Participatory Systems for Personalized Prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Machine learning models often request personal information from individuals to assign more accurate predictions across a heterogeneous population. Personalized models are not built to support *informed consent*: individuals cannot opt out of providing personal data, nor understand its effects on their predictions. In this work, we introduce a family of personalized prediction models called *participatory systems* that support informed consent. Participatory systems are interactive prediction models that let individuals opt into reporting additional personal data at prediction time, and inform them about how their data will affect their predictions. We present a model-agnostic approach for supervised learning where personal data is encoded as "group" attributes (e.g., sex, age group, HIV status). Given a pool of user-specified models, our approach can create participatory systems that differ in their training requirements and opportunities for informed consent. We conduct a comprehensive empirical study of participatory systems in clinical prediction tasks and compare them to common approaches for personalization. Our results show that our approach can produce participatory systems that exhibit large improvements in the privacy, fairness, and performance at the population and group level.

1 Introduction

Machine learning models are routinely 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 *personalized*, in that they solicit individuals for their personal data to assign more accurate predictions [1]. In the simplest, most common approach, models are personalized using *group attributes* – i.e., categorical features that encode personal characteristics. For example, models for clinical decision support include group attributes that are *protected* [e.g., `sex` 2], *sensitive* [e.g., `HIV status` 3, 4], *self-reported* [e.g., `hours_of_sleep` 2], or *costly* in that they can only be acquired with time, money, or effort [e.g., `tumor_severity` as detected via CT scan 5 or biopsy 6].

Digital platforms that solicit personal data from their individuals are designed to support *informed consent*: individuals can opt out of providing their personal data, and can see how their data will be used to support their decision [see e.g., GDPR consent banners 7, 8]. In contrast, personalized models do not provide such functionality: individuals cannot "opt-out" of reporting their personal data to a personalized model, nor tell if a model is using it to improve their predictions. This lack of functionality is alarming as standard techniques for personalization do not improve performance across all individuals who provide personal data [see 9]. In practice, a personalized model might perform worse (or just as well) as a *generic model* that did not solicit personal data for individuals with a specific personal characteristics (see Fig. 1) . In such instances, personalized models violate the promise of personalization – as individuals in this group report their personal data without receiving a tailored gain in performance in return. *These effects are prevalent, hard to detect, and hard to fix* [9] –

					Traditional groups assigned predictions from h_g			Participatory groups opt-in to predictions from personalized model h_g		
Group	Data		Predictions		Mistakes		Gain	Model Assigned	Data Reported	Gain
$\boldsymbol{g}$	$n_{\boldsymbol{g}}^{+}$	$n_{\boldsymbol{g}}^{-}$	h_0	$h_{\boldsymbol{g}}$	$R_{\boldsymbol{g}}(h_0)$	$R_{\boldsymbol{g}}(h_{\boldsymbol{g}})$	$\Delta R_{\boldsymbol{g}}(h_{\boldsymbol{g}}, h_0)$	$h_{\boldsymbol{r}}$	$\boldsymbol{r}$	$\Delta R_{\boldsymbol{g}}(h_{\boldsymbol{r}}, h_0)$
female, old	0	24	−	+	0	24	−24	h_0	∅	0
female, young	25	0	−	+	25	0	25	$h_{\boldsymbol{g}}$	sex, age	25
male, old	25	0	−	+	25	0	25	$h_{\boldsymbol{g}}$	sex, age	25
male, young	0	27	−	−	0	0	0	h_0	∅	0
all	50	51	−	−	50	24	26	$h_{\boldsymbol{g}}$	sex, age	50

Figure 1: Simple classification task where participation improves performance and reduces data use. We are given $n^+ = 50$ positive examples and $n^- = 51$ negative examples for 4 groups defined by the attributes sex × age. We fit the best linear model with a one-hot encoding of group attributes $h_{\boldsymbol{g}}$, and evaluate the gains of personalization with respect to the best linear model without group attributes h_0. In a traditional setup (left), individuals would are forced to report group membership to personalized model $h_{\boldsymbol{g}}$. Here, a personalized model $h_{\boldsymbol{g}}$ reduces overall error from 50 to 24, but assigns predictions that are identical to [male, young] and that are unnecessarily inaccurate to [female, old]. In a minimal participatory system (right), individuals opt in to receive predictions from h_0 or $h_{\boldsymbol{g}}$. This setup improves performance at the group and population levels – individuals in group [female, old] would opt for predictions from h_0, improving their gain from -24 to 0 and the overall gain from 26 to 50.

underscoring the need to let individuals opt out of personalization, and to understand its effects for people like themselves.

In this paper, we propose a new family of prediction models that operationalize these basic principles of responsible personalization. We call these systems *participatory systems* – i.e., interactive machine learning models that let individuals report additional personal data to improve their performance at prediction time. We propose a *model-agnostic* approach for prediction tasks where personal data is encoded in group attributes. Our approach starts with a user-specified pool of personalized models, which it carefully arranges within a *reporting tree* – i.e., a tree that represents the sequence of reporting decisions for a user (see Fig. 2). The resulting architecture: (1) lets individuals opt out of reporting some or all personal data; (2) provides information to support this decision (e.g., expected performance gains; change in prediction); (3) ensures that reporting data leads to an expected gain in performance. In practice, this approach has three major benefits:

Performance & Fairness: Our approach builds participatory systems that assign personalized predictions using multiple models. This architecture can use personal data in a way that produces large performance gains for each reporting group (i.e., individuals who report a specific subset of personal characteristics). In settings with heterogeneous data distributions, we can avoid performance trade-offs imposed by a single model, and further improve performance by assigning predictions to each group using a personalized model built for that group.

