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

Vinith M. Suriyakumar¹ Marzyeh Ghassemi¹* Berk Ustun²*

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.

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

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 than 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].

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 than the groups who provide personal data [see e.g., 48, 79].

When Personalization Harms Performance

<table>
<thead>
<tr>
<th>Group</th>
<th>Size</th>
<th>Training Error</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>female, &lt;30</td>
<td>48</td>
<td>38.1%</td>
<td>26.8%</td>
</tr>
<tr>
<td>male, &lt;30</td>
<td>49</td>
<td>23.9%</td>
<td>26.7%</td>
</tr>
<tr>
<td>female, 30 to 60</td>
<td>304</td>
<td>30.3%</td>
<td>29.1%</td>
</tr>
<tr>
<td>male, 30 to 60</td>
<td>447</td>
<td>15.4%</td>
<td>15.2%</td>
</tr>
<tr>
<td>female, 60+</td>
<td>123</td>
<td>19.3%</td>
<td>21.9%</td>
</tr>
<tr>
<td>male, 60+</td>
<td>181</td>
<td>11.0%</td>
<td>8.2%</td>
</tr>
<tr>
<td>Total</td>
<td>1,152</td>
<td>20.4%</td>
<td>19.4%</td>
</tr>
</tbody>
</table>

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.

2. Fair Use Conditions

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

Preliminaries  We start with a dataset $(x_i, y_i, g_i)_{i=1}^n$, where example $i$ consists of a feature vector $x_i = [x_{i,1}, \ldots, x_{i,d}] \in \mathbb{R}^d$, a label $y_i \in \mathcal{Y}$, and $k$ categorical group attributes $g_i = [g_{i,1}, \ldots, g_{i,k}] \in \mathcal{G}_1 \times \ldots \times \mathcal{G}_k = \mathcal{G}$. We refer to $g_i$ as the group membership of person $i$. For example, a female over 60 would have $g_i = \{\text{female}, \text{age} \geq 60\}$. We let $n_g := |\{i | g_i = g\}|$ denote the size of group $g$,
and \( m := |G| \) denote the number of intersectional groups.

We use the data to train a \textit{personalized} model with group attributes \( h : \mathcal{X} \times G \to \mathcal{Y} \); and a \textit{generic} model that does not \( h_0 : \mathcal{X} \to \mathcal{Y} \). We train all models via ERM with a loss function \( \ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+ \), 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 \( (x_i, y_i, g_i)_{i=1}^n \) and without them \( (x_i, y_i)_{i=1}^n \):

\[
\exists \ h_0 \in \text{argmin}_h R(h) \quad \exists \ h_0 \in \text{argmin}_h \hat{R}(h)
\]

Here, \( \mathcal{H} \) and \( \mathcal{H}_0 \) denote the class of personalized models and generic models respectively.

We measure the \textit{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 \( g \) 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 \( g' \). Given a personalized model \( h \), we denote its \textit{empirical risk} and \textit{true risk} for group \( g \) when they report \( g' \) as:

\[
\hat{R}_g(h_{g'}) := \frac{1}{n_g} \sum_{i: g_i = g} \ell(h(x_i, g'), y_i) \\
R_g(h_{g'}) := \mathbb{E}_{y \sim \mathcal{Y}}[\ell(h(x, g'), y) \mid G = g] .
\]

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

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 \( g \) between \( h \) and \( h' \) using the \textit{gain} measure:

\[
\Delta_g(h, h') := R_g(h') - R_g(h).
\]

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].

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

\textbf{Definition 1.} A personalized model \( h : \mathcal{X} \times G \to \mathcal{Y} \) guarantees the \textit{fair use of group attributes} \( G \) if it obeys:

\[
\Delta_g(h_g, h_0) \geq 0 \quad \text{for all groups } g \in G, \quad (1) \\
\Delta_g(h_g, h_{g'}) \geq 0 \quad \text{for all groups } g, g' \in G \quad (2)
\]

These conditions are collective in that performance is measured over individuals in a group. Here, condition (1) ensures \textit{rationality} for group \( g \) – i.e., that a majority of group \( g \) prefers a personalized model \( h_g \) to a generic model \( h_0 \). Condition (2) ensures \textit{envy-freeness} for group \( g \) – i.e., that majority of group \( g \) 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].

\textbf{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].

