_0) = \frac{1}{n} \sum_{i=1}^n \max_{h \in \mathcal{R}_\epsilon(h_0)} \mathbb{1}[h(x_i) \neq h_0(x_i)].$$

Note that under this definition, a prediction problem will have high ambiguity if the positive classification rate, $\frac{1}{n}|\{i : h(x_i) = 1\}|$, differs greatly between h_0 and models in $\mathcal{R}_\epsilon(h_0)$. That is, a high ambiguity may simply result from models that allocate a very different number of resources.

To define an analogous measure for the resource constrained setting, we need to compare models at the same resource cap κ. Recall that, unlike in [28], we consider $\mathcal{H}$ that is a class of prediction models returning continuous values in $\mathbb{R}$, not binary classifiers. Given a prediction model h and resource cap κ, let

$$Top_{(i,h,\kappa)} = \mathbb{1}[\tau_i(h) \leq \kappa] \tag{1}$$

be the indicator of whether instance i is "in the top-κ" when ranked according to the predicted values h. We define two notions of ambiguity in this setting.

Definition 3.2 (Top-κ ambiguity (all)). The (ϵ, κ)-*ambiguity* over a sample, S, is the proportion of examples for which the top-κ decision changes over the ϵ-Rashomon set:

$$A_{\epsilon,\tau}(h; S) := \frac{1}{n} \sum_{i=1}^{n} \max_{h \in \mathcal{R}_\epsilon(h_0)} \mathbb{1}[Top_{(i,h,\kappa)} \neq Top_{(i,h_0,\kappa)}]. \tag{2}$$

Definition 3.3 (Top-κ Ambiguity (top)). The (ϵ, κ)-*ambiguity* over a sample, S, is the proportion of top-κ examples according to h_0 that fall outside the top-κ for some models in the ϵ-Rashomon set:

$$A_{\epsilon,\tau}(h; S) := \frac{1}{\kappa} \sum_{i=1}^{n} \max_{h \in \mathcal{R}_\epsilon(h_0)} Top_{(i,h_0,\kappa)} \left(1 - Top_{(i,h,\kappa)}\right) \tag{3}$$

Before proceeding to the discussion of computational strategies for calculating the ambiguity measures it will help to have one more definition.

Definition 3.4 (Flippable point). An instance i is *flippable* in $\mathcal{R}_\epsilon(h_0)$ if either

$$Top_{(i,h_0,\kappa)} = 1 \text{ and } \max_{h \in \mathcal{R}_\epsilon(h_0)} \tau_i(h) > \kappa \quad ;\text{or}$$

$$Top_{(i,h_0,\kappa)} = 0 \text{ and } \min_{h \in \mathcal{R}_\epsilon(h_0)} \tau_i(h) \leq \kappa.$$

Note that the top-κ ambiguity (all) is simply the fraction of instances that are flippable. Top-κ ambiguity (top) is the fraction of instance in the top-κ of the baseline model h_0 that are flippable out of the top-κ by some $h \in \mathcal{R}_\epsilon(h_0)$.

3.3 Computing top-κ ambiguity for linear models

In this section we describe an approach to calculating ranked ambiguity for linear models $\mathcal{H} = \{h(x) = x^T w : w \in \mathbb{R}^{d+1}\}$ and squared error loss, $L(h; \mathcal{D}) = L(w; \mathcal{D}) = RSS(w; \mathcal{D}) = \sum_{i=1}^{n} (y_i - x_i^T w)^2$. We will use h and w notation interchangeably in the context of linear models. Unless stated otherwise, we will assume throughout this section that the design matrix X has been transformed to be orthonormal. The problem is invariant to this operation, but working with an orthonormal X helps simplify expressions and reduces notational burden.

We calculate ambiguity by formulating a mixed integer program (MIP) that identifies flippable points. Though, as we discuss in §3.4, we can determine the flippability of many points without needing to run the computationally expensive MIP. For each instance i such that $Top_{(i,h0,\kappa)} = 1$, we calculate the *maximum* attainable rank, $\max_{h \in \mathcal{R}_\epsilon(h_0)} \tau_i(h)$. For instances i outside of the top-κ, we calculate the *minimum* attainable rank, $\min_{h \in \mathcal{R}_\epsilon(h_0)} \tau_i(h)$. The following proposition helps to neatly characterize the ϵ-Rashomon set, $\mathcal{R}_\epsilon(h_0)$, for linear models, the proof of which is presented in the Appendix A.

Proposition 3.5. *Assume the design matrix $X_{n \times (d+1)}$ has been orthonormalized, and $w_0 = \text{argmin}_{w \in R^{d+1}} \|y - Xw\|_2^2$ is the least squares solution. Then*

$$\mathcal{R}_\epsilon(w_0) = \{w \in \mathbb{R}^{d+1} : RSS(w) \leq RSS(w_0) + \epsilon\} = \{w \in \mathbb{R}^{d+1} : \|w - w_0\| \leq \epsilon\}.$$

To calculate ambiguity, for each point $i \in \mathcal{D}$ we solve the following MIP formulation, which we call `FlipTopKMIP`$(h_0, x_i; \kappa, \epsilon)$.

$$\begin{aligned} \min \text{ or } \max_{I \in \mathbb{R}^{2 \times (n-1)}, w \in \mathbb{R}^{d+1}} \quad & \sum_{i' \neq i} I_{(i' > i)} & & \\ \text{s.t.} \quad & I_{(i'>i)} + I_{(i>i')} = 1 & & i' = 1, \ldots, n \setminus i \end{aligned} \tag{4a}$$

$$\begin{aligned}
(x_{i'} - x_i)^T w \;&\leq\; M * I_{(i'>i)} && i' = 1, ..., n \setminus i && \text{(4b)}\\
(x_i - x_{i'})^T w \;&\leq\; M * I_{(i'>i)} && i' = 1, ..., n \setminus i && \text{(4c)}\\
\|w - w_0\|_2^2 \leq \epsilon & && && \text{(4d)}\\
w_j \;&\in\; \mathbb{R} && j = 1, ..., d+1 &&\\
I_{(i'>i)}, I_{(i>i')} \;&\in\; \{0,1\} && i' = 1, ..., n \setminus i &&
\end{aligned}$$

where we set (see Appendix A.2 for details),

$$M = \left(\sqrt{\|w_0\|_2^2 + \epsilon}\right) \max_{i,j} \|x_j - x_i\|_2.$$

The binary variables $I_{(i'>i)}$ serve as indicators that $\hat{y}_{i'} = x_{i'}^T w \geq x_i^T w = \hat{y}_i$, which means that the objective $\sum_{i' \neq i} I_{(i'>i)} = \tau_i(w) - 1$ is simply the rank (minus 1) of point i in model w. Constraint (4d) enforces that w is in the ϵ-Rashomon set, as per Proposition (3.5).