Privacy & Harm Mitigation: Participatory systems naturally mitigate harm while promoting privacy since models that allow individuals to participate must incentivize participation. In this setup, individuals who are informed as to the gains of personalization will opt out of reporting personal data if it reduces performance (see Fig. 1). Moreover, systems can be "pruned" to avoid soliciting personal data from individuals who would not report it – thus promoting privacy via data minimization.

Flexibility: Our approach can produce three kinds of participatory systems, providing practitioners with multiple options to support informed consent (see Fig. 2). These include: (1) a minimal system, which allows individuals to opt out of an existing personalized model by training one additional model (i.e., a generic model); (2) a flat system, which allows individuals to opt into partial personalization, and further improves personalization using a specific model for each reporting group; (3) a sequential system, which allows individuals to opt into partial personalization by reporting each piece of personal data, and also improve personalization using a specific model for each reporting group.

Contributions The main contributions of this work are:

1. We introduce a new kind of prediction model that can support informed consent.
2. We develop a model-agnostic approach to built a variety of participatory systems that operationalize informed consent under different training and implementation requirements.
3. We conduct a comprehensive empirical study of participatory systems on real-world clinical applications, showing how participation can improves performance and data privacy.

4. We provide a Python package to develop and evaluate participatory personalization systems, available at: https://anonymous.4open.science/r/psc_public-164C/

2 Related Work

Data Privacy & Consent Participatory systems support key principles of responsible data use articulated in modern legislation – see e.g., guidelines in the OECD [10], GDPR [7], and California Consumer Privacy Act of 2018 [11]. These include principles like *collection limitation* (i.e., data should be collected with the consent of a data subject, and restricted to only what is necessary), and *purpose specification* (i.e., the purpose of data collection should be disclosed to individuals). A substantial body of work motivates the need for such functionality from the perspective of data subjects. For example, recent work shows that individuals care deeply about their ability to control personal data [12, 13, 14], that individual preferences with regards to sharing personal data varies considerably [15, 16], and that individuals face different costs in collecting, disclosing, or leaking information [17, 18, 19, 20, 21]. In effect, these findings show that we should not assume that data subjects would consent to sharing their personal data even in settings with legal protections [see e.g, 22, who show that underrepresented groups do not consent to report their demographic data in clinical settings].

Personalization We study personalization for prediction models with *group attributes* – i.e., categorical attributes encode personal characteristics. There is an extensive body of literature on modeling with categorical data [see e.g., 23, 24, 25], as well as stream of research on new techniques to make use of categorical attributes – e.g., methods to train models with higher-order interaction effects [26, 27, 28] or recursively partitioning data [29, 30, 31, 32]. The use of group attributes in such settings often stems from the belief that personalization can only improve performance. However, few works evaluate the gains from personalization and those that do often measure the gains at a population level rather than a group level [33, 34].

Algorithmic Fairness Our work is broadly related to research in algorithmic fairness in that we are interested in building predictive models that perform well across groups.

Participatory systems are designed for applications where models use group attributes to assign more accurate predictions over a heterogeneous population [e.g., clinical decision support and precision medicine; 35, 36, 37]. Several works discuss the need for models to account for group membership in this setting [see e.g., 24, 38, 39, 40, 41, 42, 43, 44], noting that it is otherwise impossible for a model to perform equally well for all groups.

Participatory systems provide a way to ensure the “fair use" of group attributes through mechanism design [9, 39]. Fair use conditions are preference-based notions of group fairness that incentivize truthful self-reporting for groups who report personal data in our setting [see e.g., 44, 45, 46, 47, for other preference-based notions of fairness]. These conditions differ from the traditional goal of equalizing performance across groups [see 39, 44, for a discussion]. The latter goal – *parity* – is an ill-suited for personalization because methods to achieve parity can equalize performance by reducing performance for groups who perform well, rather than by improving performance for groups who perform poorly [48, 49, 50, 51].

3 Participatory Systems

Preliminaries We consider a supervised learning task where categorical attributes encode personal characteristics. We start with a dataset of examples $(\boldsymbol{x}_i, y_i, \boldsymbol{g}_i)_{i=1}^n$ from n individuals where each example consists of d features $\boldsymbol{x}_i = [x_{i,1}, \ldots, x_{i,d}] \in \mathbb{R}^d$, a label $y_i \in \mathcal{Y}$, and k *group attributes* $\boldsymbol{g}_i = [g_{i,1}, \ldots, g_{i,k}] \in \mathcal{G}_1 \times \ldots \times \mathcal{G}_k = \mathcal{G}$ (e.g., $\boldsymbol{g}_i = [\texttt{female}, \texttt{HIV} = +]$). We refer to $\boldsymbol{g}_i$ as the *group membership* of individual i, and to the subset of examples $\{i \,|\boldsymbol{g}_i = \boldsymbol{g}\}$ as *group* $\boldsymbol{g}$. We let $n_{\boldsymbol{g}} := |\{i \,|\boldsymbol{g}_i = \boldsymbol{g}\}|$ denote the number of examples in group $\boldsymbol{g}$, and $m := |\mathcal{G}|$ denote the number of (intersectional) groups.