\textbf{Proposition 2.} Consider a prediction model where each person reports their group membership to a personalized model \( h : \mathcal{X} \times G \to \mathcal{Y} \) in deployment. Denote the reported group membership of person \( i \) as:

\[
r_i = g_i \quad \Leftrightarrow \ i \text{ reports truthfully} \\
r_i \in G \setminus \{g_i\} \quad \Leftrightarrow \ i \text{ misreports} \\
r_i = ? \quad \Leftrightarrow \ i \text{ withholds}
\]

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:

\[
g_i \in \text{argmin}_{r_i \in G \cup \{?\}} \mathbb{E}_{y \sim \mathcal{Y}}[\ell(h(x, r_i), y_i) \mid G = g_i] .
\]

Truthful self-reporting incentives reflect basic principles regarding \textit{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.

\textbf{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:

\[ H_0 : R_g(h_0) - R_g(h_g) \leq 0 \]
\[ H_A : R_g(h_0) - R_g(h_g) > 0 \]

Here, the null hypothesis \( H_0 \) assumes that group \( g \) prefers \( h_g \) to \( h_0 \). Thus, we would reject \( H_0 \) when there is enough evidence to support a rationality violation for \( g \) 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 \( m(m-1) \) 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 \( G = \{A, B, C\} \) with 1 positive and 1 negative example in which a Bayes optimal classifier \( h : \mathcal{X} \times G \rightarrow \mathcal{Y} \) should assign a personalized intercept to each group and a personalized slope to group \( B \):

\[
h(x, g) = \begin{cases} 
\text{sign}(t_A + w^T x) & \text{if } g = A \\
\text{sign}(t_B + w^T_B x) & \text{if } g = B \\
\text{sign}(t_C + w^T x) & \text{if } g = C 
\end{cases}
\]

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:

\[
h(x, g) = \begin{cases} 
\text{sign}(t_A + w^T x) & \text{if } g = A \\
\text{sign}(t_B + w^T x) & \text{if } g = B \\
\text{sign}(t_C + w^T x) & \text{if } g = C 
\end{cases}
\]
When Personalization Harms Performance

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 $x_1 \in \{0, 1\}$ or $x_2 \in \{0, 1\}$. 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 $\mathcal{G} = \{A, B\}$ choosing between $x_1$ or $x_2$ to minimize the overall error rate.

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 $\text{ERM } h \in \arg \min_{h \in \mathcal{H}} R(h)$, and evaluating its gains to personalization with respect to a generic model $h_0 \in \arg \min_{h \in \mathcal{H}_0} R(h)$ where $\mathcal{H}_0 \subseteq \mathcal{H}$. 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:

$$\hat{R}_g(h) = \hat{R}_g(h_g) \text{ for all groups } g \in \mathcal{G}.$$
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 \( h : \mathcal{X} \times \mathcal{G} \rightarrow \mathcal{Y} \) that ensures rationality and envy-freeness for group \( g \) in terms of empirical risk. Denote the empirical gains in rationality and envy-freeness for group \( g \) as:

\[
\hat{\Delta}_g := \Delta_g(h_g, h_0), \quad \hat{\gamma}_g := \min_{g' \in \mathcal{G} \setminus \{g\}} \Delta_g(h_g, h_{g'}).
\]

If \( \hat{\Delta}_g > 0 \), then rationality for group \( g \) generalizes with probability at least \( 1 - \delta \) as long as:

\[
n_g \geq \frac{4D \log \left( \frac{2n_g}{\epsilon_g^2} + 1 \right) + \log \left( \frac{8}{\delta} \right)}{\hat{\Delta}_g^2}.
\]

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

\[
n_g \geq \frac{4D \log \left( \frac{2n_g}{\epsilon_g^2} + 1 \right) + \log \left( \frac{8mn}{\delta} \right)}{\hat{\gamma}_g^2}.
\]

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](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 \( g_i = g \).

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-
When Personalization Harms Performance

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 "interventions" 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 personalization, 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 \( n = 2,006 \) 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 \( \text{sex} \in \{\text{male, female}\} \) and \( \text{race} \in \{\text{white, black, other}\} \).

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: \((\text{male, black})\) and \((\text{male, other})\).

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., \((\text{male, black})\) and \((\text{male, other})\). Checking envy-freeness reveals that groups expect better performance by misreporting group membership – e.g., \((\text{male, other})\) 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 \((\text{male, black})\) and \((\text{male, other})\) experience a fair use violation in terms of error, we could soliciting race for all \( \text{male} \) patients and reduce test error by 1\% (as the loss in accuracy for \((\text{white, male})\) are offset by the gain in accuracy for \((\text{male, black})\) and \((\text{male, other})\).