3.4 Improving efficiency by identifying provably (un)flippable points

Whereas prior related work on predictive multiplicity in binary [28] and probabilistic [42] classification has involved solving a MIP for every point in $\mathcal{D}$, we show this is not necessary in our setting. Specifically, we show that (i) one can efficiently determine that many points are provably *not flippable* over the ϵ-Rashomon set; and (ii) one can identify a subset of *flippable* points by solving a proxy optimization problem with a closed-form solution that produces a $w \in \mathcal{R}_\epsilon(w_0)$ that may flip some points into the top-κ. This means that in practice we only need to solve the computationally expensive `FlipTopKMIP` for a very small subset of points whose flippability remains undetermined following the two efficient filtering steps. Our approach is grounded in the following three results, whose proofs appear in Appendix §A.3.

Proposition 3.6 (Prediction gap bound over the ϵ-Rashomon set). *Define $\Delta_{i,i'}(w) := \hat{y}_i - \hat{y}_{i'} = x_i^T w - x_{i'}^T w$ to be the prediction gap between instances i' and i under model w. For all i, i' and $w \in \mathcal{R}_\epsilon(w_0)$,*

$$\Delta_{i,i'}(w) \leq \Delta_{i,i'}(w_0) + \sqrt{\epsilon}\|x_i - x_{i'}\|_2 =: B(i, i'; \epsilon)$$

Corollary 3.7 (Provably unflippable points). *Suppose i is not in the top-κ for model w_0; i.e., $Top_{(i,w_0,\kappa)} = 0$. If $\#\{i' : B(i,i';\epsilon) < 0\} \geq \kappa$, then $Top_{(i,w,\kappa)} = 0 \;\; \forall w \in \mathcal{R}_\epsilon(w_0)$.*

Conceptually, Proposition 3.6 establishes a bound on the gap between the predicted values of any two points over the whole ϵ-Rashomon set in terms of the prediction gap under the baseline model, w_0. Corollary 3.7 then says that if there are at least κ points, $i' \neq i$, whose predicted value is guaranteed to exceed that of point i for every model $w \in \mathcal{R}_\epsilon(w_0)$, then i is unflippable.

Proposition 3.8 (Prediction maximizing model). *The predicted value of point i, $\hat{y}_i = x_i^T w$, over the ϵ-Rashomon set is maximized at,*

$$w^* = \underset{w \in \mathcal{R}_\epsilon(w_0)}{\operatorname{argmax}} \; \hat{y}_i(w) = w_0 + \sqrt{\epsilon}\frac{x_i}{\|x_i\|_2}. \tag{5}$$

For points that are not ruled out by Corollary 3.7, Proposition 3.8 provides a candidate model within the Rashomon set that may flip a point into the top-κ. Note that this result does not preclude the possibility that $Top_{(i,w^*,\kappa)} = 0$ while also $Top_{(i,w',\kappa)} = 1$ for some other $w' \in \mathcal{R}_\epsilon(w_0)$. Taken together, these results often significantly reduce the number of points for which one needs to run the MIP in order to determine their flippability.

4 MULTI-TARGET MULTIPLICITY AND FAIRNESS

In the previous section we introduced the top-κ ambiguity measure for characterizing predictive multiplicity for a single target, y, over the ϵ-Rashomon set. As discussed at the outset, an important potential source of multiplicity is in the specification of the target outcome itself. In this section we introduce a measure of multi-target multiplicity for the setting where the multiple targets will ultimately be combined in some way to produce a single score that will be used to prioritize allocation. We also discuss group fairness by examining how the selection rate for a given group varies depending on the specific choice of combining rule.

4.1 Multi-target ambiguity and index models

Given candidate targets $\tilde{y}^{(1)}, \dots, \tilde{y}^{(K)}$ and features X, we consider a family of "combining procedures", c_α, parameterized by α that map from training data $(X, \tilde{y}^{(1)}, \dots, \tilde{y}^{(K)})$ to the space of prediction models $\mathcal{H}_\alpha = \{h_\alpha : \mathcal{X} \mapsto \mathbb{R}\}$. Under a resource constraint of κ, resources will then be allocated to the units with the κ highest values of $h_\alpha(x_i)$. We are interested in characterizing how the top-κ set varies across the parameter α governing the combining procedure, c_α. More formally, we define *multi-target ambiguity* as follows.

Definition 4.1 (Multi-target ambiguity). The (α, κ)*-multi-target– ambiguity* of a combining procedure c_α over a sample S is the proportion of examples whose top-κ decision varies depending on the choice of α,

$$A_{\alpha,\kappa}(S) := \frac{1}{n}\sum_{i=1}^{n} \max_{h_\alpha, h_{\alpha'} \in \mathcal{H}_\alpha} \mathbb{1}[Top_{(i,h_\alpha,\kappa)} \neq Top_{(i,h_{\alpha'},\kappa)}]. \tag{6}$$

Whereas in the single target case we were interested in ambiguity over the ϵ-Rashomon set, here we focus on ambiguity over the *combining procedure.* Conceptually, a point is "ambiguous" if whether it is in the top-κ depends on the particular choice of α in the combining procedure.

To make the discussion more concrete, we introduce two combining procedures inspired by existing practice, the *index model* approach and the *index variable* approach.

Definition 4.2 (Index model). The *index model* is defined as

$$\hat{y}_{IM}(x;\alpha) = c_\alpha^{IM}(X, \tilde{y}^{(1)}, \dots, \tilde{y}^{(K)})(x) = \sum_{k=1}^{K} \alpha_k \hat{y}^{(k)}(x), \tag{7}$$

where α is a weight vector in the K-simplex, $\alpha \in \mathbb{S}^K := \{\alpha \in \mathbb{R}^K : \sum_{k=1}^K \alpha = 1, \alpha_k \geq 0 \; \forall k\}$, and $\hat{y}^{(k)}(x)$ is a prediction model for target $\tilde{y}^{(k)}$.

Note that for this definition to make sense, we assume that the individual predictors $\hat{y}^{(k)}$ are first standardized to an appropriate common scale, such as by rescaling $\hat{y} \leftarrow \frac{\hat{y} - mean(\hat{y})}{sd(\hat{y})}$ or converting to percentiles prior to combining. The choice of standardization function does influence results. Choosing a single target outcome k_0 is a special case of an index model with $\alpha_{k_0} = 1$ and $\alpha_k = 0$ for $k \neq k_0$. An advantage of the index model approach is that it places no restrictions on the training procedure used to construct $\hat{y}^{(k)}$. Where appropriate, multi-task learning approaches can be used to jointly learn models across the targets.

This approach is motivated by existing practice in domains such as criminal justice and human services, where multiple so-called scales (i.e., $\hat{y}^{(k)}$'s) are constructed to predict different outcomes, and are then aggregated into prioritization schemes or decision recommendations. For instance, the Allegheny Housing Assessment (AHA) tool used to prioritize housing services for persons experiencing homelessness sums the predictions of three $\tilde{y}^{(k)}$ assessed within 12 months of the assessment date: (i) the likelihoood of inpatient mental health services; (ii) the likelihood of jail booking; and (iii) the likelihood of 4 or more ER visits [24]. In criminal justice, pre-trial risk assessment tools such as the Public Safety Assessment (PSA) separately assess risk for failure to appear, new criminal activity, and new violent criminal activity, and then combine the scales through a decision-making framework [40][1].

An alternative to index models is an index variable approach, where instead of first forming predictions and then aggregating the different scales, a composite target outcome is formed and then that target is predicted.