We use the dataset to train a *personalized model* $h_{\boldsymbol{g}} : \mathcal{X} \times \mathcal{G} \rightarrow \mathcal{Y}$. We denote the *empirical risk* and *true risk* of a model h as $\hat{R}(h)$ and $R(h)$, respectively. We fit the personalized model via empirical risk minimization with a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ so that $h_{\boldsymbol{g}} \in \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}(h)$.

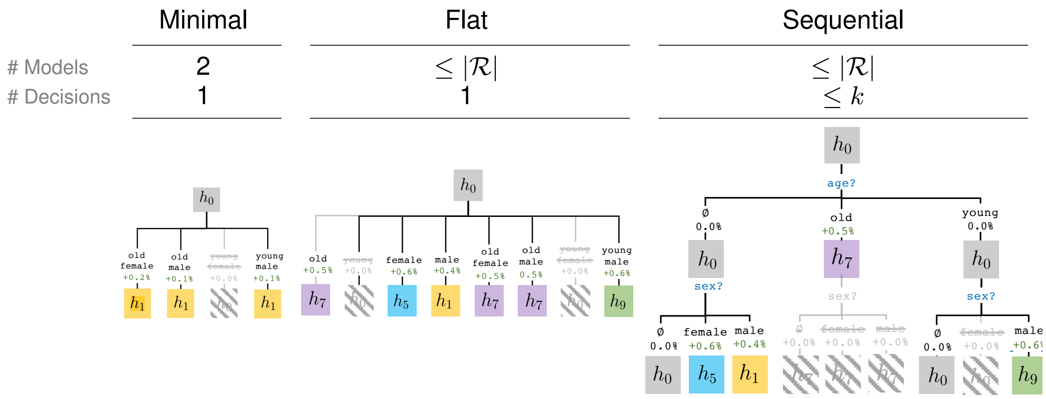

Figure 2: Participatory systems for a prediction task with $k = 2$ group attributes $\mathcal{R} = \texttt{age} \times \texttt{sex} = [\texttt{male}, \texttt{female}, \varnothing] \times [\texttt{old}, \texttt{young}, \varnothing]$. Each system allows individuals to opt out of personalization by reporting $\varnothing$, and informs their decision by revealing the gains of personalization (e.g., `+0.2%` reduction in error). Each system minimizes data use by removing reporting options that do not lead to gain (e.g., $[\texttt{young}, \texttt{female}]$ is pruned in all systems). We describe three kinds of systems with different training and implementation requirements, what individuals report, and how they report it. The minimal system allows individuals to opt into a single personalized model, while the flat and sequential systems allow for partial personalization and multiple models. In sequential systems, individuals can can make informed decisions to report each attribute.

We check that personal data improves performance for each group by comparing their performance under a personalized model $h_{\boldsymbol{g}}$ to that of a *generic model* $h_0 : \mathcal{X} \times \mathcal{Y}$ – i.e., the best model fit on a dataset without group attributes $h_0 \in \operatorname{argmin}_{h \in \mathcal{H}_0} \hat{R}(h)$.

Individuals should expect to receive tailored performance benefits in return for providing their personal data. In Definition 1, we formalize this principle in terms of collective preference guarantees.

Definition 1 (Fair Use, [9])**.** *A personalized model $h_{\boldsymbol{g}} : \mathcal{X} \times \mathcal{G} \to \mathcal{Y}$ guarantees the fair use of a group attribute $\mathcal{G}$ if it is*

$$\text{'rational' i.e.} \quad R_{\boldsymbol{g}}(h_{\boldsymbol{g}}) \leq R_{\boldsymbol{g}}(h_0) \qquad \text{for all groups } \boldsymbol{g} \in \mathcal{G}, \text{ and} \tag{1}$$

Condition (1) captures *rationality* for group $\boldsymbol{g}$: a majority of group $\boldsymbol{g}$ prefers a personalized model $h_{\boldsymbol{g}}$ to its generic counterpart h_0.

The condition is collective, in that performance is measured over individuals in a group, and weak, in that the expected performance gain is non-negative – i.e., no group will be harmed.

Fair use conditions enshrine necessary conditions for individuals would voluntarily report their true group membership to a personalized model. These conditions hold in applications where individuals prefer more accurate models. We express these preferences in terms of the *gain* $\Delta_{\boldsymbol{g}}(h, h') := R_{\boldsymbol{g}}(h') - R_{\boldsymbol{g}}(h)$, and make them explicit in Assumption 2.

Assumption 2 (Rational Preferences)**.** *Given a pair of models h and h', we assume that a group prefers to receive predictions from h to h' whenever $\Delta_{\boldsymbol{g}}(h, h') > 0$.*

Assumption 2 holds in applications where individuals prefer to receive correct predictions. These include healthcare applications, such as when estimating disease risk [52, 53, 54]. This assumption does not hold in settings where individuals may prefer to receive incorrect predictions [see e.g, “polar" clinical prediction tasks in 55]. In loan risk estimation, for example, a personalized model with group attributes could output more reliable risk predictions that are in the best interest of most individuals since they lead to lower interest rates on average. However, they may not be in the best interest of groups whose interest rates would increase as a result of reporting personal information.

Participatory Systems Participatory systems let individuals opt into personalization at prediction time. We denote an individual's choice to opt out of reporting a group attribute with $\varnothing$. We denote the *reported group membership* for individual i as $\boldsymbol{r}_i = [r_{i,1}, \ldots, r_{i,k}] \in \mathcal{R} = (\mathcal{G}_1 \cup \varnothing) \times \ldots \times (\mathcal{G}_k \cup \varnothing)$, and the number of reporting groups as $p := |\mathcal{R}|$. Thus, an individual with $\boldsymbol{g}_i = [\texttt{female}, \texttt{HIV} = +]$ who opts out of reporting their HIV status would have $\boldsymbol{r}_i = [\texttt{female}, \varnothing]$.