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 \( h_{\text{dcp}} \) – 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...
When Personalization Harms Performance

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 reducing 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.

<table>
<thead>
<tr>
<th>Group</th>
<th>Test Error</th>
<th>Test AUC</th>
<th>Test ECE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g )</td>
<td>( \Delta g )</td>
<td>( \Delta g )</td>
<td>( \Delta g )</td>
</tr>
<tr>
<td>female, black</td>
<td>55.5%</td>
<td>3.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>female, white</td>
<td>21.9%</td>
<td>2.0%</td>
<td>2.0%</td>
</tr>
<tr>
<td>female, other</td>
<td>20.4%</td>
<td>6.6%</td>
<td>9.1%</td>
</tr>
<tr>
<td>male, black</td>
<td>29.4%</td>
<td>-2.7%</td>
<td>15.6%</td>
</tr>
<tr>
<td>male, white</td>
<td>21.9%</td>
<td>8.1%</td>
<td>3.7%</td>
</tr>
<tr>
<td>male, other</td>
<td>25.3%</td>
<td>-1.9%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Total</td>
<td>27.1%</td>
<td>1.4%</td>
<td>-</td>
</tr>
</tbody>
</table>

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 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.

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.

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.
References


When Personalization Harms Performance


When Personalization Harms Performance


### A. Notation

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

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d})$</td>
<td>feature vector of example $i$</td>
</tr>
<tr>
<td>$y_i \in \mathcal{Y}$</td>
<td>label of example $i$</td>
</tr>
<tr>
<td>$g_i \in {g_{i,1}, g_{i,2}, \ldots, g_{i,k}}$</td>
<td>group membership of example $i$</td>
</tr>
<tr>
<td>$\mathcal{G} = \mathcal{G}_1 \times \mathcal{G}_2 \times \ldots \times \mathcal{G}_k$</td>
<td>space of group attributes</td>
</tr>
<tr>
<td>$m =</td>
<td>\mathcal{G}</td>
</tr>
<tr>
<td>$n_g := \sum 1[g_i = g]$</td>
<td>number of examples of group $g \in \mathcal{G}$</td>
</tr>
<tr>
<td>$n_g^+ := \sum 1[g_i = g, y_i = +1]$</td>
<td>number of examples of group $g \in \mathcal{G}$ with $y_i = +1$</td>
</tr>
<tr>
<td>$n_g^- := \sum 1[g_i = g, y_i = -1]$</td>
<td>number of examples of group $g \in \mathcal{G}$ with $y_i = -1$</td>
</tr>
<tr>
<td>$h : \mathcal{X} \times \mathcal{G} \to \mathcal{Y}$</td>
<td>personalized model</td>
</tr>
<tr>
<td>$\mathcal{H}$</td>
<td>hypothesis class of personalized models</td>
</tr>
<tr>
<td>$h_g : \mathcal{X} \times \mathcal{G} \to \mathcal{Y}$</td>
<td>personalized classifier where group membership is reported truthfully as $g$</td>
</tr>
<tr>
<td>$h_0 : \mathcal{X} \to \mathcal{Y}$</td>
<td>generic model</td>
</tr>
<tr>
<td>$\mathcal{H}_0$</td>
<td>hypothesis class of generic models</td>
</tr>
<tr>
<td>$R_g(h_{g'})$</td>
<td>true risk of model $h_0$ of group $g$ if they report $g'$</td>
</tr>
<tr>
<td>$\hat{R}<em>g(h</em>{g'})$</td>
<td>empirical risk of model $h$ of group $g$ if they report $g'$</td>
</tr>
<tr>
<td>$\Delta_g(h,h')$</td>
<td>gain (i.e., reduction in true risk) for group $g$ when using $h$ instead of $h'$</td>
</tr>
<tr>
<td>$\Delta_g(h,g_0)$</td>
<td>rationality gap for group $g$ under model $h$</td>
</tr>
<tr>
<td>$\Delta_g(h,g_{g'})$</td>
<td>envyfreeness gap for group $g$ under model $h$.</td>
</tr>
</tbody>
</table>

Table 3: Notation
When Personalization Harms Performance

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](image)