Definition 4.3 (Index variable). Given candidate targets $\tilde{y}^{(1)}, \dots, \tilde{y}^{(K)}$, features X, and weights $\alpha \in \mathbb{S}^K$, an *index variable* model, $\hat{y}_{IV}(x; \alpha)$ is defined by the minimizer,

$$\hat{h}^{(\alpha)} = \min_{h \in \mathcal{H}} L(h; \tilde{y}^{(\alpha)}), \quad \text{where} \quad \tilde{y}^{(\alpha)} = \sum_{k=1}^{K} \alpha_k \tilde{y}^{(k)}. \tag{8}$$

Conceptually, the index variable approach can be thought of as first forming a composite outcome that is believed to more comprehensively describe some latent quantity, and then finding the optimal predictor for that outcome. Both for the index model and index variable formulation, the parameter α captures potential underspecification in the choice of target. In the case of linear models, the index model and index variable approach coincide.

Proposition 4.4 (Equivalence of index model and index variable approaches for linear models.). *If we restrict consideration to linear models whose solution takes the form $\hat{y} = M_X y$ for some $n \times n$ matrix M_X that depends on X but not on y, then the index model and index variable approach are equivalent.*

Note that linear regression is a special case of a linear model, with $M_X = X(X^T X)^{-1} X^T$. Other models such as regression splines fall into this class as well.

In the remainder of this work we focus on the index model approach, as it can be analysed in a computationally tractable way for general predictors $\hat{y}^{(k)}$. Due to the equivalence result, our methods are directly applicable to the index variable approach for linear models.

4.2 Computing multi-target top-κ ambiguity for index models

As in the single target setting, we calculate ambiguity by identifying flippable points using a MIP. In this setting there is no baseline model, so the term "flippable" now refers to points where there exist two choices of combining parameters, $\alpha \neq \alpha'$, such that $Top_{(i,\alpha,\kappa)} \neq Top_{(i,\alpha',\kappa)}$. Furthermore, the optimization here is no longer over an ϵ-Rashomon set—a notion which does not naturally extend to the multiple target setting due to the absence of a baseline model—but rather over the parameters α governing the combining rule. To identify flippable points, for each point in $i \in S$ we solve the following MIP formulation, which we call `FlipTopKMultiMIP`$(x_i; \kappa)$.

$$\min_{I \in \{0,1\}^{n-1}, \alpha \in \mathbb{R}^K} \sum_{i' \neq i} I_{(i' > i)} - 0.5 \sum_{k=1}^{K} \alpha_K \quad \text{or}$$

[1] This is not an endorsement of the usage of safety assessment tools in criminal justice. There are serious concerns about risk assessment in this domain.

$$\begin{aligned}
\max_{I \in \{0,1\}^{n-1},\, \alpha \in \mathbb{R}^K} \quad & \sum_{i' \neq i} I_{(i'>i)} + 0.5 \sum_{k=1}^{K} \alpha_K \\
\text{s.t.} \quad & I_{(i'>i)} + I_{(i>i')} = 1 && i' = 1, ..., n \setminus i && \text{(9a)} \\
& \hat{y}_{IM}(x_{i'}; \alpha) - \hat{y}_{IM}(x_i; \alpha) \leq M * I_{(i'>i)} && i' = 1, ..., n \setminus i && \text{(9b)} \\
& \hat{y}_{IM}(x_i; \alpha) - \hat{y}_{IM}(x_{i'}; \alpha) \leq M * I_{(i>i')} && i' = 1, ..., n \setminus i && \text{(9c)} \\
& 0 \leq \alpha_k \leq 1 && k = 1, ..., K && \text{(9d)} \\
& 0.1 \leq \sum_{k=1}^{K} \alpha_k \leq 1 && && \text{(9e)} \\
& \alpha_k \in \mathbb{R} && K = 0, ..., d && \text{(9f)} \\
& I_{(i'>i)}, I_{(i>i')} \in \{0, 1\} && i' = 1, ..., n \setminus i && \text{(9g)}
\end{aligned}$$

where $\hat{y}_{IM}(x_i; \alpha)$ is shorthand for $\sum_{k=1}^{K} \alpha_k \hat{y}^{(k)}(x_i)$, and

$$M = \max_{i',k} \hat{y}^{(k)}(x_{i'}) - \min_{i,k} \hat{y}^{(k)}(x_i).$$

This formulation is similar to `FlipTopKMIP` from the single-target case, but the objective has an additional term that forces $\sum \alpha_k = 1$ in the solution, and the optimization here is over the combining weights α rather than the parameters of the individual predictors $\hat{y}^{(k)}$.

As in the single-target context, we can once more reduce the number of times we need to run the MIP by identifying points that provably cannot appear in the top-κ set for any choice of α, and characterize the prediction-maximizing choice of α for each point. The results and accompanying proofs are presented in Appendix §B.2.

4.3 Group-level selection rates in top-κ selection with multiple targets

As discussed at the outset, allocation equity is a key motivation for considering multiple targets. Here we will focus on a specific fairness-related measure: the prevalence of instances from a given group in the top-κ. We will aim to characterize the variation in top-κ selection for a given group $A = a$ across the combining parameters α of an index model combining procedure.[2] Letting $G_a = \{i : A_i = a\}$ denote all the instances that are in protected group $A = a$, this can be expressed as,

$$\min_{\alpha} or \max_{\alpha} \sum_{i=1}^{n} \mathbb{1}[A = a] Top_{(i,\alpha,\kappa)} = \min_{\alpha} or \max_{\alpha} \sum_{i \in G_a} Top_{(i,\alpha,\kappa)} \tag{10}$$

We can formulate this as a MIP by introducing variables $T_i \in \{0, 1\}$ that play the role of the $Top_{(i,\alpha,\kappa)}$ indicator. We call this MIP `GroupSelectRateTopKMultiMIP`$(a; \kappa)$.

$$\begin{aligned}
\min_{I \in \{0,1\}^{2n \times |G_a|},\, T \in \{0,1\}^{|G_a|},\, \alpha \in \mathbb{S}^K} \quad & \sum_{i \in G_a} T_i - 0.5 \sum_{k=1}^{K} \alpha_k \ \text{ or} \\
\max_{I \in \{0,1\}^{2n \times |G_a|},\, T \in \{0,1\}^{|G_a|},\, \alpha \in \mathbb{S}^K} \quad & \sum_{i \in G_a} T_i + 0.5 \sum_{k=1}^{K} \alpha_k \\
\text{s.t.} \quad & I_{(i'>i)} + I_{(i>i')} = 1 && \forall\ i \in G_a,\ i' = 1, ..., n \setminus i && \text{(11a)} \\
& \hat{y}_{IM}(x_{i'}; \alpha) - \hat{y}_{IM}(x_i; \alpha) \leq M_I * I_{(i'>i)} && \forall\ i \in G_a,\ i' = 1, ..., n \setminus i && \text{(11b)}
\end{aligned}$$

[2] While, in principle, one can also examine measures such as the False Positive Rate and True Positive Rate by analysing the subsample of instances for which $\tilde{y}^{(k)} = 0$ (or 1, for TPR), it is not entire clear how such quantities should be interpreted. How should one weigh a high FPR for a given target against a low FPR for a different one in a setting where the "correct" choice of target is itself in doubt?