In Fig. 2, we show three participatory systems that operationalize informed consent.

Minimal systems let individuals opt into personalization by decide whether to receive predictions from a personalized model $h_{\boldsymbol{g}}$ or its generic model h_0. This architecture allows individuals to opt out of receiving unnecessarily inaccurate predictions from a personalized model. It is is bound to improve performance at the group and population level when individuals opt into the most accurate predictions from $h_{\boldsymbol{g}}$ or h_0, and may reduce the use of personal data (as we can avoid soliciting information if it does not lead to gain).

Flat systems let individuals opt into *partial* personalization by reporting any *subset* of their group attributes – i.e., without reporting *all* of personal data. This architecture allows individuals to withhold personal data that they are unwilling or unable to share. For example, an individual with $\boldsymbol{g}_i = [\texttt{age} \geq \texttt{50}, \texttt{HIV} = +]$ can report $\boldsymbol{r}_i = [\texttt{age} \geq \texttt{50}, \emptyset]$. Flat systems can further improve performance by assigning a distinct personalized model to each reporting group. Thus, individuals can receive personalized predictions from a model that is fit to maximize performance for individuals such as themselves.

Sequential systems let individuals opt into *partial* personalization by reporting one attribute at a time. This architecture allows individuals to make a series of k decisions to report each of k group attributes. In turn, the system guides them in their decision to report or not report each group attribute by revealing: (i) the cumulative performance gain received as a result of all reporting decisions thus far; (ii) the range of additional gains in future steps. Sequential systems are well-suited for settings with *optional* information – e.g., clinical prediction models where group attributes encode the result of an optional medical procedure [e.g., the Gleason score from a prostate biopsy procedure 5]. Thus, an individual with $\boldsymbol{g}_i = [\texttt{age} \geq \texttt{50}, \texttt{HIV} = +]$ can report `age` before deciding whether to report `HIV`.

Informing Consent

Participatory systems can inform consent by providing individuals with precise information on how their decision to provide or withhold personal data their predictions and expected performance. In general, this information will change across applications – as the content and format of this information will depend on: (1) the performance metric for the task at hand, the type of participatory system, and the numeracy and technical expertise of individuals. In an online medical diagnostic built to output accurate “yes-or-no" predictions, for example, individuals would see how opting into personalization would change their prediction and their expected change in out-of-sample error. In an online medical risk assessment built to output reliable risk predictions, individuals would see how opting into personalization changes their risk prediction and their expected change in out-of-sample calibration error.

Information shown to individuals should reflect the uncertainty in estimation [see e.g., 56, 57]. Moreover, it should be tailored to technical expertise of individuals who interact with the systems. In settings where the diagnostic is soliciting information from patients, participatory systems should be grounded in best practices from uncertainty quantification and risk communication [58, 59, 60, 61, 62]. If the patient were assisted by a physician, however, we may be able to present information that is more technical.

While our approach can provide flexibility to practitioners in how they compute and present these quantities, we cannot ensure individuals who consent are truly informed.

4 Learning Participatory Systems

We present a model-agnostic procedure to construct participatory systems in Algorithm 1. Our procedure takes as input a pool of candidate models $\mathcal{M}$ and validation dataset $\mathcal{D}$. It then constructs a specific kind of participatory system by calling three routines: (1) enumerate all possible trees (Step 1); (2) assign a model to each node within the tree (Step 3); (3) prune the trees for data minimization (Step 4). The procedure uses all three routines to construct a sequential system, only calls the assignment and pruning routine to construct other systems. In what follows, we describe the key components of our approach in greater detail.

Representation We represent the participatory systems in Fig. 2 as *reporting trees*. Each reporting tree consists of nodes that represent a personalized model assigned to a specific reporting group. The tree starts with a generic model at its root, branching out as individuals opt in or out of reporting

Algorithm 1 Learning Participatory Systems

```
   Input: $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{g}_i, y_i)\}_{i=1}^n$                          validation dataset
   Input: $\mathcal{M} : \{h : \mathcal{X} \times \mathcal{R} \to \mathcal{Y}\}$                 pool of candidate models
1: $\mathcal{T} \leftarrow$ EnumerateTrees($\mathcal{G}$)                               generate all reporting trees
2: for $T \in \mathcal{T}$ do                                                  v-ary trees of models
3:     $T \leftarrow$ AssignModels($T, \mathcal{M}$)                                  assign models based on
4:     repeat
5:         for $\boldsymbol{r} \in$ leaves($T$) do                      each tree is an ordering of reporting groups
6:             $T \leftarrow$ Prune($T, \boldsymbol{r}$)                                                 prune models
7:         end for
8:     until no leaves are pruned
9: end for
Output $\mathcal{T}$, collection of participatory systems for all reporting groups $\boldsymbol{r} \in \mathcal{R}$
```

personal data. The depth of each tree reflects the number of *reporting decisions* for an individual. A flat system, which allows individuals to make 1 opt-in/out decision, corresponds to a p-ary tree of depth 1 with $p := |\mathcal{R}|$ leaves. A sequential system, which allows individuals to make k consecutive opt-in/out decisions, corresponds to a v-ary tree with depth k where k is the number of group attributes and $v := \max_t |\mathcal{G}_t|$ is the maximum number of values for any group attribute.

