When Personalization Harms Performance: Reconsidering the Use of Group Attributes in Prediction

Vinith M. Suriyakumar [1] **Marzyeh Ghassemi** [1 *] **Berk Ustun** [2 *]

Abstract

Machine learning models are often personalized with categorical attributes that define groups. In this work, we show that personalization with *group attributes* can inadvertently reduce performance at a *group level* – i.e., groups may receive unnecessarily inaccurate predictions by sharing their personal characteristics. We present formal conditions to ensure the *fair use* of group attributes in a prediction task, and describe how they can be checked by training one additional model. We characterize how fair use conditions be violated due to standard practices in model development, and study the prevalence of fair use violations in clinical prediction tasks. Our results show that personalization often fails to produce a tailored performance gain for every group who reports personal data, and underscore the need to evaluate fair use when personalizing models with characteristics that are protected, sensitive, self-reported, or costly to acquire.

1. Introduction

Machine learning models are often used to assign predictions to people – be it to predict if a patient has a rare disease, the risk that a consumer will default on a loan, or the likelihood that a student will matriculate.

Models in such applications are often *personalized* to target heterogeneous subgroups. In the most common approach, models are trained with *group attributes* – i.e., categorical attributes that define groups. In consumer finance, credit scores may include group attributes that are *protected* such as `age_group` [21]. In medicine, clinical prediction models may include group attributes that are *sensitive* (e.g., `AIDS` as in the SAPS II Score), *self-reported* (e.g. `sexual_practices` as in the Denver HIV Risk Score), or *costly to acquire* (e.g., Brief Psychiatry Rating Scale).

The widespread use of group attributes in modern prediction models reflects a belief that personalization can only improve performance. In effect, practitioners who develop clinical prediction models include protected attributes like `race` because they believe it can only improve performance [53]. Likewise, individuals report sensitive attributes like `sexual_practices` to a self-reported screening tool because they expect to receive more accurate predictions.

In this work, we formalize these expectations through a principle that we call *fair use* – i.e., that every person who reports personal characteristics should expect receive a tailored gain in performance in return. Given a model that is personalized with group attributes, we then test that it satisfies these minimal expectations. First, by testing that every group expects more accurate predictions from a personalized compared to a *generic model* trained without their group attributes. Next, by testing that the gains are *tailored*, meaning that every group prefers their personalized predictions to predictions personalized for any other group.

The vast majority of machine learning models are not designed to ensure fair use (see Fig. 1). This result stems from the fact that standard approaches to empirical risk minimization use group attributes to improve performance at a *population level*. As we will show, the resulting models may assign unnecessarily inaccurate predictions at the group level due to routine decisions such as model specification and model selection (see Fig. 2).

In practice, however, these fair use violations may inflict harm. In clinical applications, for example, inaccurate predictions lead to worse decisions and health outcomes [88]. More broadly, these effects are silent and avoidable. Silent as fair use violations would only draw attention if we were to evaluate the *gains of personalization* for *intersectional* groups. Avoidable because a fair use violation implies that a group could receive better predictions from a generic model or a personalized model for another group. Thus, one could resolve a fair use violation by assigning predictions from this better-performing model.

[*]Equal Supervision [1]MIT [2]UCSD. Correspondence to: Vinith M. Suriyakumar <vinithms@mit.edu>, Berk Ustun <berk@ucsd.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Group	Size	Training Error		Gain
g	n_g	$R(h_0)$	$R_g(h)$	$_g(h, h_0)$
female, <30	48	38.1%	26.8%	11.3%
male, <30	49	23.9%	26.7%	-2.8%
female, 30 to 60	304	30.3%	29.1%	1.2%
male, 30 to 60	447	15.4%	15.2%	0.2%
female, 60+	123	19.3%	21.9%	-2.6%
male, 60+	181	11.0%	8.2%	2.8%
Total	1,152	20.4%	19.4%	1.0%

Figure 1: Personalization can reduce performance at the group level. We train a personalized model h_g and generic model h_0 with logistic regression, personalizing h_g with a one-hot encoding of `sex` and `age_group` to screen for obstructive sleep apnea (see the `apnea` dataset in Section 4). As shown, personalization reduces training error at a population level from 20.4% to 19.4% yet *increases* error for two groups: (`female, 60+`) and (`male, <30`). These effects are also present on test data.

Group	Data		Personalized		Generic		Gain
g	n_g^+	n_g	h	$R_g(h)$	h_0	$R_g(h_0)$	$_g(h, h_0)$
female, young	0	24	+	24		0	24
male, young	25	0	+	0		25	25
female, old	25	0	+	0		25	25
male, old	0	27		0		0	0
Total	50	51		24		50	26

Figure 2: Stylized classification task where the best personalized model reduces performance for a group due to model misspecification. There are $n^+ = 50$ positive and $n^- = 51$ negative examples. We train a personalized linear classifier with a one-hot encoding of $\boldsymbol{g}$ 2 fmale, femaleg fold, youngg, and evaluate the gains to personalization with respect to a generic model h_0 without group attributes. Personalization reduces overall error from 50 to 24. However, not all groups gain from personalization: (`young, female`) receives less accurate predictions and (`old, male`) receives no gain.

Our goal in this work is to expose this effect and lay the foundations to address it. Our main contributions include:

1. We present formal conditions to ensure the fair use of group attributes in prediction task. Our conditions reflect collective preference guarantees that are necessary for truthful self-reporting, and that can be tested by training one additional model.
2. We characterize how empirical risk minimization with group attributes can violate fair use. Our analysis includes counterexamples and sufficient conditions that illustrate failure modes in model development and inform interventions to mitigate their effects.
3. We conduct a comprehensive empirical study on fair use violations in clinical prediction tasks, showing their prevalence across major model classes, personalization techniques, and prediction tasks.
4. We present a case study on personalization for a model trained to predict mortality for patients with acute kidney injury. Our study shows how a fair use audit can safeguard against incorrect "race correction" in clinical prediction models, and presents targeted interventions that reduce harm.

Related Work Personalization encompasses a broad range of techniques that use personal data. Here, we use it to describe techniques that target *groups* rather than *individuals* – i.e., "categorization" rather than "individualization" as per the taxonomy of Fan & Poole [33]. Modern approaches to personalization with group attributes use them to improve population-level performance by, e.g., automatically including higher-order interaction effects [12, 59, 86] or recursively partitioning data [30, 13, 11, 10]. In practice, few works measure the gains of personalization, and those that do measure the gains at a population level rather over the groups who provide personal data [see e.g., 48, 79].

We introduce conditions for models that use group attributes to assign more accurate predictions. Much work in algorithmic fairness discusses the need for models to account for group membership [95, 29, 22, 56, 60, 91], observing that it is otherwise impossible for a model to perform equally well for all groups [42, 94, 96, 34, 2, 67, 19]. These results highlight the need to account for group attributes in personalization. Nevertheless, methods to equalize performance are ill-suited for personalization because they can equalize performance by assigning less accurate predictions to groups for whom the model performs well, rather than by assigning more accurate predictions to groups for whom the model performs poorly [60, 46, 73, 61, 62].

We build on the work of Ustun et al. [84], who propose the preference guarantees of rationality and envy-freeness [see also 95]. Their work develops a recursive decoupling algorithm that uses preference guarantees to guide decoupling [c.f., 29, 3]. In contrast, we study these guarantees as standalone conditions to ensure personalization without harm. Our work complements an emerging stream on fair use in prediction models [see e.g., 69, 47]. More broadly, it highlights a practical application for preference-based notions of fairness [95, 84, 55, 87, 27], and represents a new use case to evaluate model performance across intersectional groups [c.f., 52, 45, 39, 90].

2. Fair Use Conditions

We present formal conditions for the fair use of group attributes in prediction tasks.

Preliminaries We start with a dataset $(\boldsymbol{x}_i, y_i, \boldsymbol{g}_i)_{i=1}^n$, where example i consists of a feature vector $\boldsymbol{x}_i = [x_{i,1}, \ldots, x_{i,d}]$ 2 R^d, a label y_i 2 Y, and k categorical *group attributes* $\boldsymbol{g}_i = [g_{i,1}, \ldots, g_{i,k}]$ 2 $\mathsf{G}_1 \quad \ldots \quad \mathsf{G}_k = \mathsf{G}$. We refer to $\boldsymbol{g}_i$ as the *group membership* of person i. For example, a female over 60 would have $\boldsymbol{g}_i$ = [`female, age 60`]. We let $n_{\boldsymbol{g}} := \mathit{jfi}\ \mathit{jg}_i = \mathit{ggj}$ denote the size of group $\boldsymbol{g}$,

and $m := |G|$ denote the number of intersectional groups.

We use the data to train a *personalized* model with group attributes $h : X \times G \to Y$; and a *generic* model that does not $h_0 : X \to Y$. We train all models via ERM with a loss function $\ell : Y \times Y \to \mathbb{R}_+$, denoting the empirical and true risks as $\hat{R}(h)$ and $R(h)$, respectively. We assume that the personalized and generic models represent the empirical risk minimizers on datasets with group attributes $(\boldsymbol{x}_i, y_i, \boldsymbol{g}_i)_{i=1}^n$ and without them $(\boldsymbol{x}_i, y_i)_{i=1}^n$:

$$h \in \operatorname*{argmin}_{h \in H} \hat{R}(h) \qquad h_0 \in \operatorname*{argmin}_{h \in H_0} \hat{R}(h)$$

Here, H and H_0 denote the class of personalized models and generic models respectively.

We measure the *gains of personalization* for a personalized model h for each group. As part of this evaluation, we measure how the model will perform for group $\boldsymbol{g}$ when they are assigned the predictions personalized for a different group – i.e., the predictions that they could receive by "misreporting" their group membership as $\boldsymbol{g}'$. Given a personalized model h, we denote its *empirical risk* and *true risk* for group $\boldsymbol{g}$ when they report $\boldsymbol{g}'$ as:

$$\hat{R}_{\boldsymbol{g}}(h_{\boldsymbol{g}'}) := \frac{1}{n_{\boldsymbol{g}}} \sum_{i:\boldsymbol{g}_i = \boldsymbol{g}} \ell\left(h(\boldsymbol{x}_i, \boldsymbol{g}'), y_i\right)$$

$$R_{\boldsymbol{g}}(h_{\boldsymbol{g}'}) := \mathbb{E}\left[\ell\left(h(\boldsymbol{x}, \boldsymbol{g}'), y\right) \mid G = \boldsymbol{g}\right].$$

We use $h_{\boldsymbol{g}'} := h(\,, \boldsymbol{g}')$ to denote a personalized model where group attributes are fixed to $\boldsymbol{g}'$.

We assume that each group prefers models that assign more accurate predictions as measured in terms of true risk, and evaluate the preferences of group $\boldsymbol{g}$ between h and h' using the *gain* measure: $\Delta_{\boldsymbol{g}}(h, h') := R_{\boldsymbol{g}}(h') - R_{\boldsymbol{g}}(h)$. This is a plausible assumption in settings where models are used to assign personalized predictions. It does not hold in settings where individuals may prefer models that [see e.g., polar prediction tasks 70].

As Collective Preference Guarantees In Definition 1, we characterize the fair use of a group attribute in terms of collective preference guarantees.

Definition 1. *A personalized model $h : X \times G \to Y$ guarantees the fair use of group attributes G if it obeys:*

$$\Delta_{\boldsymbol{g}}(h_{\boldsymbol{g}}, h_0) \geq 0 \quad \text{for all groups } \boldsymbol{g} \in G, \tag{1}$$

$$\Delta_{\boldsymbol{g}}(h_{\boldsymbol{g}}, h_{\boldsymbol{g}'}) \geq 0 \quad \text{for all groups } \boldsymbol{g}, \boldsymbol{g}' \in G \tag{2}$$

These conditions are collective in that performance is measured over individuals in a group. Here, condition (1) ensures *rationality* for group $\boldsymbol{g}$ – i.e., that a majority of group $\boldsymbol{g}$ prefers a personalized model $h_{\boldsymbol{g}}$ to a generic model h_0. Condition (2) ensures *envy-freeness* for group $\boldsymbol{g}$ – i.e., that majority of group $\boldsymbol{g}$ prefers their personalized predictions to the personalized predictions for any other group. These conditions reflect minimal expectations of groups from a personalized model.