$$
\begin{aligned}
\hat{y}_{IM}(x_i;\alpha) - \hat{y}_{IM}(x_{i'};\alpha) &\leq M_I * I_{(i>i')} && \forall\ i \in G_a,\ i' = 1,\ldots,n \setminus i && \text{(11c)}\\
\kappa - \sum_{i' \neq i} I_{(i'>i)} &\leq M_T^0 * T_i && i \in G_a && \text{(11d)}\\
\left(1 + \sum_{i' \neq i} I_{(i'>i)}\right) - \kappa &\leq M_T^1 * (1 - T_i) && i \in G_a && \text{(11e)}\\
0 \leq \alpha_k &\leq 1 && k = 1,\ldots,K && \text{(11f)}\\
0.1 \leq \sum_{k=1}^{K} \alpha_k \leq 1 & && && \text{(11g)}\\
\alpha_k &\in \mathbb{R} && k = 0,\ldots,d && \text{(11h)}\\
I_{(i'>i)}, I_{(i>i')} &\in \{0,1\} && i \neq i' = 1,\ldots,n && \text{(11i)}\\
T_i &\in \{0,1\} && i \in G_a && \text{(11j)}
\end{aligned}
$$

Here we set

$$
\begin{aligned}
M_I &= \max_{i,k} \hat{y}^{(k)}(x_i) - \min_{i,k} \hat{y}^{(k)}(x_i),\\
M_T^0 &= \kappa,\\
M_T^1 &= n - \kappa.
\end{aligned}
$$

5 EVALUATION

In this section, we apply the techniques developed above to a large healthcare dataset to better understand the opportunities afforded by multiplicity among multiple target variables. We evaluate variation in top-κ selection rates (4.3) to evaluate allocation equity over the index model formulation by re-weighting target choices. Then, we use synthetic data to systematically gain insight into the relationship between target characteristics and top-κ selection rates for individuals in a protected group. Finally, we compute predictive multiplicity measures (2),(3) for each target choice as well as the multi-target ambiguity (6) to better understand the latitude afforded by each approach. Throughout this section we solve all integer programs with Gurobi v.9.5.2 [16].

5.1 Dataset

We demonstrate our framework on a dataset released by Obermeyer et al. [32], which is unique in several ways. The original paper examines patient data for all primary care patients at a large academic hospital. However, due to the sensitivity of the data, the authors were unable to release the dataset in its original form. Instead, they created a publicly available semi-synthetic version of the dataset that is designed to closely mirror the original dataset.[3]

The released dataset contains several related but different outcomes for patients in a given year including total healthcare costs, avoidable healthcare costs (emergency visits and hospitalizations), and number of active chronic illnesses. It also contains a rich set of features about each patient, including demographics (age, sex, race) and information about the patient's health and healthcare costs in the previous year. Specifically, there are indicators for individual chronic illnesses that a patient had in the previous year, costs claimed by the patients' insurer in the previous year, biomarkers for medical tests from the previous year, and medications taken in the previous year.

In the original paper, Obermeyer et al. examine a proprietary scoring system used by the hospital to identify high-risk patients. The risk scores are generated by a model designed to predict healthcare costs in the current year based on

[3] https://gitlab.com/labsysmed/dissecting-bias

patient demographics and healthcare information available from the previous year. In particular, patients who are assigned risk scores that fall in the 97th percentile or above (i.e., the top 3% of assigned scores) are automatically identified for inclusion in the hospital's "high-risk care management" program.

The authors examine the assigned risk scores in detail and show that they contain a significant racial bias. Specifically, they find that Black patients at a given risk score have worse health outcomes, on average, than their white counterparts. The authors trace this bias back to the choice of predicting healthcare costs as the target variable. Due to differences in access to healthcare, white patients tend to have higher healthcare costs, on average, than Black patients of similar health. This difference is then reflected in the developed risk score, leading to the observed racial bias. Obermeyer et al. then go on to show that there are different target variable choices that exhibit less of a racial bias—specifically using either avoidable costs or active chronic illnesses as a target instead of total costs.

5.2 Optimizing across healthcare outcomes

We present a re-examination of this healthcare dataset to further explore the ways in which flexibility in target variable choice can be used to mitigate fairness concerns. Obermeyer et al. consider using one of each of three different target variables, which in our framework corresponds to an index model with binary α weights. For example, the cost model can be thought of as $\hat{y}_{IM} = 1 \cdot \hat{y}^{(\text{costs})} + 0 \cdot \hat{y}^{(\text{avoidable costs})} + 0 \cdot \hat{y}^{(\text{active illnesses})}$. However, these are just three extremes among the possible set of index models that can be formed with a continuous α to create a weighted average of the three available target variables.

Our analysis explores whether exercising these extra degrees of freedom can lead to more equitable outcomes. To address this, we replicate and extend the analysis in Table 2 of the original paper, using the released dataset.[4] Specifically, we train separate models to predict each of the three target variables (healthcare costs, avoidable costs, and active chronic illnesses) and use the fitted models to rank a held-out set of patients.[5] We identify the top 3% of highest-risk patients according to each of the models and look at the concentration of outcomes and the racial composition of the identified patients. For instance, when considering total costs, we compute what percent of all costs (across all patients) are covered by just the highest-risk patients. When considering active chronic illnesses, we instead compute the fraction of all illnesses (across all patients) covered by this set. We then extend these results by running the multi-target fairness mixed-integer programming (MIP), `GroupSelectRateTopKMultiMIP`, to search for an index model that maximizes the fraction of Black patients concentrated among the highest-risk set. Appendix section C outlines more details of this process.

The results are displayed in Fig. 1 and show several key observations. First, we see that, as expected, the model trained to predict a given individual outcome has the largest concentration of that outcome in the high-risk patient set. Notice that the transparent bars on the far left panel of Fig 1A show that the model trained to predict total cost is the one that has the highest concentration of total costs in the high-risk patient set. Conversely, on the far right panel we see that modeling active chronic conditions produces the highest concentration of current illnesses in the high-risk set. Second, despite these differences we see comparatively small variation in outcome concentration across different target variable choices, with less than a 5 percentage point difference across models in the first three panels. But, we do see a substantial difference in the racial composition of the high-risk set, as indicated in Fig 1B—a more than 10 percentage point difference.

[4] The original table is generated using the proprietary, unreleased data. Replicating the table with the released dataset produces similar, but not identical results for this reason.

[5] We use the train/holdout set specified by the authors in the released dataset. We train OLS linear regression models for each variable. In order to do so, we remove several co-linear features provided in the released dataset.