Generating Candidate Models We generate a pool of personalized models can be assigned to nodes in a reporting tree. This pool should contain a generic model h_0 that can be assigned to groups who opt out of reporting all attributes. In practice, we generate the pool by fitting multiple models for each reporting option – i.e., each 2^k distinct combination of group attributes that an individual could report. The models account for group membership using different personalization techniques (e.g., a one-hot encoding of group attributes, a one-hot encoding of intersectional groups, and variants of these with first degree interaction terms). By default, we include a “decoupled model" for each reporting group that is fit using only data for that group, as such models can perform well on heterogeneous subgroups [9, 38, 39].

Enumerating Reporting Trees We design a custom algorithm for the EnumerateTrees routine in Step 1 (see Appendix B). This routine is only used for sequential systems since the reporting tree is fixed for minimal and flat systems. Our routine enumerates all k-ary trees that obey user-specified constraints on ordering and data availability. Thus, one could enforce an ordering constraint to require the trees to solicit lab tests last, allowing patients to avoid lab tests based on other personal characteristics. When used to enumerate the k-ary trees for a sequential system, it outputs *all* possible v-vary trees. For a dataset with 3 binary group attributes $\mathcal{G} = \texttt{sex} \times \texttt{age\_group} \times \texttt{blood\_type}$, $\mathcal{T}$ would contain $3^1 \times 2^3 \times 1^9 = 24$ possible 3-ary trees of depth 3. Our routine can scale to datasets with ≤ 8 group attributes, but does not scale beyond this limi. In effect, enumerating p-ary trees is intractable as the number of group attributes increases as the number of possible trees is upper bounded by $|\mathcal{T}| \leq \prod_{i=1}^k i^{v^{k-i}}$.

Assigning Models to Reporting Groups We assign each reporting group a model using the AssignModels routine in Step 3. Given a reporting group, we consider all models in the pool that require any subset of personal data that an individual could report. Thus, a group who reports `age` and `sex` could be assigned predictions from a model that requires `age`, `sex`, both, or neither. This implies that we can always assign the generic model to any reporting group, meaning that every system performs at least as well as a generic model in terms of the assignment metric. By default, we assign each reporting group a model from $\mathcal{M}$ that optimizes out-of-sample performance based on a user-specified metric (e.g., 5-CV AUC). This rule can be customized to account for other criteria based on training data (e.g., one can filter $\mathcal{M}$ so that we only consider models that generalize).

Pruning for Data Minimization Algorithm 1 may output trees where reporting group should not voluntarily report personal data. This could happen in two ways:

1. A tree could assign the same model to a pair of nested reporting groups, which would correspond to a participatory system in which a group who reports personal data receives the same predictions (see e.g., a tree that assigns a generic model to $[\texttt{female}, \emptyset]$ and $[\texttt{female}, \texttt{young}]$ in Fig. 2).

2. A tree could also assign distinct models to a pair of nested groups, which would correspond to a participatory system where a model would report personal only to receive predictions that are expected to reduce performance (see e.g., Fig. 2, where $[\texttt{female}, \texttt{young}]$ receives better performance from the generic model h_0 in the flat system).

In line 4, we Prune each tree to ensure that the corresponding participatory system does not solicit data in such cases. The routine prunes a tree where a leaf that is assigned the same model as its parent by simply checking the assignment (to ensure that the participatory system will not assign the same predictions). In addition, the routine prunes a tree where a leaf that is assigned a model that performs worse than its parent (to ensure that the participatory system only solicits data that can improve predictions). In the latter case, the decision to prune is based on a one-sided hypothesis test that checks if group $\boldsymbol{g}$ prefers the parent model h to the model at the leaf h':

$$H_0 : R_{\boldsymbol{g}}(h) \leq R_{\boldsymbol{g}}(h') \quad \text{vs.} \quad H_A : R_{\boldsymbol{g}}(h) > R_{\boldsymbol{g}}(h') \tag{2}$$

Here, the null hypothesis H_0 assumes that a group prefers the parent model h over the model at the leaf h'. Thus, we reject H_0 when there is enough evidence to suggest that h' performs better for $\boldsymbol{g}$ on a held-out dataset. The testing procedure varies based on the performance metric used to evaluate the gains of personalization. In general, we can apply a bootstrap hypothesis test [63], or choose a more powerful test for common performance metrics [see e.g., the McNemar test for accuracy 64]. In settings where we must test for gains multiple times, we can control for the false discovery rate using a standard Bonferroni correction [65], which is suitable even for non-independent tests.

Discussion Model developers can easily customize the system by swapping out the criteria used to fit a pool of candidate models, to assign models to groups, and to prune trees. This flexibility provides some ability to deal with real-world constraints when training and hosting multiple models. One can train minimal system which only requires training and hosting 1 additional model. One also train flat and sequential systems with a limited number component models to match their training constraints. In terms of scalability, the primary bottleneck in building participatory systems is data rather than computation. In a setting with $k = 20$ binary attributes, for example, we could have – at most – 2^{20} intersectional groups and $(2 + 1)^{20}$ reporting groups. Assuming 30 samples per intersectional group, we would need $\approx$ 30M samples to build a participatory system with $k = 20$ binary attributes.

5 Experiments

We benchmark participatory systems on real-world datasets for clinical decision support. Our goals are to compare participatory systems against other kinds of personalized models in terms of performance, data use, and opportunities for informed consent. We include additional results in Appendix A, and software and scripts to reproduce the results in our anonymized repository.