These conditions can be adapted to different supervised learning tasks by choosing a suitable risk metric. Since fair use conditions reflect the expected gains from personalization, a "suitable" metric should represent an exact measure of model performance rather than a surrogate measure optimized for training. In classification tasks where we want accurate predictions, this would be the error rate. In tasks where we want reliable risk estimates, it would be the expected calibration error [66].

As Prerequisites for Truthful Self-Reporting In copyright law, fair use conditions characterize when we can use copyrighted material without permission from copyright owners [93, 68]. In this setting, fair use conditions characterize when we can use personal data without asking permission from the owners of that data. In particular, fair use conditions are necessary for "truthful self-reporting" [see e.g., 77, 50, 40].

Proposition 2. *Consider a prediction model where each person reports their group membership to a personalized model $h : X \times G \to Y$ in deployment. Denote the reported group membership of person i as:*

$$\begin{array}{lcll} \boldsymbol{r}_i & = & \boldsymbol{g}_i & , \quad i \text{ reports truthfully} \\ \boldsymbol{r}_i & \in & G \setminus \{\boldsymbol{g}_i\} & , \quad i \text{ misreports} \\ \boldsymbol{r}_i & = & ? & , \quad i \text{ withholds} \end{array}$$

If a personalized model guarantees the fair use of G then each person would choose to report truthfully as this strategy would maximize their expected performance:

$$\boldsymbol{g}_i \in \operatorname*{argmin}_{\boldsymbol{r}_i \in G \cup \{?\}} \mathbb{E}\left[\ell\left(h(\boldsymbol{x}, \boldsymbol{r}_i), y_i\right) \mid G = \boldsymbol{g}_i\right].$$

Truthful self-reporting incentives reflect basic principles regarding *consent* in data privacy. In effect, a personalized model that violates fair use uses group membership in a way that is coercive. If groups were allowed to report personal information to a personalized model at prediction time, group who experience a fair use violation would not report group membership voluntarily or truthfully, choosing to withhold or misreport instead. If a model obeys fair use, individuals may still withhold group membership because the gain is insufficient. In light of this, fair use conditions should be viewed as minimal requirements to flag harm rather than a "rubber stamp" for consent.

Use Cases Fair use conditions should hold in prediction tasks where individuals are entitled to control or report their own data. In such tasks, we should ensure fair use conditions for group attributes that encode:

Immutable Attributes: Group attributes often encode characteristics like `sex` [see e.g., 71]. In this setting, fair use conditions ensure that individuals will not receive unnecessarily inaccurate predictions due to immutable characteristics.

Sensitive Information: Models that use attributes like `hiv_status` should guarantee a tailored gain in performance for the sensitive group, `hiv_status = +`. Otherwise, they require individuals to disclose information that may be harmful when leaked [see e.g., 7].

Self-Reported Information: Models are often personalized using information that individuals report directly – see e.g., self-reported screening tests for mental illnesses [54, 83]. These models should obey fair use conditions to incentivize truthful self-reporting as per Proposition 2.

Costly Information: Group attributes can encode characteristics that must be collected at test time – e.g., an attribute like `pcr_test` whose value requires a medical test. Models that ensure fair use with respect to `pcr_test` guarantee that patients with a specific outcome will not receive a less accurate prediction after taking a test.

Testing for Fair Use We can evaluate fair use conditions by training a generic model in addition to a personalized model. Given a personalized model and its generic counterpart, we can check the conditions in Definition 1 on a sample by computing the relevant performance gains. This procedure will return point estimates that should be paired with a measure of uncertainty to guide model development. In some tasks, a significant fair use violation may warrant a new model. In others, we may wish to ensure a significant gain to use a group attribute in the first place.

In practice, we check for a rationality violation using a one-sided hypothesis test of the form:

$$H_0 : R_{\boldsymbol{g}}(h_0) \quad R_{\boldsymbol{g}}(h_{\boldsymbol{g}}) \quad 0$$
$$H_A : R_{\boldsymbol{g}}(h_0) \quad R_{\boldsymbol{g}}(h_{\boldsymbol{g}}) > 0$$

Here, the null hypothesis H_0 assumes that group $\boldsymbol{g}$ prefers $h_{\boldsymbol{g}}$ to h_0. Thus, we would reject H_0 when there is enough evidence to support a rationality violation for $\boldsymbol{g}$ on held-out data. We can test all conditions in Definition 1 by repeating this test for all m groups to check rationality, and repeating analogous tests for all $m(m \quad 1)$ pairs of groups to check envy-freeness. In general, one can test these hypotheses for any performance metric using a bootstrap hypothesis test [25], and control the false discovery rate using a Bonferroni correction [28]. In practice, one should draw on more powerful tests when working with salient performance metrics [e.g., the McNemar test for accuracy 26].

3. Failure Modes and Guarantees

Practitioners naturally presume that training a model with group attributes will provide a uniform performance gain for all groups. Here, we characterize how empirical risk minimization may fail to improve performance at a group level through counterexamples and sufficient conditions. We include additional examples and proofs in Appendix B.

3.1. Failure Modes

We characterize common practices that lead personalization to reduce performance at a group level. We present examples related to model misspecification and model selection as they motivate interventions for model development in Section 4. We include examples related to generalization, distributional shifts, and training with a surrogate loss function in Appendix B.1.

Misspecification We start with misspecification – i.e., when a model that cannot capture the influence of group membership in a conditional data distribution. A common form of misspecification occurs when we personalize simple models with a one-hot encoding [85]. In such cases, models exhibit fair use violations on data distributions that exhibit *intersectionality*. Consider, for example, a logistic regression model with a one-hot encoding that assigns higher risk to patients who are `young`, and to patients who are `female`. This model would exhibit a fair use violation for patients who are `young` *and* `female` if their true risk were lower due to an interaction effect among group attributes (see Fig. 2).

Misspecification can also stem from group-specific interaction effects – e.g., tasks where group attributes act as mediators or moderators [see e.g., 9]. In Example 1, we show an example that exhibits the hallmarks of personalization: a generic model performs poorly on "heterogeneous" groups A and C, and a personalized model that targets these groups improves performance at a population-level.

Example 1. *Consider a 2D classification task with groups* $\mathsf{G} = \mathsf{f}A, B, C\mathsf{g}$ *with 1 positive and 1 negative example in which a Bayes optimal classifier* $h : \mathsf{X} \quad \mathsf{G} \; ! \; \mathsf{Y}$ *should assign a personalized intercept to each group* and *a personalized slope to group* B*:*

$$h(\boldsymbol{x}, \boldsymbol{g}) = \begin{cases} \text{sign}\left(t_A + \boldsymbol{w}^\top \boldsymbol{x}\right) & \text{if } \boldsymbol{g} = A \\ \text{sign}\left(t_B + \boldsymbol{w}_B^\top \boldsymbol{x}\right) & \text{if } \boldsymbol{g} = B \\ \text{sign}\left(t_C + \boldsymbol{w}^\top \boldsymbol{x}\right) & \text{if } \boldsymbol{g} = C \end{cases}$$

Here, ERM with a standard one-hot encoding of G *would return a personalized model that assigns a personalized intercept for each group, but the same slope to all three groups:*

$$h(\boldsymbol{x}, \boldsymbol{g}) = \begin{cases} \text{sign}\left(t_A + \boldsymbol{w}^\top \boldsymbol{x}\right) & \text{if } \boldsymbol{g} = A \\ \text{sign}\left(t_B + \boldsymbol{w}^\top \boldsymbol{x}\right) & \text{if } \boldsymbol{g} = B \\ \text{sign}\left(t_C + \boldsymbol{w}^\top \boldsymbol{x}\right) & \text{if } \boldsymbol{g} = C \end{cases}$$

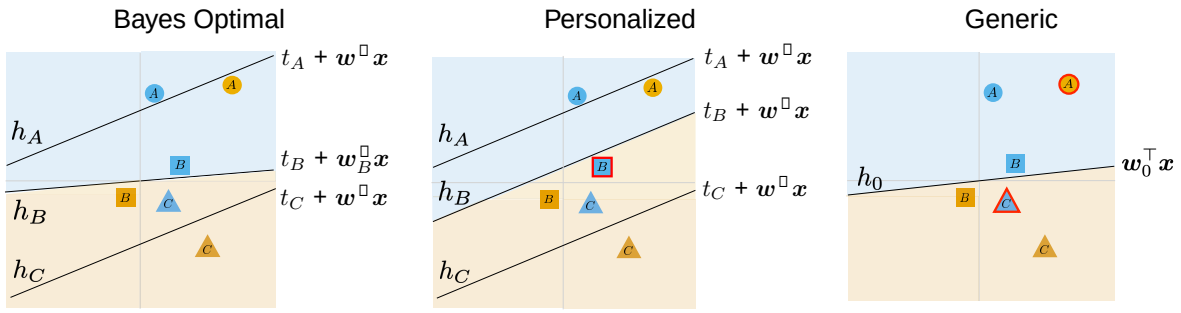

Figure 3: ERM returns a misspecified personalized model that assigns a personalized intercept for each group but the same slope for all groups. It does not capture the personalized slope needed to accurately model group B. The model improves overall performance by assigning more accurate predictions to groups A and C. However, it performs *worse* for group B.

The model would improve overall performance by assigning more accurate predictions to groups A and C. However, it would perform worse *for group B (Figure 3).*

Resolving violations from model misspecification is difficult since it requires interventions that can resolve them for all groups. In practice, one could fit models from a class that is rich enough to capture these effects, or train a separate model for each group. Both approaches are challenging when working with multiple groups. The first requires that we either specify interactions for each group and fit these terms correctly. The second requires that we train models using a limited amount of data for each group.

Model Selection Model development often involves choosing a model from candidate models – e.g., when setting a regularization penalty to avoid overfitting or to induce sparsity. Common criteria for model selection guide these decisions on the basis of population-level performance [e.g., mean K-CV test error 5]. As shown in Example 2, the resulting model may improve performance for one group while reducing performance for another group in tasks with heterogeneous data distributions.

Example 2. *Consider a classification task where a personalized model must use either x_1 2 f0, 1g or x_2 2 f0, 1g. We are given 60 examples from group A and 90 examples from group B. We train a personalized model with a one-hot encoding of G = fA, Bg choosing between x_1 or x_2 to minimize the overall error rate.*

				Generic		Personalized with x_1			Personalized with x_2		
Group	(x_1, x_2)	n^+	n	h_0	$R(h_0)$	h_1	$R(h_1)$		h_2	$R(h_2)$	
A	$(0,0)$	10	0	+	0	+	0	0	+	10	10
A	$(0,1)$	10	0	+	0	+	0	0	+	0	0
A	$(1,0)$	0	20	+	20		0	20		0	20
A	$(1,1)$	20	0	+	0		20	20	+	0	0
B	$(0,0)$	5	0	+	0		5	5	+	0	0
B	$(0,1)$	0	20	+	20		0	20	+	20	0
B	$(1,0)$	20	0	+	0	+	0	0	+	0	0
B	$(1,1)$	30	0	+	0	+	0	0	+	0	0
Total		95	40		40		25	15		30	10
Group A		40	20		20		20	0		10	10
Group B		55	20		20		5	15		20	0

The generic model h_0 is the same whether it uses x_1 or x_2. However, the personalized model violates fair use for either group A when it uses x_1, and violates fair use for group B when it uses x_2. In this case, ERM returns the personalized model that benefits the majority group (A).

Example 2 could arise, for example, when developing a clinical prediction model using features that encode the outcome of competing diagnostics. More broadly, Example 2 highlights how fair use violations may be unavoidable when we must assign predictions with a single model – as the task shows that models trained with x_1 and x_2 would lead to fair use violations on A or B respectively.