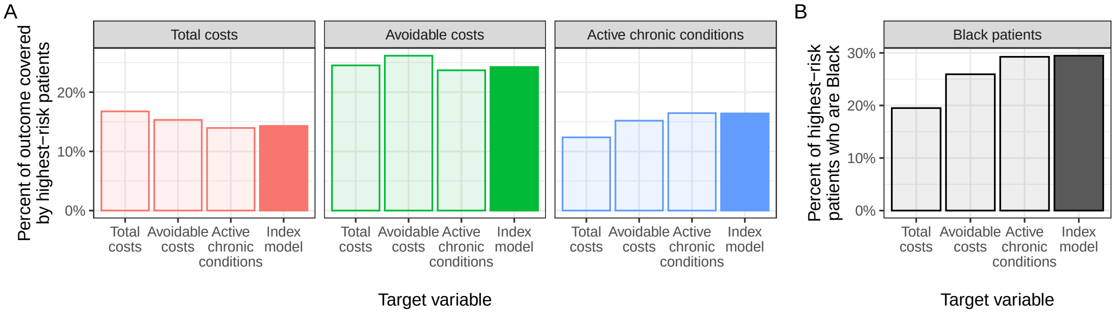


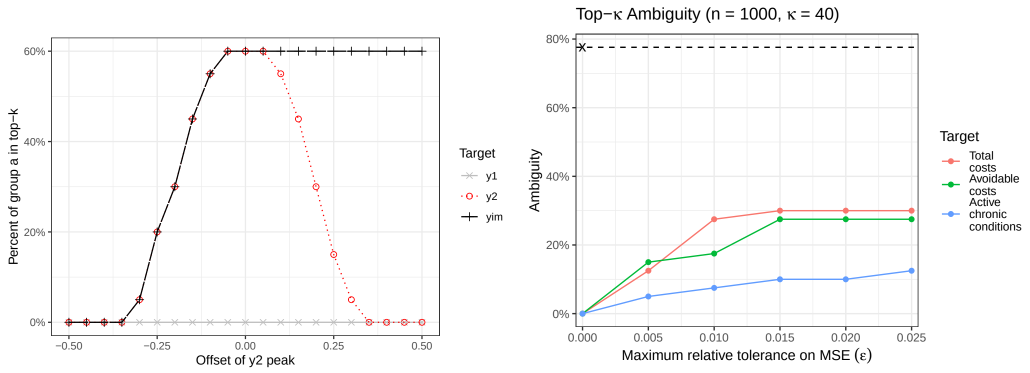
Fig. 1. A. The concentration of various outcomes under models optimized for different targets. Each panel shows the percent of an outcome captured by the highest-risk patients relative to the entire outcome distribution across all patients. Each bar represents one type of model. The transparent bars depict models trained to predict individual targets, whereas the solid bars depict the index model, which re-weights the individual predictions to maximize fairness. B. The percent of black patients among highest-risk patients identified by each model.

The index model, in comparison, is shown in the solid bars of Fig 1. By comparing the solid bars to the transparent ones, we see that the index model does a reasonable job of capturing each of the individual targets that it is comprised of, but also produces a high-risk set with a high concentration of black patients, as per the objective of the multi-target group selection formulation (10). In effect, this represents a "best of both worlds" solution: we are able to fit separate models that are useful for predicting the three outcomes that may be of interest on their own (i.e., for budgeting purposes), but we also arrive at a way of ranking patients that results in a more equitable allocation of a scarce resource via the index model.

5.3 Exploring the conditions for effective multi-target optimization

In the example above, we found it was possible to learn an index model that combined individual target variables from the healthcare dataset to improve group selection rates. In this section, we use synthetic data to gain a better understanding of the conditions for which we should (or should not) expect to see such gains in other datasets. To do this, we produce toy data to systematically control the relationship between a protected group attribute a, a feature x and the different choices of target $\tilde{y}^{(k)}$. We then vary these relationships and examine how this effects the group selection rate that an index model can achieve.

Specifically, we construct a dataset with one feature (age) and two target variables $\tilde{y}^{(1)}$ and $\tilde{y}^{(2)}$, along with a protected attribute (race). We construct a scenario where race is non-monotonically correlated with age, with a high concentration of Black patients in the middle age range compared to the rest of the population. We then construct one target variable $\tilde{y}^{(1)}$ that is negatively correlated with age and one target variable $\tilde{y}^{(2)}$ whose correlation with age varies from strongly positive to strongly negative, controlled by a parameter b that sets the location of its peak relative to the peak of $\tilde{y}^{(2)}$. In this setting, prioritizing middle-aged patients maximizes the fraction of black patients in the high-risk set, but fitting a model to $\tilde{y}^{(1)}$ prioritizes young patients, resulting in a lower rate of Black patients among the selected set. Conversely, when b is large and positive, $\tilde{y}^{(2)}$ is positively correlated with age, and so fitting a model to it will prioritize older patients, also leading to sub-optimal group selection. However, as we show in Fig. 2A, an index model can be fit over a wide range of b values such that the group selection rate is maximized. The intuition is that the index model can learn to average out unhelpful correlation structure between the protected attribute and the target variables.



(A) (B)

Fig. 2. (A) Group selection rates on synthetic data as the relationship between two different targets, $\tilde{y}^{(1)}$ and $\tilde{y}^{(2)}$, is varied. While $\tilde{y}^{(1)}$ (solid grey) is not helpful in maximizing group selection and $\tilde{y}^{(2)}$ (dotted red) is helpful over a limited range, an index model (dashed black) can find a helpful weighting of the two targets over a wide range of conditions. (B) Comparison of multiplicity within vs. between targets. Ambiguity within each individual target is shown by the colored lines and different relative mean squared error tolerances. Ambiguity across the three targets is shown by the black 'x' and dotted line.

5.4 Multi-target versus individual-target multiplicity

Finally, we compare the latitude afforded by across-target multiplicity (index model) to that for within-target multiplicity (predictive multiplicity). To do this we return to the original dataset released by Obermeyer et al. and work with a subset of the features and examples for computational efficiency. We evaluate predictive multiplicity by computing the single-target top-κ ambiguity for each choice of target variable Eq (3) by running `FlipTopKMIP` for different error tolerances ϵ. This allows us to determine the proportion of top-κ points that can be flipped. The results are shown in Fig. 2B, with each color corresponding to a choice of target variable. From this, we see that single-target ambiguity rises quickly with ϵ and then plateaus. This is a result of the ranked output setting. The total cost variable has the highest ambiguity, at about 35%, whereas active chronic conditions plateau just above 10%.

We compute the multi-target top-κ ambiguity (6) by running `FlipTopKMultiMIP` across the three different target variables. This results in a multi-target ambiguity of 78%, as indicated by the black "x" and dashed horizontal line in Fig. 2B. From this we see that the across-target multiplicity is substantially higher than the within-target multiplicity—a much higher proportion of points can be flipped into the top-κ set by re-weighting predictions for the different targets than by entertaining slightly sub-optimal model fits for the individual targets. On the other hand, this points to more predictive consistency over the individual Rashomon sets compared to the option of combining of target choices.

6 CONCLUDING REMARKS

In this paper, we have examined the benefits of exploring the roads not taken in problem formulation. In particular, we have developed two frameworks for leveraging flexibility in predictive modeling under resource constraints to address fairness concerns. First, we show how to measure and exploit multiplicity for a given target variable in settings where decision makers face constraints that limit the total number of people who can receive a scarce resource. Second, we show that when faced with a choice of multiple target variables, practitioners can develop index models that address fairness concerns by re-weighting and combining predictions for each target. Our empirical results show that both of these methods are effective for addressing fairness concerns in a large healthcare dataset. Notably we find that the latitude afforded by re-weighting predictions across target variables is substantially larger than the flexibility provided by leveraging within-target multiplicity. This may represent a "best of both worlds" solution: we are able to fit separate models for predicting outcomes that may be interesting to model in their own right, but we can also combine the predictions from these models to allocate resources more equitably.