5.1 Setup

We consider six datasets for clinical decision support shown in Table 2 which include group attributes like sex, age group, and HIV status. We focus on clinical prediction models since they often use characteristics that should be optional (e.g., characteristics that are protected, self-reported, sensitive, or costly). We minimally process each dataset to handle missing data and repair class imbalances at the group level. We split each dataset into a training sample (60%) used to train models, a validation sample (20%) used to assign and prune models, and a test sample (20%) used to evaluate performance.

We use each dataset to train six kinds of personalized models: (1) 1Hot, a model fit with a one-hot encoding of group attributes; (2) mHot, a model fit with a one-hot encoding of intersectional groups; (3) Impute, a 1Hot model where individuals can opt out of personalization by imputing their group membership; (4) Minimal, a minimal system composed of 1Hot and its generic counterpart; (5) Flat, a flat system composed of 1Hot, mHot, and their generic counterparts; and (5) Seq: a sequential system composed of 1Hot, mHot, and their generic counterparts. We fit all models – i.e., the personalized models and the components of participatory systems – from a single hypothesis class. We report results for logistic regression below given their widespread use in clinical applications[1] , and include results for random forests in Appendix A.3.

[1] In practice, most clinical prediction models are built using logistic regression and a one-hot encoding of group attributes [see e.g., 25, 66, 67]. These simple models are well-suited for this setting since they perform

We assume that individuals report all group attributes in models that do not allow them to opt out (e.g., for 1Hot, mHot). When a model or system does allow individuals to opt out, we assume that individuals will report their group attributes when they are informed that it is will strictly improves expected performance for their reporting group (i.e., they see a positive gain for a performance metric on validation data).

Metric	**Definition**	**Description**
Overall Performance	$\sum_{\boldsymbol{g} \in \mathcal{G}} \frac{1}{n_{\boldsymbol{g}}} R_{\boldsymbol{g}}(h_{\boldsymbol{g}})$	Population-level performance of a personalized system/model. This is computed as a weighted average over all intersectional groups
Overall Gain	$\sum_{\boldsymbol{g} \in \mathcal{G}} \frac{1}{n_{\boldsymbol{g}}} (R_{\boldsymbol{g}}(h_0) - R_{\boldsymbol{g}}(h_{\boldsymbol{g}})$	Population-level gain in performance of a personalized system/model over its generic counterpart
Group Gains	$\min_{g \in \mathcal{G}} / \max_{g \in \mathcal{G}} R_{\boldsymbol{g}}(h_0) - R_{\boldsymbol{g}}(h_{\boldsymbol{g}})$	Range of group-level gains of a personalized system/model over its generic counterpart across all groups
# Violations	$H_0 : R_{\boldsymbol{g}}(h_{\boldsymbol{g}}) \leq R_{\boldsymbol{g}}(h_0)$	The number of groups who receive unnecessarily poor predictions by a personalized system/model. Each violation is a rejection of H_0 at a 10% significance level using a bootstrap hypothesis test with 100 resamples
Data Reduction	$\sum_{\boldsymbol{g} \in \mathcal{G}} \frac{1}{n_{\boldsymbol{g}}} A_{\boldsymbol{g}} / A_{h_{\boldsymbol{g}}}.$	The number of attributes that a system/model will not request from an average user. Here, $A_{h_{\boldsymbol{g}}}$ is the number of attributes requested by $h_{\boldsymbol{g}}$ for group $\boldsymbol{g}$, and $A_{\boldsymbol{g}}$ is the maximum number of attributes that $\boldsymbol{g}$ could report.
Opportunities for Informed Consent	$\sum_{\boldsymbol{g} \in \mathcal{G}} \frac{1}{n_{\boldsymbol{g}}} I_{\boldsymbol{g}} / A_{\boldsymbol{g}}$	The number of opt-in decisions that a system/model provides an average user. Here, $I_{\boldsymbol{g}}$ is the number of opt-in/out decisions that a system provides for group $\boldsymbol{g}$, and $A_{\boldsymbol{g}}$ is the maximum number of attributes that $\boldsymbol{g}$ could report.

Table 1: Caption

5.2 Results

Our results in Table 2 show that participatory systems can operationalize informed consent in ways that improve performance at both the population level and the group level. In particular, participatory systems achieve the best overall and group-level performance on 6/6 datasets. In contrast, traditional approaches perform worse, but assign unnecessarily inaccurate personalized predictions to specific groups on at least 3/6 datasets (see # violations in red). For example, on the saps dataset, we find that mHot improves Test AUC at a population level but reduces Test AUC for the worst-off group by -0.002, produce a statistically significant rationality violation. This means that at least one group would have been better off with the generic model using a hypothesis test with 10% significance. Our results for Minimal show that we can reap benefits in such cases: when a personalized model assigns unnecessarily inaccurate predictions, a minimal system that allows individuals to opt out can improve performance and reduce data collection.

well across multiple performance metrics for clinical decision support (i.e., accuracy, AUC) and generalize in small-sample regimes that arise when working with intersectional groups.