**Figure 4:** Fair use violations resulting from empirical risk minimization with a surrogate loss function. We consider a classification task with two features \( x = (x_1, x_2) \) and one group attribute \( g \in \{A, B\} \) in which we fit a linear SVM \( h_g \) but evaluate 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 \( (h_0; \text{grey}) \) 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 \( x_O \). 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)

<table>
<thead>
<tr>
<th>Group</th>
<th>Training Data</th>
<th>Data Distribution</th>
<th>Predictions</th>
<th>Observed Performance</th>
<th>True Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( g_1 )</td>
<td>( g_2 )</td>
<td>( n^+ )</td>
<td>( n^- )</td>
<td>( n^+ )</td>
</tr>
<tr>
<td>0 0</td>
<td>0</td>
<td>65</td>
<td>60</td>
<td>130</td>
<td>120</td>
</tr>
<tr>
<td>1 0</td>
<td>60</td>
<td>65</td>
<td>120</td>
<td>130</td>
<td>+</td>
</tr>
<tr>
<td>0 1</td>
<td>60</td>
<td>65</td>
<td>130</td>
<td>120</td>
<td>+</td>
</tr>
<tr>
<td>1 1</td>
<td>70</td>
<td>55</td>
<td>140</td>
<td>110</td>
<td>+</td>
</tr>
<tr>
<td>Total</td>
<td>255</td>
<td>245</td>
<td>520</td>
<td>480</td>
<td>245</td>
</tr>
</tbody>
</table>

**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.
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 $h_0(x) = h(x, g)$ for all $x \in X$ and all groups $g \in G$.

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 $h \in \arg\min_{h \in H} \hat{R}(h)$, and evaluating its gains to personalization with respect to a generic model $h_0 \in \arg\min_{h \in H_0} \hat{R}(h)$ where $H_0 \subseteq H$. The personalized model $h$ obeys fair use in terms of empirical risk so long as:

$$\hat{R}_g(h) = R_g(h_g)$$

for all groups $g \in G$.

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

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

$$\hat{\epsilon}_g := \Delta_g(h_g, h_0), \quad \hat{\gamma}_g := \min_{g' \in G \setminus \{g\}} \Delta_g(h_g, h_{g'})$$

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

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

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

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

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.
We now consider the bound on envy-freeness Eq. (4). Given that

\[ 1 - \delta > 0, \]

the corresponding gap in true risk will satisfy \( \Delta_g(h_a, h_b) > 0 \) with probability at least \( 1 - \delta \) so long as:

\[
\sqrt{\frac{4D \left( \log \frac{2n_g}{D} + 1 \right) + \log \left( \frac{8}{\delta} \right)}{n_g}} \leq \hat{\Delta}_g(h_a, h_b).
\] (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 \( h \in \mathcal{H} \) from hypothesis class \( \mathcal{H} \) with VC-dimension \( D \), and any \( \delta > 0 \), the generalization error of \( h \) on group \( g \) with \( n_g \) will obey the following inequality [64] with probability at least \( 1 - \frac{\delta}{2} \):

\[
\left| \hat{R}_g(h) - R_g(h) \right| \leq \sqrt{\frac{D \left( \log \frac{2n_g}{D} + 1 \right) + \log \frac{8}{\delta}}{n_g}}.
\] (6)

We denote the quantity on the right hand side of Eq. (6) as the bounding function \( B(n_g, \mathcal{H}, \delta) := \sqrt{\frac{D \left( \log \frac{2n_g}{D} + 1 \right) + \log \frac{8}{\delta}}{n_g}} \).

Given the bounding function \( B(n_g, \mathcal{H}, \delta) \), Lemma 6 states that for any \( \delta > 0 \), with probability at least \( 1 - \delta \),

\[
2B(n_g, \mathcal{H}, \delta) \leq \hat{\Delta}_g(h_a, h_b) \implies R_g(h_b) - R_g(h_a) \geq 0
\]

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 \( 2B(n_g, \mathcal{H}, \delta) \leq \hat{\Delta}_g(h_a, h_b) \). Then we can observe that the right hand side is bounded as follows:

\[
R_g(h_b) - R_g(h_a) = R_g(h_b) - \hat{R}_g(h_a) + \hat{R}_g(h_a) - \hat{R}_g(h_a) + \hat{R}_g(h_b) - \hat{R}_g(h_b) \\
\geq -B(n_g, \mathcal{H}, \delta) - B(n_g, \mathcal{H}, \delta) \quad \because \hat{\Delta}_g(h_a, h_b) := \Delta_g(h_a, h_b) \\
\geq -2B(n_g, \mathcal{H}, \delta) + \hat{\Delta}_g(h_a, h_b) \\
\geq 0
\]