ACKNOWLEDGMENTS

We thank David C. Parkes and MSR FATE team members for feedback and helpful discussions. JWD is supported by a Ford Foundation Pre-doctoral Fellowship and the NSF Graduate Research Fellowship Program under Grant No. DGE1745303. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Junaid Ali, Preethi Lahoti, and Krishna P. Gummadi. 2021. *Accounting for Model Uncertainty in Algorithmic Discrimination*. Vol. 1. Association for Computing Machinery. 336–345 pages. https://doi.org/10.1145/3461702.3462630

[2] Solon Barocas and Andrew D Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016), 671.

[3] Emily Black, Hadi Elzayn, Alexandra Chouldechova, Jacob Goldin, and Daniel Ho. 2022. Algorithmic fairness and vertical equity: Income fairness with IRS tax audit models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1479–1503.

[4] Emily Black and Matt Fredrikson. 2021. Leave-one-out Unfairness. (2021).

[5] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model multiplicity: Opportunities, concerns, and solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 850–863.

[6] Leo Breiman. 2001. Statistical modeling: The two cultures. *Statist. Sci.* 16, 3 (2001), 199–215. https://doi.org/10.1214/ss/1009213726

[7] A. Feder Cooper, Solon Barocas, Christopher De Sa, and Siddhartha Sen. 2023. Variance, Self-Consistency, and Arbitrariness in Fair Classification. https://doi.org/10.48550/ARXIV.2301.11562

[8] Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. 2022. A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms. *arXiv preprint arXiv:2206.14983* (2022).

[9] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. 2021. Characterizing Fairness Over the Set of Good Models Under Selective Labels. (2021). http://arxiv.org/abs/2101.00352

[10] Maria De-Arteaga, Stefan Feuerriegel, and Maytal Saar-Tsechansky. 2022. Algorithmic fairness in business analytics: Directions for research and practice. *Production and Operations Management* 31, 10 (2022), 3749–3770.

[11] Jiayun Dong and Cynthia Rudin. 2019. Variable Importance Clouds: A Way to Explore Variable Importance for the Set of Good Models. (2019). http://arxiv.org/abs/1901.03209

[12] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv* (2020).

[13] Sina Fazelpour and David Danks. 2021. Algorithmic bias: Senses, sources, solutions. *Philosophy Compass* 16, 8 (2021), e12760.

[14] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* 20, Vi (2019).

[15] Saul Gass and Thomas Saaty. 1955. The computational algorithm for the parametric objective function. *Naval research logistics quarterly* 2, 1-2 (1955), 39–45.

[16] Gurobi Optimization, LLC. 2023. Gurobi Optimizer Reference Manual. https://www.gurobi.com

[17] David J Hand. 1994. Deconstructing statistical questions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 157, 3 (1994), 317–338.

[18] David J. Hand. 2006. Classifier Technology and the Illusion of Progress. *Statist. Sci.* 21, 1 (2006), 1 – 14. https://doi.org/10.1214/088342306000000060

[19] David J Hand. 2016. *Measurement: A very short introduction*. Oxford University Press.

[20] Hsiang Hsu and Flavio du Pin Calmon. 2022. Rashomon Capacity: A Metric for Predictive Multiplicity in Classification. https://doi.org/10.48550/ARXIV.2206.01295

[21] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 375–385. https://doi.org/10.1145/3442188.3445901

[22] Heinrich Jiang and Ofir Nachum. 2020. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 702–712.

[23] Pauline T Kim. 2022. Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action. *California Law Review* 110 (2022), 1539.

[24] Chamari I Kithulgoda, Rhema Vaithianathan, and Dennis P Culhane. 2022. Predictive risk modeling to identify homeless clients at risk for prioritizing services using routinely collected data. *Journal of Technology in Human Services* 40, 2 (2022), 134–156.

[25] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2022. The challenge of understanding what users want: Inconsistent preferences and engagement optimization. *arXiv preprint arXiv:2202.11776* (2022).

[26] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

[27] Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. 2020. Participatory problem formulation for fairer machine learning through community based system dynamics. *arXiv preprint arXiv:2005.07572* (2020).

[28] Charles Marx, Flavio P. Calmon, and Berk Ustun. 2019. Predictive multiplicity in classification.

[29] Smitha Milli, Luca Belli, and Moritz Hardt. 2021. From optimizing engagement to measuring value. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 714–722.

[30] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163.

[31] Sendhil Mullainathan and Ziad Obermeyer. 2021. On the inequity of predicting a while hoping for B. In *AEA Papers and Proceedings*, Vol. 111. 37–42.

[32] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. https://doi.org/10.1126/science.aax2342

[33] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* (2019), 39–48. https://doi.org/10.1145/3287560.3287567

[34] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. On counterfactual explanations under predictive multiplicity. *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence, UAI 2020* 124 (2020), 839–848.

[35] Stephen Pfohl, Yizhe Xu, Agata Foryciarz, Nikolaos Ignatiadis, Julian Genkins, and Nigam Shah. 2022. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1039–1052.

[36] Foster Provost and Tom Fawcett. 2013. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* O'Reilly Media, Inc.

[37] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence* 3, 10 (2021), 896–904.

[38] Aaron Roth, Alexander Tolbert, and Scott Weinstein. 2022. Reconciling Individual Probability Forecasts. https://doi.org/10.48550/ARXIV.2209.01687

[39] Lesia Semenova, Cynthia Rudin, and Ronald Parr. 2019. A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. (2019), 1–64. http://arxiv.org/abs/1908.01755

[40] Arnold Ventures. 2022. What is the PSA? https://advancingpretrial.org/psa/about/

[41] Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H. Chi. 2021. Understanding and Improving Fairness-Accuracy Trade-offs in Multi-Task Learning. *CoRR* abs/2106.02705 (2021). arXiv:2106.02705 https://arxiv.org/abs/2106.02705

[42] Jamelle Watson-Daniels, David C. Parkes, and Berk Ustun. 2022. Predictive Multiplicity in Probabilistic Classification. (2022), 1–24. http://arxiv.org/abs/2206.01131

A TECHNICAL DETAILS FOR RANKED SINGLE TARGET MULTIPLICITY

This section presents the derivations and proofs of technical results appearing in Section 3.

A.1 ϵ-Rashomon set for linear regression

In the §3 on single-target resource constrained predictive multiplicity, we repeatedly use the fact that, for orthonormal design matrices, X, the ϵ-Rashomon set is given by

$$\mathcal{R}_\epsilon(w_0) = \{w \in \mathbb{R}^{d+1} : \|w_0 - w\|_2^2 \leq \epsilon\}.$$

Here we provide a quick proof, which follows as a Corollary of Theorem 10 in Semenova et al. [39].