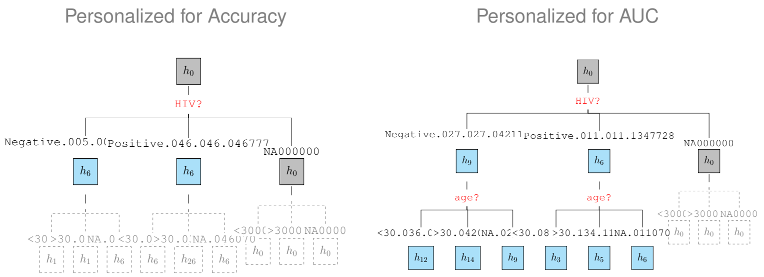

Figure 3: Sequential systems for the saps dataset optimized for error rate (left) and AUC (right). The systems differ structurally because models are assigned and pruned based on different criteria (error rate vs AUC). The left system might be suitable for diagnosis, while the right system might be suitable for prioritization in an ICU setting. The left system achieves 16.6% test error while the right system achieves 0.960 test AUC. We provide additional information about these models and others in Appendix A.

		Static		Imputed	Participatory		
Dataset	Metrics	1Hot	mHot	Impute	Minimal	Flat	Seq
`cardio_eicu` $n = 1341, d = 49$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $m = 4$ Pollard et al. [68]	Overall Performance	0.858	0.857	0.858	0.858	**0.923**	**0.923**
	Overall Gain	0.001	-0.000	0.001	0.001	**0.067**	**0.067**
	Group Gains	-0.001 – 0.002	-0.001 – 0.002	-0.001 – 0.002	-0.001 – 0.002	0.008 – 0.094	0.008 – 0.094
	# Violations	2	1	3	1	**0**	**0**
	Data Reduction	0.0%	0.0%	NA%	0.0%	50.0%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	50.0%	100.0%
`cardio_mimic` $n = 5289, d = 49$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $m = 4$ Johnson et al. [69]	Overall Performance	0.876	0.876	0.876	0.877	**0.896**	**0.896**
	Overall Gain	-0.000	-0.000	-0.000	0.000	**0.020**	**0.020**
	Group Gains	-0.000 – 0.001	-0.000 – 0.001	-0.000 – 0.001	-0.000 – 0.001	0.005 – 0.034	0.005 – 0.034
	# Violations	**0**	2	**0**	**0**	**0**	**0**
	Data Reduction	0.0%	0.0%	NA%	0.0%	37.5%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	40.0%	100.0%
`lungcancer` $n = 120641, d = 84$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $m = 6$ NCI [70]	Overall Performance	0.855	0.855	0.855	0.855	**0.861**	**0.861**
	Overall Gain	0.001	0.001	0.001	0.001	**0.007**	**0.007**
	Group Gains	-0.000 – 0.000	-0.000 – 0.000	-0.000 – 0.000	-0.000 – 0.000	0.001 – 0.012	0.001 – 0.012
	# Violations	2	2	2	1	**0**	**0**
	Data Reduction	0.0%	0.0%	NA%	0.0%	29.2%	16.7%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	35.3%	100.0%
`saps` $n = 7797, d = 36$ $\mathcal{G} = \{\texttt{HIV}, \texttt{age}\}$ $m = 4$ Allyn et al. [71]	Overall Performance	0.875	0.877	0.875	0.875	**0.960**	**0.960**
	Overall Gain	0.010	0.011	0.010	0.009	**0.095**	**0.095**
	Group Gains	-0.000 – 0.015	-0.002 – 0.019	-0.000 – 0.015	0.000 – 0.015	0.035 – 0.139	0.026 – 0.139
	# Violations	**0**	1	**0**	**0**	**0**	**0**
	Data Reduction	0.0%	0.0%	NA%	0.0%	25.0%	31.3%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	33.3%	100.0%
`sleepapnea` $n = 1152, d = 26$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $m = 6$ Ustun et al. [72]	Overall Performance	0.774	0.774	0.774	0.775	**0.850**	**0.850**
	Overall Gain	-0.002	-0.002	-0.002	-0.001	**0.074**	**0.074**
	Group Gains	-0.002 – 0.002	-0.002 – 0.003	-0.002 – 0.002	-0.002 – 0.002	0.004 – 0.115	0.004 – 0.115
	# Violations	2	3	2	1	**0**	**0**
	Data Reduction	0.0%	0.0%	NA%	0.0%	50.0%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	50.0%	100.0%
`support` $n = 9105, d = 55$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $m = 6$ Knaus et al. [73]	Overall Performance	0.707	0.706	0.707	0.706	**0.712**	**0.712**
	Overall Gain	0.002	0.001	0.002	0.001	**0.007**	**0.007**
	Group Gains	-0.000 – 0.003	-0.000 – 0.003	-0.000 – 0.003	0.000 – 0.003	-0.000 – 0.023	-0.000 – 0.023
	# Violations	**0**	**0**	**0**	**0**	**0**	**0**
	Data Reduction	0.0%	0.0%	NA%	0.0%	66.7%	33.3%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	60.0%	100.0%

Table 2: Performance, data use, and consent, of personalized models and systems for all 6 datasets. We describe addition information about the metrics and datasets in Table 1 and Appendix A.1. We provide additional results using random forests in Appendix A.

On the Benefits of Complex Participatory Architectures Our results highlight some of the benefits of using a flat or sequential system. We find that flat and sequential systems can further improve performance over minimal systems – bestowing gains ranging from 0.006 AUC on `lungcancer` to 0.085 AUC on `saps`. These systems can also solicit less personal data and provide more opportunities for consent. For example, the flat and sequential systems lead to a data reduction of 50% and 25.0% on `cardio_eicu`, meaning that they require 50% to 75% of the data solicited by a traditional system. In this dataset, sequential systems provide additional opportunities for consent (e.g., 100% compared to 50.0% for a flat system).