3.2. Sufficient Conditions

We present sufficient conditions for ERM with group attributes to output a model that obeys fair use in training (Proposition 3) and testing (Proposition 4).

Proposition 3. *Consider training a personalized model by ERM h 2 argmin$_{h2H}$ $\hat{R}(h)$, and evaluating its gains to personalization with respect to a generic model h_0 2 argmin$_{h2H_0}$ $\hat{R}(h)$ where H_0 H. The personalized model h obeys fair use in terms of empirical risk so long as the model achieves the same risk as a model that specifically targets the group. That is:*

$$\hat{R}_{\boldsymbol{g}}(h) = \hat{R}_{\boldsymbol{g}}(h_{\boldsymbol{g}}) \text{ for all groups } \boldsymbol{g} \in \mathsf{G}.$$

Proposition 3 holds for settings where we fit personalized models from a class H that extends the generic model class H_0 (see Definition 5). This requirement implies that we should fit personalized models from model classes that are rich enough to target each intersectional group. When we personalize a linear classifier via "score correction" [85], we should include a correction term for each group. Otherwise, we may violate fair use due to model misspecification when using a one-hot encoding as in Fig. 2. Likewise, if we personalize a model with interaction terms, we should include an interaction for each group. More broadly, the conditions

in Proposition 3 are met when, for example, we use the data from each group to train a model for each group. Given that these are sufficient conditions, it is still possible to achieve fair use even when they don't hold.

Proposition 4. *Consider a personalized model* $h : X \; G \; ! \; Y$ *that ensures rationality and envy-freeness for group* $\boldsymbol{g}$ *in terms of empirical risk. Denote the empirical gains in rationality and envy-freeness for group* $\boldsymbol{g}$ *as:*

$$\hat{\epsilon}_{\boldsymbol{g}} := \hat{} _{\boldsymbol{g}}(h_{\boldsymbol{g}}, h_0), \qquad \hat{\gamma}_{\boldsymbol{g}} := \min_{\boldsymbol{g}' 2 G / f \boldsymbol{g} \boldsymbol{g}} \hat{} _{\boldsymbol{g}}(h_{\boldsymbol{g}}, h_{\boldsymbol{g}'}).$$

If $\hat{\epsilon}_{\boldsymbol{g}} > 0$*, then rationality for group* $\boldsymbol{g}$ *generalizes with probability at least* $1 \quad \delta$ *as long as:*

$$n_{\boldsymbol{g}} \quad \frac{4D \log\left(\frac{2n_{\boldsymbol{g}}}{D} + 1\right) + \log\left(\frac{8}{\delta}\right)}{\hat{\epsilon}^2_{\boldsymbol{g}}}$$

If $\hat{\gamma}_{\boldsymbol{g}} > 0$*, then envy-freeness for group* $\boldsymbol{g}$ *generalizes with probability at least* $1 \quad \delta$ *as long as:*

$$n_{\boldsymbol{g}} \quad \frac{4D \log\left(\frac{2n_{\boldsymbol{g}}}{D} + 1\right) + \log\left(\frac{8m}{\delta}\right)}{\hat{\gamma}^2_{\boldsymbol{g}}}$$

Proposition 4 characterizes the sample complexity of generalization for personalized models that satisfy fair use conditions on training data. The bounds apply to a general class of personalized models, and can be strengthened by assuming a finite hypothesis class [e.g, in 84], or by accounting for distributional differences between groups [e.g., 92]. The result holds in tasks where personalization leads to strictly positive gains with respect to rationality and envy-freeness on the training data, which is not guaranteed in practice and must be checked in practice.

4. Empirical Study

In this section, we present an empirical study of fair use in clinical prediction models – i.e. a class of models that routinely include group attributes and where fair use violations inflict harm. Our goals are to discuss the prevalence of fair use violations, the impact of standard personalization techniques, and the potential to resolve them through interventions in model development. We provide code to reproduce these results at `https://github.com/ustunb/fairuse` and include additional results in Appendix D.

4.1. Setup

We work with 6 datasets for clinical prediction tasks listed in Table 1 and Appendix C. We minimally process each dataset to impute the values of missing points (using mean value imputation), and repair class imbalances across intersectional groups (to eliminate "trivial" fair use violations that occur due to class imbalance). We split each dataset into a training sample (80%) to fit models, and a test sample (20%) to evaluate the gains of personalization.

We train 9 personalized models for each dataset. Each model belongs to one of 3 model classes: *logistic regression* (LR), *random forests* (RF), and *neural nets* (NN), and encodes group attributes using one of 3 personalization techniques:

One-Hot Encoding (1Hot): We train a model with features that include dummy variables for each group attribute.

Intersectional Encoding (All): We train with features that include dummy variables for each intersectional group.

Decoupling (DCP): We train a separate model for each intersectional group using only data from this group $\boldsymbol{g}_i = \boldsymbol{g}$.

These three techniques reflect the increasingly complex approaches available to practitioners to account for group membership in a prediction model as measured in terms of the interactions between group attributes and other features: 1Hot reflect no interactions; All reflect interactions between group attributes; and DCP reflects all possible interactions between group attributes and features.

We evaluate the gains of personalization for each model in terms of three performance metrics, reflecting common metrics that are encountered in different tasks: (1) *error rate*, which reflects the accuracy of yes-or-no predictions, e.g., for a diagnostic test [32]; (2) *area under ROC curve* (AUC), which measures accuracy in ranking, e.g., for triage [e.g., 97]; (3) *expected calibration error* (ECE), which measures the reliability of risk predictions for a risk score [14, 81].

4.2. Results

We summarize our results for logistic regression in Table 1 and for neural networks and random forests in Appendix D.

On the Prevalence of Fair Use Violations Our results show that we train models that improve population level performance across prediction tasks in terms of training loss (guaranteed), training performance (expected), and test performance (expected). Yet personalized models that improve performance at a population level can also reduce performance for specific groups. These violations arise across datasets, personalization techniques, and model classes.

We consider the standard configuration used to develop clinical prediction models – i.e., a logistic regression model with a one-hot encoding of group attributes (LR+1Hot). In this case, we find that at least one group experiences a statistically significant fair use violation in terms of error on 4/6 datasets (5/6 for AUC and ECE). On `saps`, for example, LR + 1Hot exhibits a statistically significant gain from personalization for patients over 30 who are HIV negative. Conversely, in `cardio_eicu` when training LR+All we de-

Dataset	Metrics	Test Error			Test AUC			Test ECE		
		1Hot	All	DCP	1Hot	All	DCP	1Hot	All	DCP
apnea	Personalized	34.2%	33.8%	26.2%	0.750	0.750	0.803	7.5%	5.5%	7.2%
n = 1152, d = 26	Gain	-1.0%	-0.7%	7.0%	0.001	0.000	0.053	-1.5%	0.6%	-1.1%
G = $f_{\texttt{age},\texttt{sex}}g$	Best/Worst Gain	0.0% / -9.6%	1.7% / -7.8%	21.7% / -7.8%	0.002 / -0.001	0.001 / -0.010	0.119 / -0.005	0.9% / -8.6%	0.8% / -4.6%	1.7% / -6.6%
m = 6	Rat. Gains/Viols	6/4	5/3	2/2	1/4	1/2	4/4	3/3	3/3	3/3
Ustun et al. [82]	EF Gains/Viols	0/0	1/0	5/4	0/6	0/6	0/0	3/0	0/0	4/4
cardio_eicu	Personalized	29.1%	29.1%	29.5%	0.768	0.767	0.762	4.4%	4.6%	8.9%
n = 1341, d = 49	Gain	-0.4%	-0.4%	-0.9%	0.000	-0.001	-0.007	0.4%	0.2%	-4.1%
G = $f_{\texttt{age},\texttt{sex}}g$	Best/Worst Gain	0.0% / -3.1%	0.2% / -3.1%	13.0% / -8.6%	0.002 / -0.001	0.001 / -0.001	0.096 / -0.104	1.6% / -1.5%	0.9% / -0.2%	-0.9% / -6.2%
m = 4	Rat. Gains/Viols	4/2	4/2	2/2	2/3	2/3	1/1	1/1	0/0	4/4
Pollard et al. [75]	EF Gains/Viols	1/0	1/0	3/3	0/4	0/4	1/1	0/0	0/0	1/1
cardio_mimic	Personalized	23.3%	23.4%	21.4%	0.854	0.854	0.870	2.1%	2.3%	2.3%
n = 5289, d = 49	Gain	0.3%	0.3%	2.2%	0.001	0.001	0.017	-0.4%	-0.5%	-0.6%
G = $f_{\texttt{age},\texttt{sex}}g$	Best/Worst Gain	0.9% / -0.1%	0.9% / -0.1%	7.9% / -0.0%	0.001 / -0.000	0.001 / -0.000	0.053 / 0.006	0.5% / 0.4%	0.6% / -0.2%	0.8% / -2.3%
m = 4	Rat. Gains/Viols	1/0	1/0	0/0	2/2	2/2	4/4	0/0	0/0	2/2
Johnson et al. [49]	EF Gains/Viols	1/0	1/0	4/4	0/4	0/4	0/0	1/1	0/0	3/3
heart	Personalized	19.7%	19.7%	15.8%	0.870	0.846	0.817	8.4%	17.8%	17.5%
n = 181, d = 26	Gain	-1.3%	-1.3%	2.6%	-0.007	-0.030	-0.060	2.8%	-6.6%	-6.3%
G = $f_{\texttt{sex},\texttt{age}}g$	Best/Worst Gain	0.0% / -6.8%	0.1% / -9.9%	10.6% / -8.4%	0.008 / -0.036	0.017 / -0.055	0.039 / -0.190	3.7% / -0.5%	-1.2% / -3.2%	10.1% / -4.6%
m = 4	Rat. Gains/Viols	4/1	4/1	2/1	1/3	0/3	1/1	1/1	3/3	1/1
Detrano et al. [24]	EF Gains/Viols	3/0	3/0	2/2	0/4	0/4	2/2	1/1	0/0	2/2
mortality	Personalized	23.6%	23.4%	20.2%	0.848	0.848	0.880	2.0%	2.1%	2.5%
n = 25366, d = 468	Gain	-0.2%	0.0%	3.2%	0.000	0.001	0.033	0.2%	0.1%	-0.3%
G = $f_{\texttt{age},\texttt{sex}}g$	Best/Worst Gain	0.8% / -2.5%	2.1% / -0.4%	20.8% / -0.6%	0.004 / -0.001	0.004 / -0.000	0.114 / 0.011	1.6% / 0.0%	2.9% / -0.5%	11.2% / -2.5%
m = 6	Rat. Gains/Viols	4/4	2/2	1/1	3/3	4/4	6/6	0/0	0/0	3/3
Johnson et al. [49]	EF Gains/Viols	2/0	2/1	6/6	0/6	0/6	0/0	4/0	3/3	5/5
saps	Personalized	18.9%	18.9%	18.5%	0.890	0.890	0.888	1.6%	1.6%	1.9%
n = 7797, d = 36	Gain	0.0%	0.0%	0.4%	0.001	0.001	-0.001	0.0%	0.0%	-0.3%
G = $f_{\texttt{hiv},\texttt{age}}g$	Best/Worst Gain	16.4% / -12.2%	0.7% / -12.2%	3.5% / -23.3%	0.013 / -0.000	0.013 / -0.000	0.017 / -0.246	2.9% / -2.1%	2.5% / -1.3%	9.4% / -19.1%
m = 4	Rat. Gains/Viols	2/2	3/2	2/1	1/3	1/3	2/2	2/2	2/2	2/2
Allyn et al. [4]	EF Gains/Viols	2/1	2/2	2/2	0/4	0/4	1/2	2/2	2/2	3/3

Table 1: Performance of personalized logistic regression models on all datasets. We show the gains of personalization in terms of test AUC, ECE, and error. We report: model performance at the population level, the overall gain of personalization, the range of gains over m intersectional groups, and the number of rationality and envy-freeness gains/violations (evaluated using a bootstrap hypothesis test (Section 2) at a 10% significance level). We include results for other model classes in Appendix D.

tect a fair use violation for old females (see e.g., 4/2 Rat. Gains/Viols. respectively for test error in Table 1).