Proof. Unpenalized linear regression is a special case of ridge regression

$$\min_w L(w; \lambda) = \min_w (y - Xw)^T (y - Xw) + \lambda \|w\|_2^2,$$

with $\lambda = 0$. Part 1 of Theorem 10 of Semenova et al. [39] shows that the ϵ-Rashomon set for ridge regression is,

$$\mathcal{R}_\epsilon(w_0; X, \lambda) = \{w : (w - w_0)^T \left(X^T X + \lambda I_{d+1}\right) (w - w_0) \leq \epsilon\}.$$

For orthonormal designs, $X^T X = I_{d+1}$. This, combined with taking $\lambda = 0$ to recover the unpenalized linear regression setting gives the stated result. □

A.2 M bound in constraints (4b) and (4c)

To ensure that $I_{(i' > i)}$ whenever $(x_{i'} - x_i)^T w = \hat{y}(x_{i'}) - \hat{y}(x_i) > 0$ we need to choose M so that

$$M \geq \hat{y}(x_{i'}) - \hat{y}(x_i) \quad \forall i', i, \text{ and } \quad \forall w \in \mathcal{R}_\epsilon(w_0)$$

Proposition A.1.

$$\hat{y}(x_{i'}) - \hat{y}(x_i) \leq \left(\sqrt{\|w_0\|_2^2 + \epsilon}\right) \max_{i,j} \|x_j - x_i\|_2 \quad \forall i', i, \quad \textit{and} \quad \forall w \in \mathcal{R}_\epsilon(w_0)$$

Proof.

$$\max_{w \in \mathcal{R}_\epsilon(w_0)} \hat{y}_{i'} - \hat{y}_i = \max_{w \in \mathcal{R}_\epsilon(w_0)} (x_{i'} - x_i)^T w$$

By Cauchy-Schwartz,

$$(x_{i'} - x_i)^T w \leq \|x_{i'} - x_i\|_2 \|w\|_2 \leq \|x_{i'} - x_i\|_2 \max_{w \in \mathcal{R}_\epsilon(w_0)} \|w\|_2.$$

Noting that

$$\|w\|_2 = \sqrt{\|w_0 + (w - w_0)\|_2^2} \leq \sqrt{\|w_0\|_2^2 + \|(w - w_0)\|_2^2} \leq \sqrt{\|w_0\|_2^2 + \epsilon} \quad \forall w \in R_\epsilon(w_0),$$

we therefore get that,

$$M_i = \max_{i'} \max_{w \in \mathcal{R}_\epsilon} \hat{y}_{i'} - \hat{y}_i \leq \left(\sqrt{\|w_0\|_2^2 + \epsilon}\right) \max_{i'} \|x_{i'} - x_i\|_2.$$

Taking the maximum over all i' gives the desired result,

$$M = \max_{i,i'} \max_{w \in \mathcal{R}_\epsilon(w_0)} \hat{y}_{i'} - \hat{Y}_i \leq \left(\sqrt{\|w_0\|_2^2 + \epsilon}\right) \max_{i,i'} \|x_{i'} - x_i\|_2.$$

□

Note that the proof shows that one can set M_i differently for each point i we are aiming to flip in the given run of the MIP.

A.3 Identifying certifiably (un)flippable points without solving a MIP

Proof of Proposition 3.6.

$$\begin{aligned}\Delta_{i,i'}(w) &= x_i^T w - x_{i'}^T w = (x_i - x_i)^\top w \\ &= (x_i - x_i)^\top w + (x_i - x_{i'})^\top \hat{w} - (x_i - x_i)^\top \hat{w} \\ &= (x_i - x_{i'})^\top (w - \hat{w}) + (x_i - x_{i'})^\top \hat{w}\end{aligned}$$

By Cauchy-Schwartz,

$$\begin{aligned}\left|(x_i - x_{i'})^T (w - \hat{w})\right| &\leq \|x_i - x_{i'}\|_2 \|w - \hat{w}\|_2 \\ &\leq \sqrt{\epsilon}\|x_i - x_{i'}\|_2,\end{aligned}$$

where in the second step we use the fact that $w \in \mathcal{R}_\epsilon(w_0)$.

Thus $\forall w \in \mathcal{R}_\epsilon(w_0)$,

$$\Delta_{i,i'}(w) \leq \Delta_{i,i'}(\hat{w}) + \sqrt{\epsilon}\|x_i - x_{i'}\|_2 = B(i, i'; \epsilon).$$

So if $B(i, i'; \epsilon) < 0$, we have $\Delta_{i,i'}(w) < 0 \; \forall w \in \mathcal{R}_\epsilon(w_0)$.

□

Proof of Corollary 3.7. $\{i' : B(i, i'; \epsilon) < 0)\} \geq \kappa$ means that there are at least κ points for which $\Delta_{i,i'}(w) < 0 \, \forall w \in \mathcal{R}_\epsilon(w_0)$, so i cannot be in the top-κ set for any model in the ϵ-Rashomon set. □

Proof of Proposition 3.8. Let

$$w^* = w_0 + \sqrt{\epsilon}\frac{x_i}{\|x_i\|_2}.$$

We will show that $\forall w \in \mathcal{R}_\epsilon(w_0)$, $\hat{y}_i(w) \leq \hat{y}_i(w^*)$. By construction,

$$\hat{y}_i(w^*) = x^T w_0 + \sqrt{\epsilon}\|x_i\|_2.$$

By Cauchy-Schwartz, for any $w \in \mathcal{R}_\epsilon(w_0)$

$$\begin{aligned}\hat{y}_i(w) &= x_i^T w_0 + x_i^T (w - w_0) \\ &\leq x_i^T w_0 + \|x_i\|_2 \|w - w_0\|_2 \\ &\leq x_i^T w_0 + \sqrt{\epsilon}\|x_i\|_2 \\ &= \hat{y}_i(w^*)\end{aligned}$$

□

B TECHNICAL DETAILS FOR MULTI-TARGET MULTIPLICITY AND FAIRNESS

B.1 Equivalence of index model and index variable approaches for linear models

Proof of Proposition 4.4. Starting with the index variable definition, we get that

$$\hat{y}_{IV}^{(\alpha)} = M_X \tilde{y}^{(\alpha)} = M_X \left(\sum_{k=1}^{K} \alpha_k \tilde{y}^{(k)} \right) = \sum_{k=1}^{K} \alpha_k M_X \tilde{y}^{(k)} = \sum_{k=1}^{K} \alpha_k \hat{y}^{(k)} = \hat{y}_{IM}^{(\alpha)}$$

□

B.2 Identifying certifiably (un)flippable points in the multi-target setting without solving a MIP

Proposition B.1 (Prediction gap bound for index models.). *Let $\Delta_{i,i'}(\alpha) := \hat{y}_{IM}(x_i; \alpha) - \hat{y}_{IM}(x_{i'}; \alpha)$ to be the prediction gap between instances i' and i under combining parameters α. For all i, i' and $\alpha, \alpha' \in \mathbb{S}^K$,*

$$\Delta_{i,i'}(\alpha) \leq \Delta_{i,i'}(\alpha') + \sum_{k=1}^{K} |\hat{y}^{(k)}(x_i) - \hat{y}^{(k)}(x_{i'})| =: B_{IM}(i, i'; \alpha).$$