Thus we have that \( R_g(h_b) - R_g(h_a) \geq 0 \) whenever \( 2B(n_g, \mathcal{H}, \delta) \leq \hat{\Delta}_g(h_a, h_b) \). This completes the proof. \( \square \)

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 model \( h_g \) and the generic model \( h_0 \) to obtain:

\[
\sqrt{\frac{4D \left( \log \frac{2n_g}{D} + 1 \right) + \log \left( \frac{8}{\hat{\epsilon}_g} \right)}{n_g}} \leq \hat{\epsilon}_g
\]

\[
n_g \geq \frac{4D \left( \log \frac{2n_g}{D} + 1 \right) + \log \left( \frac{8}{\hat{\epsilon}_g} \right)}{\hat{\epsilon}_g^2}
\]

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 \( g, g' \in G \). This produces \( m - 1 \) preferences to generalize. Given that \( m - 1 \leq m \), we apply Lemma 6 with probability \( 1 - \frac{\gamma}{m} \). Doing so and inverting for \( n_g \) proves the result.

\[
\sqrt{\frac{4D \left( \log \frac{2n_g}{D} + 1 \right) + \log \left( \frac{8m}{\hat{\gamma}_g} \right)}{n_g}} \leq \hat{\gamma}_g
\]

\[
n_g \geq \frac{4D \left( \log \frac{2n_g}{D} + 1 \right) + \log \left( \frac{8m}{\hat{\gamma}_g} \right)}{\hat{\gamma}_g^2}
\]

\( \square \)
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.

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

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 includes 26 features that cover information that is readily available in an electronic health record (e.g. BMI, comorbidities, 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 database [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($y = +1$) = 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%.
We present tables that summarize the gains of personalization for neural networks and random forests. The following tables show additional results showing the gains of personalization when training personalized neural nets and random forests.