On the Robustness of Personalization Techniques Our results show there is no one personalization technique that can avoid fair use violations, as demonstrated by the fact that the personalization technique that minimizes fair use violations varies across datasets, model classes, and prediction tasks. In Table 1, for example, we find that the best technique to personalize a logistic regression model for `cardio_eicu` is to use an intersectional encoding, but to train decoupled models for `mortality`. These strategies change across model classes – as the ideal strategies for neural networks are decoupling and intersectional encoding, respectively `cardio_eicu` and `mortality` (see Appendix D). Even configurations that exhibit few violations across datasets may fail critically across groups. For example, LR+DCP for `saps` leads to a 10% increase in error for `HIV+ & >30`. Overall, these results suggest that the most reliable way to avoid a fair use violation is to check.

On Detecting Violations Our results underscore the need for reliable procedures to spot fair use violations or claim gains from personalization. We can often find reliable instances of benefit or harm but we sometimes are unable to do so. An actionable finding from evaluating the gains of personalization is a group does not experience a meaningful gain nor harm due to personalization. We note a number of cases across datasets, personalization techniques, and model classes where we note no meaningful gain or harm. Often times this is because the effect size is small or the group sample sizes are too small.

In such cases where we are unable to detect any impact from personalizationo, one may wish to intervene to avoid soliciting unnecessary data. For example, when group attributes encode information that is sensitive or must be collected at prediction time (e.g., `HIV`), we may prefer to avoid soliciting information unless it is demonstrably useful for prediction.

On Resolving Violations Our results show that routine decisions in model development can induce considerable differences in group-level performance. This suggests that we can reduce fair use violations through "interventions" in model development. We studied the effectiveness of this approach through an ablation study where we repeated our experiments with interventions that address failure modes in Section 3, namely: using an intersectional one-hot encoding, decoupled training, and equalizing sample sizes.

Our results show that interventions can often reduce fair use violations. For example, we can eliminate all fair use violations for `cardio_mimic` in our standard configuration

by decoupled training. However, there is no "silver bullet" intervention that resolves fair use violations across all datasets and model classes. In general, the best intervention varies across model classes and datasets. In some cases, the best intervention may fail to resolve all fair use violations as resolving a violation for one group may induce a violation on another group. In `cardio_eicu`, for example, a logistic regression model with a one-hot encoding will exhibit a violation on old males. Switching an intersectional encoding will fix this violation but introduce another for old females.

5. Mortality in Acute Kidney Injury

In this section, we audit fair use for mortality prediction model for patients with acute kidney injury. Our results demonstrate how evaluating the gains to personalization can inform model development and improve simple interventions to mitigate harm.

5.1. Setup

We consider a mortality prediction task for critically-ill patients who receive continuous renal replacement therapy. The data contains $n = 2,066$ patients from MIMIC III and IV [49] and includes $d = 78$ features related to their health, lab tests, length of stay, and potential for organ failure. Here, $y_i = +1$ if patient i dies in the ICU and $\Pr(y_i = +1) = 51.1\%$. We train personalized models using the setup in Section 4.1, and evaluate fair use for groups defined by the attributes $\texttt{sex} \in \{\texttt{male}, \texttt{female}\}$ and $\texttt{race} \in \{\texttt{white}, \texttt{black}, \texttt{other}\}$.

5.2. Results

We show performance for the personalized logistic regression model with a one-hot encoding in Table 2, and present results for other configurations in Appendix D. Our findings show that personalization yields uneven gains at a group level, leading to fair use violations across prediction tasks and model classes. In this case, the gains in error across range from -5.2% to 6.8%, and two groups experience statistically significant fair use violations: (`male, black`) and (`male, other`).

On the Use of Race Clinical prediction models include group attributes whenever there is a plausible biological relationship between group membership and the outcome of interest or social determinants of health. These norms have led to development of models that use race an ethnicity [31, 88, 36, 44, 58, 65, 37, 51]. Recently, Vyas et al. [88] discuss how such models can inflict harm and urge physicians to check if "race correction is based on robust [statistical] evidence." Our results highlight how a fair use audit can yield evidence that serves to guide the use "race correction" in such cases. Here, checking rationality shows that a race-specific model can reduce performance for specific groups – e.g., (`male, black`) and (`male, other`). Checking envy-freeness reveals that groups expect better performance by misreporting group membership – e.g., (`male,other`) would experience a 5.6% gain in test error by reporting any other race.

In tasks where race improves performance, race may act as a proxy for broader social determinants of health. Thus, a model that includes race may act as a "smoke screen" in that it attributes differences in health outcomes to an immutable factor, and perpetuates inaction on the root causes of health disparities [72]. Given these uncertainties, we advocate that race should only be included in clinical model when there is evidence of gain. Regardless of its use in prediction, collecting information about race and ethnicity is necessary to measure model performance across these groups. In such cases, one should be careful to disclose the purposes of data collection – stating that it will be used to evaluate performance but not to assign personalized predictions. In tasks where race does not improve model performance, models may exhibit differences in performance across racial groups – as data may encode proxies of race in redacted notes [1], or even band-pass filtered images [38].

Interventions We build on our results to discuss interventions that can resolve fair use violations and broaden the gains to personalization by using multiple models. These are simple interventions that have the benefit of being broadly applicable – i.e., we can use them to mitigate harm from fair use violations for any prediction task and model class.

Assigning a Generic Model. We assign groups who experience a fair use violation the predictions from a generic model h_0. This intervention will resolve all fair use violations in a way that strictly improves performance. In this case, it resolves all rationality violations (2/3/2 in terms of error/AUC/ECE respectively). We also observe a potential to reduce data usage in deployment: seeing how both (`male, black`) and (`male, other`) experience a fair use violation in terms of error, we could soliciting race for all `male` patients and reduce test error by 1% (as the loss in accuracy for (`white, male`) are offset by the gain in accuracy for (`male,black`) and (`male, other`).

Assigning a Decoupled Model. We assign groups who experience a fair use violation predictions from the best of a generic model, personalized model, or a *decoupled model* h_g^{dcp} – i.e., a model trained using only data from their group. While this approach may not resolve fair use violations, it can produce surprisingly large gains as decoupling effectively personalizes the entire model development pipeline. Our results in Table 2 show the potential gains of this intervention across all performance metrics. Focusing on error, we see that one can: (1) eliminate fair use violations

Group	Test Error		Intervention		Test AUC		Interventions		Test ECE		Interventions	
g	$R_g(h_g)$	g	Assign h_0	Assign h_g^{dcp}	$R_g(h_g)$	g	Assign h_0	Assign h_g^{dcp}	$R_g(h_g)$	g	Assign h_0	Assign h_g^{dcp}
female, black	55.5%	3.5%	3.5%	33.1%	0.443	0.010	0.010	0.315	32.2%	2.1%	2.1%	11.9%
female, white	21.9%	2.0%	2.0%	2.0%	0.845	0.004	0.004	0.057	10.1%	2.0%	2.0%	0.03%
female, other	20.4%	6.6%	6.6%	9.1%	0.861	-0.003	0.000	0.038	14.7%	1.8%	1.8%	5.3%
male, black	29.4%	-2.7%	0.0%	15.6%	0.799	0.020	0.020	0.096	18.1%	-0.0%	0.0%	6.4%
male, white	21.9%	8.1%	8.1%	3.7%	0.767	0.006	0.006	0.104	10.6%	-1.4%	0.0%	1.4%
male, other	25.3%	-1.9%	0.0%	1.3%	0.835	-0.003	0.000	0.017	13.5%	0.0%	0.0	0.0%
Total	27.1%	1.4%	-	-	0.803	0.010	-	-	4.7%	0.2%	-	-

Table 2: Fair use evaluation of a personalized logistic regression model with a one-hot encoding of group attributes. As shown, personalization can improve overall performance while reduces performance for specific groups (red). This result holds across all performance metrics. In such cases, we can resolve fair use violations and improve the gains from personalization by assigning personalized predictions to each group with multiple models. By this we mean selecting from one of three available models which provides the most accurate predictions for a group: a generic model h_0, the personalized model h_g, or a decoupled model h^{dcp}. We highlight cases where assigning predictions from one of these models led to a gain in green, and where it resolved a violation in yellow.

for (`male,black`) and (`male,other`); (2) greatly improve accuracy for (`female,black`) who experience a gain of **37.3%** from a previous accuracy of less than 50%; and (3) improve overall gains by 6.2%. We observe similar effects across other configurations and model classes.

6. Concluding Remarks

Machine learning models that are personalized with group attributes can fail to improve performance for all groups who provide personal data. Our results underscore the need to evaluate fair use when developing models with group attributes that are protected, sensitive, self-reported, or costly to acquire [e.g., 78, 89, 20, 23, 15, 17, 53]. Evaluating fair use is a routine procedure that whose results can be summarized and communicated in a model report [63, 6, 8, 16, 18] – and that can be used to flag instances where personalization reduces performance for specific groups and guide interventions that broaden the gains of personalization.

Limitations Our work assumes that a gain in performance is a suitable "stand-in" for preference or harm, which holds in tasks where every group benefits from a more accurate model. This assumption may not hold when, for example, models are trained to use proxy labels, or groups may prefer a specific prediction over the most accurate prediction.

In closing, we caution that fair use should be considered a safeguard against "worsenalization" rather than a rubber stamp for consent. In effect, fair use is not an individual-level guarantee. The gains associated with fair use conditions reflect average measures of performance over individuals in a group. In tasks where these gains are reported to individuals, they should be presented alongside information that summarizes the impact of personalization on their prediction – e.g., the degree of change in individual predictions due to personalization, and the degree of representation in the sample used to evaluate the gains of personalization.

Acknowledgements

We thank the following individuals for helpful discussions: Flavio Calmon, Katherine Heller, Sanmi Koyejo, Ziad Obermeyer, Charlie Marx, Stephen Pfohl, Emma Pierson, Kush Varshney, and Haoran Zhang. This work was supported by funding from the National Science Foundation IIS 2040880, the NIH Bridge2AI Center Grant U54HG012510, and the Wellcome Trust.

References

[1] Adam, H., Yang, M. Y., Cato, K., Baldini, I., Senteio, C., Celi, L. A., Zeng, J., Singh, M., and Ghassemi, M. Write it like you see it: Detectable differences in clinical notes by race lead to differential model recommendations. In *AIES 2022: Conference on Artificial Intelligence, Ethics and Society*, 2022.

[2] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2018.

[3] Alabi, D., Immorlica, N., and Kalai, A. Unleashing linear optimizers for group-fair learning and optimization. In *Conference On Learning Theory*, pp. 2043–2066, 2018.

[4] Allyn, J., Ferdynus, C., Bohrer, M., Dalban, C., Valance, D., and Allou, N. Simplified acute physiology score ii as predictor of mortality in intensive care units: a decision curve analysis. *PloS one*, 11(10):e0164828, 2016.

[5] Arlot, S. and Celisse, A. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[6] Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., et al. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6–1, 2019.

[7] Bansal, G., Gefen, D., et al. The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. *Decision support systems*, 49(2):138–150, 2010.

[8] Barocas, S., Guo, A., Kamar, E., Krones, J., Morris, M. R., Vaughan, J. W., Wadsworth, D., and Wallach, H. Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. *arXiv preprint arXiv:2103.06076*, 2021.

[9] Baron, R. M. and Kenny, D. A. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

[10] Bertsimas, D. and Kallus, N. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

[11] Bertsimas, D., Dunn, J., and Mundru, N. Optimal prescriptive trees. *INFORMS Journal on Optimization*, 1(2):164–183, 2019.

[12] Bien, J., Taylor, J., and Tibshirani, R. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111, 2013.

[13] Biggs, M., Sun, W., and Ettl, M. Model distillation for revenue optimization: Interpretable personalized pricing. *arXiv preprint arXiv:2007.01903*, 2020.