Proof. For any two instances $x_i, x_{i'} \in \mathcal{X}$ and combining parameter vectors $\alpha, \alpha' \in \mathbb{S}^K$,

$$\begin{aligned}
\Delta_{i,i'}(\alpha) &= \Delta_{i,i'}(\alpha') + \sum_{k=1}^{K} (\alpha_k - \alpha'_k) \left(\hat{y}^{(k)}(x_i) - \hat{y}^{(k)}(x_{i'}) \right) \\
&\leq \Delta_{i,i'}(\alpha') + \sum_{k=1}^{K} \left| \hat{y}^{(k)}(x_i) - \hat{y}^{(k)}(x_{i'}) \right| \\
&= B_{IM}(i, i'; \alpha)
\end{aligned}$$

□

Corollary B.2 (Points that cannot appear in top-κ set for any index model). *Suppose i is not in the top-κ for an index model with parameter $\tilde{\alpha}$; i.e., $Top_{(i,\tilde{\alpha},\kappa)} = 0$. If $\#\{i' : B(i, i'; \tilde{\alpha}) < 0\} \geq \kappa$, then $Top_{(i,w,\kappa)} = 0 \;\; \forall \alpha \in \mathbb{S}^K$.*

Proof. $\{i' : B_{IM}(i, i'; \tilde{\alpha}) < 0)\} \geq \kappa$ means that there are at least κ points, i', for which $\Delta_{i,i'}(\alpha) < 0 \, \forall \alpha \in \mathbb{S}^K$, so i cannot be in the top-κ set for any index model. □

Proposition B.1 establishes a bound on the gap between the predicted values of any two points *for all* $\alpha \in \mathbb{S}^K$ in terms of the prediction gap under *any one* choice of combining parameters α. Corollary B.2 then allows us to determine when i cannot be in the top-κ of *any* index model $\alpha \in \mathbb{S}^K$ based on the prediction gap for a *given* $\tilde{\alpha}$.

Proposition B.3 (Prediction maximizing index model). *The predicted value of point i is maximized at $\alpha^* \in \mathbb{S}^K$ where $\alpha^*_{k^*} = 1$ for $k^* = \text{argmax}_k \, \hat{y}^{(k)}(x_i)$ and $\alpha^*_k = 0$ for $k \neq k^*$.*

Proof.

$$\max_{\alpha \in \mathbb{S}^K} \hat{y}_{IM}^{(\alpha)}(x_i) = \max_{\alpha \in \mathbb{S}^K} \sum_{k=1}^{K} \alpha_k \hat{y}^{(k)} \leq \max_k \hat{y}^{(k)} \sum_{k=1}^{K} \alpha_k = \max_k \hat{y}^{(k)},$$

which is achieved at the stated value of the combining parameter vector, α^*. □

Proposition B.3 provides a candidate α for which a given point may be in the index model's top-κ. Note that this result does not preclude the possibility that $Top_{(i,\alpha^*,\kappa)} = 0$ while also $Top_{(i,\alpha',\kappa)} = 1$ for some other $\alpha' \in \mathbb{S}^K$. This result suggests the simple strategy of first identifying points whose top-κ decision varies between the single-target prediction models $\hat{y}^{(k)}$.

C MAXIMIZING FAIRNESS WITH MULTIPLE TARGETS

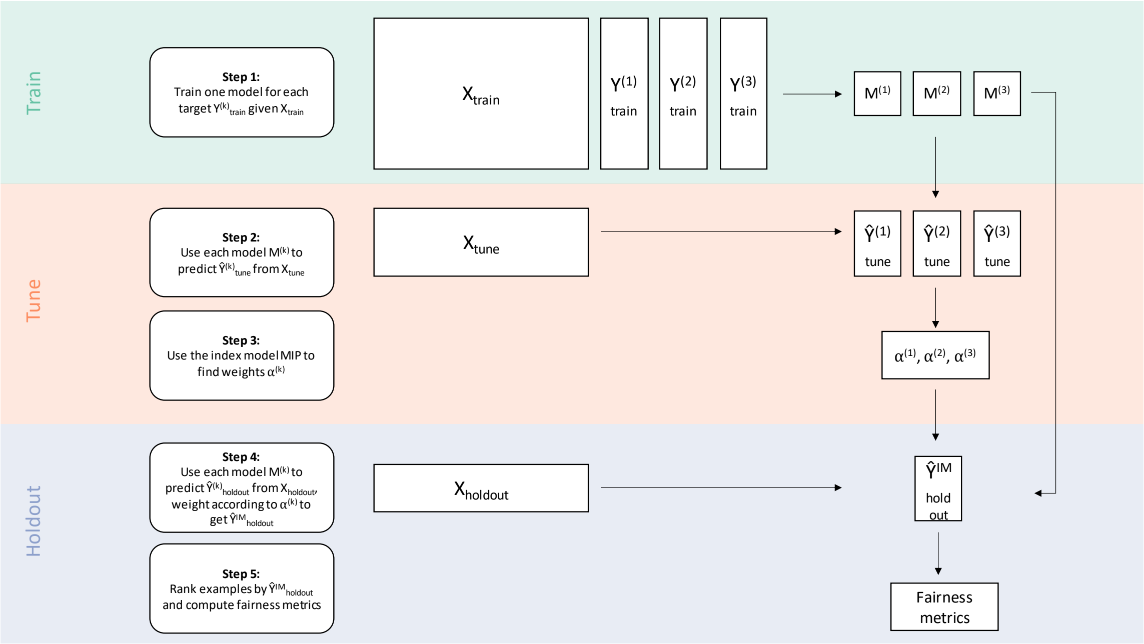

Train
Tune
Holdout
Step 1:
Train one model for each target $Y^{(k)}_{train}$ given X_{train}
X_{train}
$Y^{(1)}$ train
$Y^{(2)}$ train
$Y^{(3)}$ train
$M^{(1)}$
$M^{(2)}$
$M^{(3)}$
Step 2:
Use each model $M^{(k)}$ to predict $\hat{Y}^{(k)}_{tune}$ from X_{tune}
X_{tune}
$\hat{Y}^{(1)}$ tune
$\hat{Y}^{(2)}$ tune
$\hat{Y}^{(3)}$ tune
Step 3:
Use the index model MIP to find weights $\alpha^{(k)}$
$\alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}$
Step 4:
Use each model $M^{(k)}$ to predict $\hat{Y}^{(k)}_{holdout}$ from $X_{holdout}$, weight according to $\alpha^{(k)}$ to get $\hat{Y}^{IM}_{holdout}$
$X_{holdout}$
$\hat{Y}^{IM}$ hold out
Step 5:
Rank examples by $\hat{Y}^{IM}_{holdout}$ and compute fairness metrics
Fairness metrics

Fig. 3. Workflow for maximizing fairness with multiple targets. Training data is used to fit separate models for each target variable. In the tune phase, the fitted models are used to forecast each target variable, and the index model MIP is run to find a fairness-maximizing weighted combination of the targets. Finally, in the holdout phase a separate dataset is used to calculate the weight index model predictions, from which fairness metrics are computed.