### 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metrics</th>
<th>Test Error</th>
<th>Test AUC</th>
<th>Test IVE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1Hot</td>
<td>All</td>
<td>DCP</td>
</tr>
<tr>
<td>apnea</td>
<td>Personalized</td>
<td>35.0%</td>
<td>48.4%</td>
<td>41.5%</td>
</tr>
<tr>
<td>n = 1152, d = 26</td>
<td>Gain</td>
<td>-1.7%</td>
<td>-15.1%</td>
<td>-8.1%</td>
</tr>
<tr>
<td>$G = {a, o, s}$</td>
<td>Best/Worst Gain</td>
<td>15.7% / -4.5%</td>
<td>-6.1% / -34.4%</td>
<td>-2.2% / -50.5%</td>
</tr>
<tr>
<td>m = 6</td>
<td>Rat. Gains/Vio</td>
<td>4/3</td>
<td>66/6</td>
<td>66</td>
</tr>
<tr>
<td>Johnson et al. [82]</td>
<td>EF Gains/Vio</td>
<td>3/3</td>
<td>0/0</td>
<td>2/1</td>
</tr>
<tr>
<td>cardio_aici</td>
<td>Personalized</td>
<td>31.5%</td>
<td>31.8%</td>
<td>36.6%</td>
</tr>
<tr>
<td>n = 1141, d = 49</td>
<td>Gain</td>
<td>1.6%</td>
<td>1.3%</td>
<td>-3.5%</td>
</tr>
<tr>
<td>$G = {a, o, s}$</td>
<td>Best/Worst Gain</td>
<td>8.4% / -0.5%</td>
<td>5.5% / -1.5%</td>
<td>0.0% / -10.3%</td>
</tr>
<tr>
<td>m = 4</td>
<td>Rat. Gains/Vio</td>
<td>0/0</td>
<td>2/1</td>
<td>3/3</td>
</tr>
<tr>
<td>Pollard et al. [75]</td>
<td>EF Gains/Vio</td>
<td>28/30</td>
<td>2/2</td>
<td>22/2</td>
</tr>
<tr>
<td>cardio_mimic</td>
<td>Personalized</td>
<td>23.7%</td>
<td>24.0%</td>
<td>23.9%</td>
</tr>
<tr>
<td>n = 5209, d = 49</td>
<td>Gain</td>
<td>0.6%</td>
<td>0.2%</td>
<td>-0.4%</td>
</tr>
<tr>
<td>$G = {a, o, s}$</td>
<td>Best/Worst Gain</td>
<td>2.0% / -1.1%</td>
<td>2.3% / -2.4%</td>
<td>1.4% / -3.1%</td>
</tr>
<tr>
<td>m = 4</td>
<td>Rat. Gains/Vio</td>
<td>1/1</td>
<td>2/2</td>
<td>2/2</td>
</tr>
<tr>
<td>Johnson et al. [49]</td>
<td>EF Gains/Vio</td>
<td>1/1</td>
<td>0/0</td>
<td>3/3</td>
</tr>
<tr>
<td>novelty</td>
<td>Personalized</td>
<td>58.0%</td>
<td>28.1%</td>
<td>36.2%</td>
</tr>
<tr>
<td>n = 181, d = 26</td>
<td>Gain</td>
<td>1.3%</td>
<td>25.0%</td>
<td>13.2%</td>
</tr>
<tr>
<td>$G = {a, o, s}$</td>
<td>Best/Worst Gain</td>
<td>12.0% / -12.8%</td>
<td>29.7% / 16.6%</td>
<td>28.1% / 7.1%</td>
</tr>
<tr>
<td>m = 4</td>
<td>Rat. Gains/Vio</td>
<td>2/1</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td>Detrano et al. [24]</td>
<td>EF Gains/Vio</td>
<td>2/1</td>
<td>2/1</td>
<td>3/1</td>
</tr>
<tr>
<td>kidney</td>
<td>Personalized</td>
<td>29.5%</td>
<td>31.7%</td>
<td>30.9%</td>
</tr>
<tr>
<td>n = 2066, d = 78</td>
<td>Gain</td>
<td>2.3%</td>
<td>4.5%</td>
<td>-3.7%</td>
</tr>
<tr>
<td>$G = {a, o, s}$</td>
<td>Best/Worst Gain</td>
<td>1.2% / -7.8%</td>
<td>5.2% / -6.8%</td>
<td>-1.6% / -16.3%</td>
</tr>
<tr>
<td>m = 6</td>
<td>Rat. Gains/Vio</td>
<td>5/4</td>
<td>5/5</td>
<td>6/6</td>
</tr>
<tr>
<td>Zhang et al. [39]</td>
<td>EF Gains/Vio</td>
<td>1/0</td>
<td>1/0</td>
<td>4/4</td>
</tr>
<tr>
<td>mortality</td>
<td>Personalized</td>
<td>20.4%</td>
<td>21.6%</td>
<td>17.7%</td>
</tr>
<tr>
<td>n = 25866, d = 468</td>
<td>Gain</td>
<td>0.1%</td>
<td>-1.1%</td>
<td>2.8%</td>
</tr>
<tr>
<td>$G = {a, o, s}$</td>
<td>Best/Worst Gain</td>
<td>5.2% / -1.7%</td>
<td>-0.6% / -3.2%</td>
<td>12.9% / 0.0%</td>
</tr>
<tr>
<td>m = 6</td>
<td>Rat. Gains/Vio</td>
<td>2/2</td>
<td>6/6</td>
<td>0/0</td>
</tr>
<tr>
<td>Johnson et al. [49]</td>
<td>EF Gains/Vio</td>
<td>5/1</td>
<td>2/2</td>
<td>6/6</td>
</tr>
<tr>
<td>sago</td>
<td>Personalized</td>
<td>33.9%</td>
<td>22.5%</td>
<td>48.9%</td>
</tr>
<tr>
<td>n = 7797, d = 36</td>
<td>Gain</td>
<td>7.7%</td>
<td>39.0%</td>
<td>12.7%</td>
</tr>
<tr>
<td>$G = {a, o, s}$</td>
<td>Best/Worst Gain</td>
<td>13.1% / 0.0%</td>
<td>54.8% / 1.4%</td>
<td>22.0% / 0.0%</td>
</tr>
<tr>
<td>m = 4</td>
<td>Rat. Gains/Vio</td>
<td>2/0</td>
<td>1/1</td>
<td>1/1</td>
</tr>
</tbody>
</table>

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 of AUC or risk calibration. We observe this effect in the Table below. For example, using an intersection 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.
Table 6: Performance of personalized random forests models on all datasets. We describe the metrics shown for each model and dataset in Table 1.