[14] Blaha, M. J. The critical importance of risk score calibration: time for transformative approach to risk score validation?, 2016.

[15] Bouwmeester, W., Zuithoff, N. P., Mallett, S., Geerlings, M. I., Vergouwe, Y., Steyerberg, E. W., Altman, D. G., and Moons, K. G. Reporting and methods in clinical prediction research: a systematic review. *PLoS medicine*, 9(5): e1001221, 2012.

[16] Bynum, L., Loftus, J., and Stoyanovich, J. Disaggregated interventions to reduce inequality. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–13. 2021.

[17] Cabitza, F. and Campagner, A. The need to separate the wheat from the chaff in medical informatics, 2021.

[18] Cabrera, Á. A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., and Chau, D. H. Fairvis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 46–56. IEEE, 2019.

[19] Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 319–328. ACM, 2019.

[20] Collins, G. S., Reitsma, J. B., Altman, D. G., and Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *Journal of British Surgery*, 102(3):148–158, 2015.

[21] Commission, F. T. Equal credit opportunity act. https://www.fdic.gov/resources/supervision-and-examinations/consumer-compliance-examination-manual/documents/5/v-7-1.pdf, 2020.

[22] Corbett-Davies, S. and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[23] Cowley, L. E., Farewell, D. M., Maguire, S., and Kemp, A. M. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. *Diagnostic and prognostic research*, 3(1):1–23, 2019.

[24] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., and Froelicher, V. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310, 1989.

[25] DiCiccio, T. J. and Efron, B. Bootstrap confidence intervals. *Statistical science*, pp. 189–212, 1996.

[26] Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[27] Do, V., Corbett-Davies, S., Atif, J., and Usunier, N. Online certification of preference-based fairness for personalized recommender systems. *arXiv preprint arXiv:2104.14527*, 2021.

[28] Dunn, O. J. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

[29] Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 119–133. PMLR, 2018.

[30] Elmachtoub, A. N., Gupta, V., and Hamilton, M. The value of personalized pricing. *Available at SSRN 3127719*, 2018.

[31] Eneanya, N. D., Yang, W., and Reese, P. P. Reconsidering the consequences of using race to estimate kidney function. *Jama*, 322(2):113–114, 2019.

[32] Eusebi, P. Diagnostic accuracy measures. *Cerebrovascular Diseases*, 36(4):267–272, 2013.

[33] Fan, H. and Poole, M. S. What is personalization? perspectives on the design and implementation of personalization in information systems. *Journal of Organizational Computing and Electronic Commerce*, 16(3-4):179–202, 2006.

[34] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM, 2015.

[35] Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., and Saria, S. The clinician and dataset shift in artificial intelligence. *The New England journal of medicine*, 385(3):283–286, 2021.

[36] Flamm, B. L. and Geiger, A. M. Vaginal birth after cesarean delivery: an admission scoring system. *Obstetrics & Gynecology*, 90(6):907–910, 1997.

[37] Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., and Mulvihill, J. J. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute*, 81(24):1879–1886, 1989.

[38] Gichoya, J. W., Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L.-C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.-C., et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 2022.

[39] Globus-Harris, I., Kearns, M., and Roth, A. An algorithmic framework for bias bounties. 2022.

[40] Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[41] Guo, L. L., Pfohl, S. R., Fries, J., Posada, J., Fleming, S. L., Aftandilian, C., Shah, N., and Sung, L. Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. *Applied Clinical Informatics*, 12(04):808–815, 2021.

[42] Hardt, M., Price, E., Srebro, N., et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

[43] Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

[44] Haukoos, J. S., Lyons, M. S., Lindsell, C. J., Hopkins, E., Bender, B., Rothman, R. E., Hsieh, Y.-H., MacLaren, L. A., Thrun, M. W., Sasson, C., et al. Derivation and validation of the denver human immunodeficiency virus (hiv) risk score for targeted hiv screening. *American journal of epidemiology*, 175(8):838–846, 2012.

[45] Hébert-Johnson, Ú., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the International Conference on Machine Learning*, pp. 1944–1953, 2018.

[46] Hu, L. and Chen, Y. Fair Classification and Social Welfare. *arXiv preprint arXiv:1905.00147*, 2019.

[47] James, H., Nagpal, C., Heller, K., and Ustun, B. Participatory systems for personalized prediction. *arXiv preprint arXiv:2302.03874*, 2023.

[48] Jaques, N., Taylor, T. S., Nosakhare, N. E., Sano, S. A., and Picard R, . P. R. Multi-task learning for predicting health, stress, and happiness. *Neural Information Processing Systems (NeurIPS) Workshop on Machine Learning for Healthcare*, 2016.

[49] Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[50] Jovanovic, B. Truthful disclosure of information. *The Bell Journal of Economics*, pp. 36–44, 1982.

[51] Kanis, J. A. et al. Assessment of osteoporosis at the primary health-care level. Technical report, Technical report, 2007.

[52] Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2018.

[53] Kent, D. M., Paulus, J. K., Van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P., Patrick-Lake, B., Morton, S., Pencina, M., et al. The predictive approaches to treatment effect heterogeneity (path) statement. *Annals of internal medicine*, 172(1):35–45, 2020.

[54] Kessler, R. C., Adler, L., Ames, M., Demler, O., Faraone, S., Hiripi, E., Howes, M. J., Jin, R., Secnik, K., Spencer, T., et al. The world health organization adult adhd self-report scale (asrs): a short screening scale for use in the general population. *Psychological medicine*, 35(2):245–256, 2005.

[55] Kim, M. P., Korolova, A., Rothblum, G. N., and Yona, G. Preference-informed fairness. *arXiv preprint arXiv:1904.01793*, 2019.

[56] Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. Algorithmic Fairness. In *AEA Papers and Proceedings*, volume 108, pp. 22–27, 2018.

[57] Le Gall, J.-R., Lemeshow, S., and Saulnier, F. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24): 2957–2963, 1993.

[58] Levey, A. S., Stevens, L. A., Schmid, C. H., Zhang, Y., Castro III, A. F., Feldman, H. I., Kusek, J. W., Eggers, P., Van Lente, F., Greene, T., et al. A new equation to estimate glomerular filtration rate. *Annals of internal medicine*, 150 (9):604–612, 2009.

[59] Lim, M. and Hastie, T. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, 2015.

[60] Lipton, Z., McAuley, J., and Chouldechova, A. Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems 31*, pp. 8135–8145, 2018.

[61] Martinez, N., Bertran, M., and Sapiro, G. Fairness with minimal harm: A pareto-optimal approach for healthcare. *arXiv preprint arXiv:1911.06935*, 2019.

[62] Martinez, N., Bertran, M., and Sapiro, G. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pp. 6755–6764. PMLR, 2020.

[63] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

[64] Mitchell, T. M. et al. *Machine learning*, volume 1. McGraw-hill New York, 2007.

[65] Moore, C. L., Bomann, S., Daniels, B., Luty, S., Molinaro, A., Singh, D., and Gross, C. P. Derivation and validation of a clinical prediction rule for uncomplicated ureteral stone—the stone score: retrospective and prospective observational cohort studies. *Bmj*, 348, 2014.

[66] Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[67] Narasimhan, H. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 1646–1654, 2018.

[68] Netanel, N. W. Making sense of fair use. *Lewis & Clark L. Rev.*, 15:715, 2011.

[69] Paes, L. M., Long, C. X., Ustun, B., and Calmon, F. On the epistemic limits of personalized prediction. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Snp3iEj7NJ.

[70] Paulus, J. K. and Kent, D. M. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ digital medicine*, 3(1):1–8, 2020.

[71] Paulus, J. K., Wessler, B. S., Lundquist, C., Lai, L. L., Raman, G., Lutz, J. S., and Kent, D. M. Field synopsis of sex in clinical prediction models for cardiovascular disease. *Circulation: Cardiovascular Quality and Outcomes*, 9(2_suppl_1): S8–S15, 2016.

[72] Perez-Rodriguez, J. and de la Fuente, A. Now is the time for a postracial medicine: Biomedical research, the national institutes of health, and the perpetuation of scientific racism. *The American Journal of Bioethics*, 17(9):36–47, 2017.

[73] Pfohl, S., Marafino, B., Coulet, A., Rodriguez, F., Palaniappan, L., and Shah, N. H. Creating fair models of atherosclerotic cardiovascular disease risk. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 271–278, 2019.

[74] Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[75] Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[76] Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. Mit Press, 2008.

[77] Savage, L. J. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66 (336):783–801, 1971.

[78] Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.

[79] Taylor, S., Jaques, N., Nosakhare, E., Sano, A., and Picard, R. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing*, 11(2):200–213, 2017.

[80] Ustun, B. and Rudin, C. Supersparse Linear Integer Models for Optimized Medical Scoring Systems. *Machine Learning*, 102(3):349–391, 2016.

[81] Ustun, B. and Rudin, C. Learning optimized risk scores. *Journal of Machine Learning Research*, 20(150):1–75, 2019.

[82] Ustun, B., Westover, M. B., Rudin, C., and Bianchi, M. T. Clinical prediction models for sleep apnea: the importance of medical history over symptoms. *Journal of Clinical Sleep Medicine*, 12(02):161–168, 2016.

[83] Ustun, B., Adler, L. A., Rudin, C., Faraone, S. V., Spencer, T. J., Berglund, P., Gruber, M. J., and Kessler, R. C. The world health organization adult attention-deficit/hyperactivity disorder self-report screening scale for dsm-5. *Jama psychiatry*, 74(5):520–527, 2017.

[84] Ustun, B., Liu, Y., and Parkes, D. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pp. 6373–6382, 2019.

[85] van den Goorbergh, R., van Smeden, M., Timmerman, D., and Van Calster, B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, 29(9):1525–1534, 06 2022. ISSN 1527-974X. doi: 10.1093/jamia/ocac093. URL https://doi.org/10.1093/jamia/ocac093.

[86] Vaughan, G., Aseltine, R., Chen, K., and Yan, J. Efficient interaction selection for clustered data via stagewise generalized estimating equations. *Statistics in Medicine*, 39(22): 2855–2868, 2020.

[87] Viviano, D. and Bradic, J. Fair policy targeting. *arXiv preprint arXiv:2005.12395*, 2020.

[88] Vyas, D. A., Eisenstein, L. G., and Jones, D. S. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.

[89] Wallace, E., Smith, S. M., Perera-Salazar, R., Vaucher, P., McCowan, C., Collins, G., Verbakel, J., Lakhanpaul, M., and Fahey, T. Framework for the impact analysis and implementation of clinical prediction rules (cprs). *BMC medical informatics and decision making*, 11(1):1–7, 2011.

[90] Wang, A., Ramaswamy, V. V., and Russakovsky, O. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. *arXiv preprint arXiv:2205.04610*, 2022.

[91] Wang, H., Ustun, B., and Calmon, F. P. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2019.

[92] Wang, H., Hsu, H., Diaz, M., and Calmon, F. P. To split or not to split: The impact of disparate treatment in classification. *IEEE Transactions on Information Theory*, 67(10): 6733–6757, 2021.

[93] Yankwich, L. R. What is fair use? *The University of Chicago Law Review*, 22(1):203–215, 1954.

[94] Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

[95] Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., and Weller, A. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pp. 228–238, 2017.

[96] Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 962–970. PMLR, 20–22 Apr 2017.

[97] Zhan, Q., Sierra, E., Malmsten, J., Ye, Z., Rosenwaks, Z., and Zaninovic, N. Blastocyst score, a blastocyst quality ranking tool, is a predictor of blastocyst ploidy and implantation potential. *F&S Reports*, 1(2):133–141, 2020.

[98] Zhang, H., Morris, Q., Ustun, B., and Ghassemi, M. Learning optimal predictive checklists. *Advances in Neural Information Processing Systems*, 34, 2021.

A. Notation

We provide a list of the notation used throughout the paper in Table 3.

Symbol	Meaning
$\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d})$	feature vector of example i
$y_i \in Y$	label of example i
$\boldsymbol{g}_i \in \{g_{i,1}, g_{i,2}, \ldots, g_{i,k}\}$	group membership of example i
$G = G_1 \times G_2 \times \ldots \times G_k$	space of group attributes
$m = \lvert G \rvert$	number of intersectional groups
$n_{\boldsymbol{g}} := \sum 1[\boldsymbol{g}_i = \boldsymbol{g}]$	number of examples of group $\boldsymbol{g} \in G$
$n^+_{\boldsymbol{g}} := \sum 1[\boldsymbol{g}_i = \boldsymbol{g}, y_i = +1]$	number of examples of group $\boldsymbol{g} \in G$ with $y_i = +1$
$n^-_{\boldsymbol{g}} := \sum 1[\boldsymbol{g}_i = \boldsymbol{g}, y_i = -1]$	number of examples of group $\boldsymbol{g} \in G$ with $y_i = -1$
$h : X \times G \to Y$	personalized model
H	hypothesis class of personalized models
$h_{\boldsymbol{g}} : X \times G \to Y$	personalized classifier where group membership is reported truthfully as $\boldsymbol{g}$
$h_0 : X \to Y$	generic model
H_0	hypothesis class of generic models
$R_{\boldsymbol{g}}(h_{\boldsymbol{g}'})$	true risk of model h_0 of group $\boldsymbol{g}$ if they report $\boldsymbol{g}'$
$\hat{R}_{\boldsymbol{g}}(h_{\boldsymbol{g}'})$	empirical risk of model h of group $\boldsymbol{g}$ if they report $\boldsymbol{g}'$
$\gamma_{\boldsymbol{g}}(h, h')$	gain (i.e., reduction in true risk) for group $\boldsymbol{g}$ when using h instead of h'
$\alpha_{\boldsymbol{g}}(h_{\boldsymbol{g}}, h_0)$	rationality gap for group $\boldsymbol{g}$ under model h
$\beta_{\boldsymbol{g}}(h_{\boldsymbol{g}}, h_{\boldsymbol{g}'})$	envyfreeness gap for group $\boldsymbol{g}$ under model h

Table 3: Notation

B. Supporting Material for Section 3

B.1. Additional Failure Modes of Personalization

We describe additional mechanisms that lead personalized models to exhibit fair use violations. The mechanisms below reflect failure modes that arise in later stages of the machine learning pipeline, and that are more difficult to address through interventions.

ERM with a Surrogate Loss Function Consider a setting where we want a personalized model that maximizes classification accuracy – i.e., one that minimizes the 0–1 loss. If we fit this classifier using a linear SVM – e.g., by solving an ERM problem that optimizes the hinge loss – the approximation error between the 0-1 loss and the hinge loss can produce a fair use violation (see Figure 4). This example is specifically designed to avoid fair use violations that stem from model misspecification.

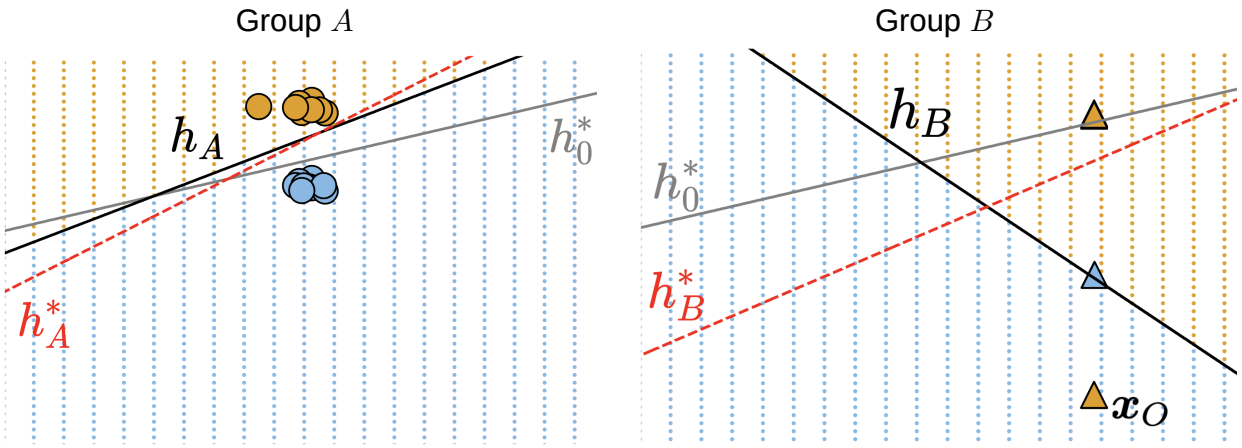

Figure 4: Fair use violations resulting from empirical risk minimization with a surrogate loss function. We consider a classification task with two features $\boldsymbol{x} = (x_1, x_2)$ and one group attribute $\boldsymbol{g} \in \{A, B\}$ in which we fit a linear SVM $h_{\boldsymbol{g}}$ but evaluate the the gains of personalization in terms of the error rate (i.e., hinge loss vs. 0-1 loss). We plot the data for group A and group B separately, and show the generic classifier (h_0; grey) and the personalized classifiers for the corresponding group (h_A or h_B; black). In this case, the personalized model produces a fair use violation for Group B due to an outlier $\boldsymbol{x}_O$. As a baseline for comparison, we show the personalized models that we would obtain by optimizing an exact loss function (i.e., 0-1 loss, which matches the performance metric that we use to evaluate the gains for personalization). As shown, we would expect to avoid this violation had we fit a model by optimizing the 0–1 loss directly.

Generalization & Dataset Shifts Fair use violations can arise in deployment. Small samples may distort the relative prevalence of each group, leading ERM to return a personalized model or suboptimal generic model. In Fig. 5, we show how fair use violations occur when sampling bias results in a difference in the training data distribution and the true distribution. Here, we sample data from the true distribution where the small sample size or sampling bias results in a label shift for one specific group. Likewise, violations can arise as a result of changes in the data distribution [i.e., dataset shift 76, 35, 41] (see Fig. 6)

Group		Training Data		Data Distribution		Predictions		Observed Performance			True Performance		
g_1	g_2	n^+	n	n^+	n	h_0	$h_{\boldsymbol{g}}$	$R_{\boldsymbol{g}}(h_0)$	$R_{\boldsymbol{g}}(h)$	$_{\boldsymbol{g}}(h_{\boldsymbol{g}}, h_0)$	$R_{\boldsymbol{g}}(h_0)$	$R_{\boldsymbol{g}}(h)$	$_{\boldsymbol{g}}(h_{\boldsymbol{g}}, h_0)$
0	0	65	60	130	120	+	+	60	60	0	120	120	0
1	0	60	65	120	130	+		65	60	5	130	120	10
0	1	60	65	130	120	+		65	60	5	120	130	10
1	1	70	55	140	110	+	+	55	55	0	110	110	0
	Total	255	245	520	480			245	235	10	480	470	0

Figure 5: Fair use violations can arise when personalizing models on small samples. Here, we show a 2D classification task in which a personalized model only exhibits fair use violations in deployment. Here, group $(1, 0)$ experiences an gain once the model is deployment. In contrast, group $(0, 1)$ experiences a fair use violation as a result of sampling error.

Group		Training Data		Data Distribution		Predictions		Observed Performance			True Performance		
g_1	g_2	n^+	n	n^+	n	h_0	h	$R_g(h_0)$	$R_g(h)$	$_g(h_g, h_0)$	$R_g(h_0)$	$R_g(h)$	$_g(h_g, h_0)$
0	0	20	0	20	0	+	+	0	0	0	0	0	0
1	0	5	25	5	25	+		25	5	20	25	5	20
0	1	5	25	30	25	+		25	5	20	20	30	10
1	1	20	0	20	0	+	+	0	0	0	0	0	0
	Total	50	50	75	45			50	10	40	45	35	10

Figure 6: Label shift produces a fair use violation. Here, we train a linear classifier on a dataset with [one binary feature and one binary group attribute]. As shown, personalization leads to overall improvement reducing aggregate reduce from 50 to 24 and group-specific improvement on the training data. However, not all groups perform equally well in deployment. While groups $(0, 1)$ and $(1, 1)$ see improvements, a violation (red) occurs for group $(1, 0)$ due to the label shift where positive examples in the true distribution for group $(0, 1)$ (highlighted in yellow) are undersampled in the training data.

B.2. Missing Proofs

We provide the proofs for our sufficient conditions described in Section 3. We start with a simple condition to ensure the empirical risk minimizer over H can return a model that assigns the same predictions as a generic model for every group.

Definition 5. *A personalized model class* H extends *a generic model class* H_0 *if for every personalized model* $h \in H$*, there exists a generic model* $h_0 \in H_0$ *such that* $h_0(\boldsymbol{x}) = h(\boldsymbol{x}, \boldsymbol{g})$ *for all* $\boldsymbol{x} \in X$ *and all groups* $\boldsymbol{g} \in G$*.*

This is a basic condition that is often satisfied in practice, and can be guaranteed by practitioners during model specification. Intuitively the condition is meant to rule out instances where a personalized model exhibits a rationality violation because it is required to account for group membership (see e.g., Example 2).

Proposition 3 Consider training a personalized model by ERM $h \in \text{argmin}_{h \in H} \hat{R}(h)$, and evaluating its gains to personalization with respect to a generic model $h_0 \in \text{argmin}_{h \in H_0} \hat{R}(h)$ where $H_0 \subseteq H$. The personalized model h obeys fair use in terms of empirical risk so long as:

$$\hat{R}_{\boldsymbol{g}}(h) = \hat{R}_{\boldsymbol{g}}(h_{\boldsymbol{g}}) \text{ for all groups } \boldsymbol{g} \in G.$$

Proof. Say that we have a personalized model $h \in \text{argmin}_{h \in H} \hat{R}(h)$ that obeys $\hat{R}_{\boldsymbol{g}}(h) = \hat{R}_{\boldsymbol{g}}(h_{\boldsymbol{g}})$ for all groups $\boldsymbol{g} \in G$. This implies that $\hat{R}_{\boldsymbol{g}}(h_{\boldsymbol{g}}) \leq \hat{R}_{\boldsymbol{g}}(h)$ for any model $h \in H$ and any group $\boldsymbol{g} \in G$. Since $h_0 \in H$, we have that $\hat{R}_{\boldsymbol{g}}(h_{\boldsymbol{g}}) \leq \hat{R}_{\boldsymbol{g}}(h_0)$ for all groups $\boldsymbol{g} \in G$. Thus, rationality holds for all groups $\boldsymbol{g} \in G$. Likewise, since $h_{\boldsymbol{g}'} \in H$, we have that $\hat{R}_{\boldsymbol{g}}(h_{\boldsymbol{g}}) \leq \hat{R}_{\boldsymbol{g}}(h_{\boldsymbol{g}'})$ for all groups $\boldsymbol{g}, \boldsymbol{g}' \in G$. Thus, envy-freeness holds for all groups $\boldsymbol{g} \in G$.

□

Proposition 4 Consider a personalized model $h : X \times G \to Y$ that ensures rationality and envy-freeness for group $\boldsymbol{g}$ in terms of empirical risk. Denote the empirical gains in rationality and envy-freeness for group $\boldsymbol{g}$ as:

$$\hat{\epsilon}_{\boldsymbol{g}} := \hat{}_{\boldsymbol{g}}(h_{\boldsymbol{g}}, h_0), \qquad \hat{\gamma}_{\boldsymbol{g}} := \min_{\boldsymbol{g}' \in G / \{\boldsymbol{g}\}} \hat{}_{\boldsymbol{g}}(h_{\boldsymbol{g}}, h_{\boldsymbol{g}'})$$

If $\hat{\epsilon}_{\boldsymbol{g}} > 0$, then rationality for group $\boldsymbol{g}$ generalizes with probability at least $1 - \delta$ as long as:

$$n_{\boldsymbol{g}} \geq \frac{4D \log\left(\frac{2n_{\boldsymbol{g}}}{D} + 1\right) + \log\left(\frac{8}{\delta}\right)}{\hat{\epsilon}_{\boldsymbol{g}}^2} \tag{3}$$

If $\hat{\gamma}_{\boldsymbol{g}} > 0$, then envy-freeness for group $\boldsymbol{g}$ generalizes with probability at least $1 - \delta$ as long as:

$$n_{\boldsymbol{g}} \geq \frac{4D \log\left(\frac{2n_{\boldsymbol{g}}}{D} + 1\right) + \log\left(\frac{8m}{\delta}\right)}{\hat{\gamma}_{\boldsymbol{g}}^2} \tag{4}$$

The proof of Proposition 4 are based on a generalized version of a lemma from Ustun et al. [84] which assumes that the dimension of the hypothesis class is finite whereas we use VC-dimension instead of the dimension of the hypothesis class.

Lemma 6 (Generalization of Gains)**.** *Consider a pair of classifiers h_a and h_b from a hypothesis class H with VC-dimension D. If the empirical risk of each classifier on group g satisfy $\hat{}_g(h_a, h_b) := \hat{R}_g(h_b) \quad \hat{R}_g(h_a) > 0$, then for any $\delta > 0$, the corresponding gap in true risk will satisfy $\quad_g(h_a, h_b) > 0$ with probability at least $1 \quad \delta$ so long as:*

$$\sqrt{\frac{4D\left(\log\frac{2n_g}{D} + 1\right) + \log\left(\frac{8}{\delta}\right)}{n_g}} \quad \hat{}_g(h_a, h_b). \tag{5}$$

Proof. The proof applies a standard concentration inequality [64] to bound the generalization error of a classifier over groups as follows. Given a classifier $h \; 2 \; H$ from hypothesis class H with VC-dimension D, and any $\delta > 0$, the generalization error of h on group $g \; 2 \; G$ with n_g will obey the following inequality [64] with probability at least $1 \quad \frac{\delta}{2}$:

$$\left|\hat{R}_g(h) \quad R_g(h)\right| \quad \sqrt{\frac{D\left(\log\frac{2n_g}{D} + 1\right) + \log\frac{8}{\delta}}{n_g}}. \tag{6}$$

We denote the quantity on the right hand side of Eq. (6) as the bounding function $B(n_g, H, \delta) := \sqrt{\frac{D\left(\log\frac{2n_g}{D}+1\right)+\log\frac{8}{\delta}}{n_g}}$. Given the bounding function $B(n_g, H, \delta)$, Lemma 6 states that for any $\delta > 0$, with probability at least $1 \quad \delta$,

$$2B(n_g, H, \delta) \quad \hat{}_g(h_a, h_b) \quad =) \quad R_g(h_b) \quad R_g(h_a) \quad 0$$

We will prove the statement by showing that the condition on the left hand side implies the condition on the right hand side. Assume that the condition on the left hand side holds so that $2B(n_g, H, \delta) \quad \hat{}_g(h_a, h_b)$. Then we can observe that the right hand side is bounded as follows:

$$\begin{aligned} R_g(h_b) \quad R_g(h_a) &= R_g(h_b) \quad R_g(h_a) + \hat{R}_g(h_a) \quad \hat{R}_g(h_a) + \hat{R}_g(h_b) \quad \hat{R}_g(h_b) \\ &= \underbrace{R_g(h_b) \quad \hat{R}_g(h_b)}_{B(n_g,H,\delta)} + \underbrace{\hat{R}_g(h_a) \quad R_g(h_a)}_{B(n_g,H,\delta)} + \underbrace{\hat{R}_g(h_b) \quad \hat{R}_g(h_a)}_{:=\hat{}_g(h_a,h_b)} \\ &\quad 2B(n_g, H, \delta) + \hat{}_g(h_a, h_b) \\ &\quad 0 \end{aligned}$$

Thus we have that $R_g(h_b) \quad R_g(h_a) \quad 0$ whenever $2B(n_g, H, \delta) \quad \hat{}_g(h_a, h_b)$. This completes the proof. □

We now present the proof to Proposition 4.

Proof. We recover the bounds by applying Lemma 6. We start with the bound on rationality in Eq. (3). Given that $\hat{\epsilon}_g > 0$, we apply Lemma 6 to the personalized and model h_g and the generic model h_0 to obtain:

$$\sqrt{\frac{4D\left(\log\frac{2n_g}{D} + 1\right) + \log\left(\frac{8}{\delta}\right)}{n_g}} \quad \hat{\epsilon}_g$$

$$n_g \quad \frac{4D\left(\log\frac{2n_g}{D} + 1\right) + \log\left(\frac{8}{\delta}\right)}{\hat{\epsilon}_g^2}$$

We now consider the bound on envy-freeness Eq. (4). Given that $\hat{\gamma}_g > 0$, we apply Lemma 6 to the personalized model h_g and $h_{g'}$ for all $g, g' \; 2 \; G$. This produces $m \quad 1$ preferences to generalize. Given that $m \quad 1 \quad m$, we apply Lemma 6 with probability $1 \quad \frac{\delta}{m}$. Doing so and inverting for n_g proves the result.

$$\sqrt{\frac{4D\left(\log\frac{2n_g}{D} + 1\right) + \log\frac{8m}{\delta}}{n_g}} \quad \hat{\gamma}_g$$

$$n_g \quad \frac{4D\left(\log\frac{2n_g}{D} + 1\right) + \log\left(\frac{8m}{\delta}\right)}{\hat{\gamma}_g^2}$$

□

C. Additional Information on Datasets

In this Appendix, we include additional information on the datasets used in Section 4 and Section 5. We present a summary of the goals and characteristics for each dataset in Table 4. We include a brief description of each dataset and preprocessing steps taken below.

Dataset	n	d	Group Attributes – G	Prediction Task	Reference
apnea	1,152	26	Age Sex = {<30, 30 to 60, 60+} {Male, Female}	patient has obstructive sleep apnea	Ustun et al. [82]
cardio_eicu	1,341	49	Age Sex = {Young, Old} {Male, Female}	patient with cardiogenic shock dies	Pollard et al. [75]
cardio_mimic	5,289	49	Age Sex = {Young, Old} {Male, Female}	patient with cardiogenic shock dies	Johnson et al. [49]
heart	181	26	Age Sex = {Young, Old} {Male, Female}	patient has heart disease	Detrano et al. [24]
kidney	2,066	78	Sex Race = {Male, Female} {White, Black, Other}	mortality of patient on CRRT	Zhang et al. [98]
mortality	21,139	484	Age Sex = {< 30, 30 to 60, 60+} {Male, Female}	mortality of patient in ICU	Harutyunyan et al. [43]
saps	7,797	36	Age HIV = { 30, 30+} {Positive, Negative}	mortality of patient in ICU	Le Gall et al. [57]

Table 4: Clinical prediction tasks considered in Section 4 and Section 5. We state conditions for $y_i = +1$ for each dataset. All datasets used are publicly available. Datasets based on MIMIC-III [49] (`kidney`, `mortality`) and eICU [75] (`cardio`) are hosted on PhysioNet under the PhysioNet Credentialed Health Data License. The `heart` dataset is hosted on the UCI ML Repository under an Open Data license. The `apnea` and `saps` datasets must be requested from the authors of the papers listed under references [57, 82]. In cases where data access requires consent or approval from the data holders, we have followed the proper procedure to obtain such consent.

`apnea` We use the obstructive sleep apnea dataset from Ustun et al. [82] [see also 80]. The dataset contains a cohort of 1,152 patients of which $\Pr(y = +1) = 23\%$ have OSA and includdes 26 features that cover information that is readily available in an electronic health record (e.g. BMI, comobordities, age, sex).

`cardio_eicu` & `cardio_mimic` Cardiogenic shock is a serious acute condition where the heart cannot provide sufficient blood to the vital organs. We create a cohort of patients who have cardiogenic shock during an ICU stay from the eICU Collaborative Research Database V2.0[75] and MIMIC-III databases [49], respectively. The goal is to predict mortality for a patient with cardiogenic shock. As features include summarize statistics for vitals and lab tests (e.g. systolic BP, heart rate, hemoglobin count) obtained up to 24 hours prior to the onset of cardiogenic shock. The final dataset contains 8,815 patients and $\Pr(y_i = +1) = 13.5\%$.

`heart` We use the Heart dataset from the UCI Machine Learning Repository, where the goal is to predict the presence of heart disease which covers a cohort of 303 patients, of which $\Pr(y_i = +1) = 54.5\%$ have heart disease. We use all available features, treating `cp`, `thal`, `ca`, `slope` and `restecg` as categorical, and all remaining features as continuous.

`kidney` We use MIMIC-III and MIMIC-IV [49] to define a cohort of patients who were given *continuous renal replacement therapy* (CRRT) at any point during their ICU stay. For patients with multiple ICU stays, we select their first one. We define the target as whether the patient dies during the course of their selected hospital admission. As features, we select the most recent instances of relevant lab measurements (e.g. sodium, potassium, creatinine) prior to the CRRT start time, along with the patient's age, the number of hours they have been in ICU when CRRT was administered, and their Sequential Organ Failure Assessment (SOFA) score at admission. We treat all variables as continuous with the exception of the SOFA score, which we treat as ordinal. This results in a dataset of 1,722 CRRT patients, with $\Pr(y_i = +1) = 51.1\%$.

`mortality` We define a cohort of patients for in-hospital mortality prediction task following Harutyunyan et al. [43]. We select the first ICU stay longer than 48 hours for patients in MIMIC-III[49], and predict in-hospital mortality for this visit. As features, we include periodic lab and vital measurements used by Harutyunyan et al. [43] into four 12-hour time-bins, and compute the mean in each time-bin. This results in a cohort of 21,139 patients where $\Pr(y_i = +1) = 13.2\%$.

`saps` The Simplified Acute Physiology Score II (SAPS II) is a risk score developed to predict ICU mortality [57]. This study contains a cohort of critically-ill patients from 137 medical centers across 12 countries. For each patient we have access to demographics, comorbidities, and vitals which are used to predict the risk of mortality in the ICU. The final dataset contains 7,797 patients where percentage of patients in the dataset who experience mortality is $\Pr(y_i = +1) = 21.8\%$.

D. Additional Experimental Results

We include additional results showing the gains of personalization when training personalized neural nets and random forests. We present tables that summarize the gains of personalization for neural networks and random forests. The following tables are analogous to Table 1, except that they also include results for the `kidney` dataset in Section 5.

Neural Nets We trained neural networks with two hidden layers of size 5 and 2 and learning rate of 1^{-3}. We applied Platt scaling [74] to ensure that the models assigned calibrated probabilities. As in Section 4.2 and Section 5, we can identify significant fair use violations and gains as noted by the gains and violations.

		Test Error			Test AUC			Test ECE		
Dataset	Metrics	1Hot	All	DCP	1Hot	All	DCP	1Hot	All	DCP
`apnea`	Personalized	35.0%	48.4%	41.5%	0.704	0.502	0.622	4.8%	2.4%	5.3%
$n = 1152, d = 26$	Gain	-1.7%	-15.1%	-8.1%	-0.012	-0.215	-0.095	0.8%	3.2%	0.4%
$G = f_{\texttt{age},\texttt{sex}}$	Best/Worst Gain	15.7% / -4.5%	-6.1% / -34.4%	-2.2% / -50.5%	0.097 / -0.040	-0.052 / -0.496	-0.068 / -0.328	8.0% / -9.5%	25.2% / 3.3%	9.8% / -5.7%
$m = 6$	Rat. Gains/Viols	4/3	6/6	6/6	2/2	0/0	0/0	2/2	0/0	2/2
Ustun et al. [82]	EF Gains/Viols	3/0	0/0	2/1	4/5	6/6	4/4	1/1	0/0	1/1
`cardio_eicu`	Personalized	31.5%	31.8%	36.6%	0.739	0.738	0.687	4.5%	5.5%	5.4%
$n = 1341, d = 49$	Gain	1.6%	1.3%	-3.5%	0.001	-0.001	-0.051	2.3%	1.4%	1.5%
$G = f_{\texttt{age},\texttt{sex}}$	Best/Worst Gain	8.4% / -0.5%	5.5% / -1.3%	0.0% / -10.3%	0.067 / -0.003	0.029 / -0.012	-0.000 / -0.091	2.6% / -1.2%	2.4% / -1.9%	5.4% / -2.8%
$m = 4$	Rat. Gains/Viols	0/0	2/1	3/3	3/3	1/1	0/0	1/1	1/1	2/2
Pollard et al. [75]	EF Gains/Viols	2/0	3/0	2/2	4/4	3/4	2/2	2/2	1/1	1/1
`cardio_mimic`	Personalized	23.7%	24.0%	23.9%	0.849	0.849	0.836	3.1%	4.7%	3.3%
$n = 5289, d = 49$	Gain	0.6%	0.2%	0.4%	0.004	0.004	-0.009	1.1%	-0.4%	1.0%
$G = f_{\texttt{age},\texttt{sex}}$	Best/Worst Gain	2.0% / -1.1%	2.3% / -2.4%	1.4% / -1.3%	0.018 / -0.005	0.012 / -0.000	0.003 / -0.015	2.1% / -0.4%	1.4% / -2.3%	2.5% / -0.2%
$m = 4$	Rat. Gains/Viols	1/1	2/2	2/2	3/3	3/3	1/1	1/1	2/2	0/0
Johnson et al. [49]	EF Gains/Viols	1/1	0/0	3/3	3/3	3/3	0/0	0/0	0/0	1/1
`heart`	Personalized	50.0%	26.3%	38.2%	0.451	0.771	0.554	21.3%	19.5%	18.1%
$n = 181, d = 26$	Gain	1.3%	25.0%	13.2%	-0.096	0.225	0.007	-7.8%	-5.9%	-4.5%
$G = f_{\texttt{sex},\texttt{age}}$	Best/Worst Gain	12.0% / -12.8%	29.7% / 16.6%	28.1% / 7.1%	0.046 / -0.387	0.393 / 0.119	0.257 / -0.023	-0.1% / -27.2%	16.8% / -5.7%	6.2% / -14.8%
$m = 4$	Rat. Gains/Viols	2/1	0/0	0/0	1/1	4/4	1/2	3/3	1/1	1/1
Detrano et al. [24]	EF Gains/Viols	2/1	2/1	3/1	1/4	0/3	1/2	1/1	0/0	1/1
`kidney`	Personalized	29.5%	31.7%	30.9%	0.758	0.774	0.762	5.6%	6.8%	7.3%
$n = 2066, d = 78$	Gain	-2.3%	-4.5%	-3.7%	-0.013	0.004	-0.009	0.3%	-0.9%	-1.4%
$G = f_{\texttt{sex},\texttt{ethnicity}}$	Best/Worst Gain	1.2% / -7.8%	5.2% / -6.8%	-1.6% / -16.3%	0.047 / -0.144	0.049 / -0.103	0.032 / -0.135	4.6% / -7.8%	1.9% / -5.6%	1.0% / -5.9%
$m = 6$	Rat. Gains/Viols	5/4	5/5	6/6	2/2	4/4	2/2	2/2	4/4	5/5
Zhang et al. [98]	EF Gains/Viols	3/0	1/0	4/4	4/6	3/5	2/2	1/0	0/0	3/3
`mortality`	Personalized	20.4%	21.6%	17.7%	0.870	0.869	0.895	2.8%	4.7%	3.0%
$n = 25366, d = 468$	Gain	0.1%	-1.1%	2.8%	-0.003	-0.004	0.022	0.6%	-1.3%	0.5%
$G = f_{\texttt{age},\texttt{sex}}$	Best/Worst Gain	5.2% / -1.7%	-0.6% / -3.2%	12.9% / 0.0%	0.032 / -0.018	-0.000 / -0.022	0.042 / 0.005	2.7% / -0.8%	2.9% / -1.8%	8.3% / 0.1%
$m = 6$	Rat. Gains/Viols	2/2	6/6	0/0	3/3	0/0	6/6	3/3	3/3	0/0
Johnson et al. [49]	EF Gains/Viols	5/1	2/2	6/6	0/4	4/4	0/0	4/1	3/3	6/6
`saps`	Personalized	53.9%	22.5%	48.9%	0.521	0.872	0.758	43.6%	9.4%	31.5%
$n = 7797, d = 36$	Gain	7.7%	39.0%	12.7%	0.328	0.679	0.565	1.7%	36.0%	13.9%
$G = f_{\texttt{hiv},\texttt{age}}$	Best/Worst Gain	13.1% / 0.0%	54.8% / 1.4%	22.0% / 0.0%	0.727 / 0.197	0.757 / 0.638	0.743 / -0.273	13.2% / 1.6%	45.1% / -2.9%	49.9% / 6.4%
$m = 4$	Rat. Gains/Viols	2/0	1/0	1/0	4/4	4/4	3/3	0/0	1/1	0/0
Allyn et al. [4]	EF Gains/Viols	4/1	3/0	3/2	1/3	1/3	3/4	0/0	1/1	1/1

Table 5: Gains of personalization for neural network models on test data.

Random Forests We trained random forests with the following hyperparameters: 100 estimators, max depth of 20, minimum samples per split is 5, and minimum number of samples in each leaf is 2. We expect these models to perform well in terms of error rare but not necessarily in terms in terms of AUC or risk calibration. We observe this effect in the Table below. For example, using an intersectional encoding with random forests minimizing fair use violations in terms of error rate as measured on multiple datasets (e.g. `apnea`, `kidney`). As noted with other model classes, we can find statistically significant violations.

		Test Error			Test AUC			Test ECE		
Dataset	Metrics	1Hot	All	DCP	1Hot	All	DCP	1Hot	All	DCP
apnea	Personalized	29.7%	31.0%	26.5%	0.751	0.757	0.815	8.2%	7.2%	8.0%
$n = 1152, d = 26$	Gain	1.8%	0.3%	5.3%	-0.004	-0.001	0.055	-2.3%	-1.1%	-1.0%
$G = f_{\text{age,sex}}g()$	Best/Worst Gain	4.7% / -4.4%	1.4% / -3.8%	17.0% / -6.0%	0.061 / -0.021	0.019 / -0.015	0.104 / -0.008	2.2% / -3.9%	2.1% / -2.0%	2.8% / -2.6%
$m = 6$	Rat. Gains/Viols	2/2	4/1	1/1	1/2	1/2	5/5	3/3	3/3	2/2
Ustun et al. [82]	EF Gains/Viols	3/0	3/1	5/5	2/6	2/4	1/1	0/0	1/1	0/0
cardio_eicu	Personalized	30.8%	30.5%	27.1%	0.770	0.769	0.801	8.0%	8.5%	9.4%
$n = 1341, d = 49$	Gain	0.4%	-0.3%	3.9%	0.003	0.003	0.032	-0.5%	-0.3%	-1.9%
$G = f_{\text{age,sex}}g$	Best/Worst Gain	3.6% / -3.6%	0.0% / -0.9%	16.4% / 0.4%	0.016 / -0.008	0.013 / -0.012	0.121 / 0.007	0.7% / -2.0%	0.5% / -0.0%	4.2% / -1.4%
$m = 4$	Rat. Gains/Viols	2/2	4/1	0/0	2/2	2/2	4/4	3/3	0/0	2/2
Pollard et al. [75]	EF Gains/Viols	1/0	2/2	4/4	2/3	1/1	0/0	1/1	1/1	0/0
cardio_mimic	Personalized	24.0%	23.7%	20.9%	0.849	0.850	0.871	10.0%	11.0%	11.7%
$n = 5289, d = 49$	Gain	-0.3%	0.3%	2.9%	0.001	-0.002	0.023	-0.8%	-0.9%	-2.3%
$G = f_{\text{age,sex}}g$	Best/Worst Gain	0.9% / -1.3%	0.6% / -0.1%	5.8% / 1.1%	0.003 / -0.002	0.004 / -0.004	0.047 / 0.007	0.5% / -1.7%	-0.0% / -1.6%	-0.1% / -4.6%
$m = 4$	Rat. Gains/Viols	2/2	1/0	0/0	3/3	1/1	4/4	3/3	3/3	3/3
Johnson et al. [49]	EF Gains/Viols	2/0	3/3	4/4	2/2	1/1	0/0	0/0	1/1	0/0
heart	Personalized	18.4%	21.1%	21.1%	0.899	0.896	0.936	9.2%	10.6%	13.5%
$n = 181, d = 26$	Gain	1.3%	2.6%	-1.3%	0.001	-0.000	0.035	2.0%	4.5%	1.1%
$G = f_{\text{sex,age}}g$	Best/Worst Gain	5.9% / 0.0%	10.8% / 0.0%	16.3% / -18.6%	0.004 / -0.067	0.016 / -0.063	0.094 / 0.001	7.1% / -3.7%	4.3% / -4.6%	11.3% / -13.8%
$m = 4$	Rat. Gains/Viols	3/0	3/0	3/2	0/3	1/3	3/4	2/2	2/2	2/2
Detrano et al. [24]	EF Gains/Viols	4/0	3/1	1/1	0/4	1/4	0/0	2/2	4/4	2/2
kidney	Personalized	30.1%	30.5%	22.3%	0.773	0.773	0.860	7.5%	7.5%	13.2%
$n = 2066, d = 78$	Gain	-0.4%	-1.2%	7.8%	-0.003	-0.005	0.083	0.8%	1.3%	-5.1%
$G = f_{\text{sex,ethnicity}}g$	Best/Worst Gain	0.5% / -3.4%	0.0% / -3.4%	17.8% / 3.1%	0.008 / -0.022	0.015 / -0.008	0.143 / 0.062	1.8% / -1.8%	1.7% / -2.6%	-1.7% / -8.1%
$m = 6$	Rat. Gains/Viols	4/2	6/4	0/0	2/2	2/2	6/6	3/3	1/1	6/6
Zhang et al. [98]	EF Gains/Viols	5/0	1/0	6/6	3/5	3/3	0/0	2/0	0/0	0/0
mortality	Personalized	27.1%	26.9%	24.6%	0.803	0.806	0.841	11.0%	10.9%	12.0%
$n = 25366, d = 468$	Gain	0.2%	0.4%	2.4%	-0.004	0.002	0.035	-0.5%	-0.6%	-0.9%
$G = f_{\text{age,sex}}g$	Best/Worst Gain	0.8% / -1.1%	1.0% / -0.7%	22.2% / 0.2%	0.004 / -0.011	0.009 / -0.011	0.186 / 0.013	0.2% / -1.2%	0.3% / -1.3%	0.7% / -7.8%
$m = 6$	Rat. Gains/Viols	4/4	2/2	0/0	1/1	3/3	6/6	3/3	4/4	4/4
Johnson et al. [49]	EF Gains/Viols	5/0	6/0	6/6	0/6	0/6	0/0	3/0	6/0	0/0
saps	Personalized	19.6%	19.8%	19.1%	0.880	0.880	0.882	5.0%	4.9%	4.7%
$n = 7797, d = 36$	Gain	0.0%	0.0%	0.6%	-0.001	0.000	0.002	-0.5%	-0.5%	0.3%
$G = f_{\text{hiv,age}}g$	Best/Worst Gain	0.1% / -0.2%	22.2% / -0.4%	4.6% / 0.0%	0.000 / -0.001	0.002 / -0.023	0.023 / -0.187	1.9% / -0.6%	0.3% / -1.0%	8.7% / -0.4%
$m = 4$	Rat. Gains/Viols	3/1	1/1	1/0	0/2	1/2	2/2	2/2	2/2	1/1
Allyn et al. [4]	EF Gains/Viols	2/1	3/1	3/3	0/2	1/3	1/2	1/0	1/1	3/3

Table 6: Performance of personalized random forests models on all datasets. We describe the metrics shown for each model and dataset in Table 1.