On the Beneficiaries of Participation The ranges of group gains suggest that most groups benefit from participatory systems, as opposed to groups who are assigned unnecessarily inaccurate predictions by a static system. For example, on $5/6$ datasets, both the worse-case and best-case gains improve for the flat system compared with the static or imputed systems. This translates to better predictions for individuals in intersectional groups defined by sex, age, and HIV status. These gains are likely a consequence of added capacity provided by the use of multiple models in the flat and sequential systems.

On the Potential for Data Reduction Our results highlight how participatory systems can reap the benefits of personalization without requiring all individuals to report personal data. In practice, the potential for data reduction varies across datasets and our choice of performance metric. In Fig. 3, we show a pair of sequential systems for the `saps` dataset. Here, a system built to optimize error has fewer nodes than one built to optimize for AUC since as can prune more nodes when we measure gains in terms of the error rate (see e.g., our results for error rate in Appendix A). In practice, this means that we can avoid requesting `age` entirely if we care about error rate.

On the Pitfalls of Imputation Imputation provides an alternative solution to build predictive models where individuals to opt out of personalization. In theory, imputation could resolve fair use violations when a harmed group is imputed the value of a group that they would have been better off reporting (e.g., in Fig. 1). Here, we impute group membership using mean imputation as an illustrative example. Our results for Impute demonstrate the potential pitfalls of this approach. Although the imputed system does not introduce additional fair use violations and maintains performance across all datasets, we still observe fair use violations on $3/6$ datasets. This suggests that limiting the system to a single model, even with careful imputation, may not achieve the capacity required to mitigate fair use violations.

6 Concluding Remarks

This work describes methods for building participatory systems and demonstrates their benefits on real-world clinical prediction tasks. Participatory systems allow individuals to consent to the use of their personal data in an informed manner. We caution that presenting individuals with information does not necessarily mean that individuals will understand the information that is presented to them. Effectively informing individuals remains a key consideration when implementing participatory systems in practice and an avenue for future work.

One possible limitation of our approach is that it precludes the ability to improve the system over time by collecting additional data in deployment and using it to update the model. This is because participation allows individuals might opt out of reporting personal data. One solution is to allow individuals to report additional information voluntarily for model improvement.

References

[1] Haiyan Fan and Marshall Scott Poole. 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.

[2] Jessica K Paulus, Benjamin S Wessler, Christine Lundquist, Lana LY Lai, Gowri Raman, Jennifer S Lutz, and David M Kent. Field synopsis of sex in clinical prediction models for cardiovascular disease. *Circulation: Cardiovascular Quality and Outcomes*, 9(2_suppl_1):S8–S15, 2016.

[3] Elizabeth C George, A Sarah Walker, Sarah Kiguli, Peter Olupot-Olupot, Robert O Opoka, Charles Engoru, Samuel O Akech, Richard Nyeko, George Mtove, Hugh Reyburn, et al. Predicting mortality in sick african children: the feast paediatric emergency triage (pet) score. *BMC medicine*, 13(1):1–12, 2015.

[4] Christopher C Moore, Riley Hazard, Kacie J Saulters, John Ainsworth, Susan A Adakun, Abdallah Amir, Ben Andrews, Mary Auma, Tim Baker, Patrick Banura, John A Crump, Martin P Grobusch, Michaëla A M Huson, Shevin T Jacob, Olamide D Jarrett, John Kellett, Shabir Lakhi, Albert Majwala, Martin Opio, Matthew P Rubach, Jamie Rylance, W Michael Scheld, John Schieffelin, Richard Ssekitoleko, India Wheeler, and Laura E Barnes. Derivation and validation of a universal vital assessment (uva) score: a tool for predicting mortality in adult hospitalised patients in sub-saharan africa. *BMJ Global Health*, 2(2), 2017. doi: 10.1136/bmjgh-2017-000344. URL https://gh.bmj.com/content/2/2/e000344.

[5] Vivek Narain, Fernando J Bianco Jr, David J Grignon, Wael A Sakr, J Edson Pontes, and David P Wood Jr. How accurately does prostate biopsy gleason score predict pathologic findings and disease free survival? *The Prostate*, 49(3):185–190, 2001.

[6] Chi Zhang, Ling Qin, Kang Li, Qi Wang, Yan Zhao, Bin Xu, Lianchun Liang, Yanchao Dai, Yingmei Feng, Jianping Sun, et al. A novel scoring system for prediction of disease severity in covid-19. *Frontiers in cellular and infection microbiology*, 10:318, 2020.

[7] GDPR. 2018 reform of eu data protection rules, 2018. URL https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf.

[8] Margot E Kaminski. The right to explanation, explained. *Berkeley Tech. LJ*, 34:189, 2019.

[9] Vinith M Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. When personalization harms: Reconsidering the use of group attributes in prediction. In *arXiv preprint*, 2022.

[10] OECD. Recommendation of the council concerning guidelines governing the protection of privacy and transborder flows of personal data, 2013. URL https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0188.

[11] P. Bukaty. *The California Consumer Privacy Act (CCPA): An implementation guide*. IT Governance Publishing, 2019. ISBN 9781787781337. URL https://books.google.com/books?id=vGWfDwAAQBAJ.

[12] Catherine L Anderson and Ritu Agarwal. The digitization of healthcare: boundary risks, emotion, and consumer willingness to disclose personal health information. *Information Systems Research*, 22(3): 469–490, 2011.

[13] Brooke Auxier, Lee Rainie, Monica Anderson, Andrew Perrin, Madhu Kumar, and Erica Turner. Americans and privacy: Concerned, confused and feeling lack of control over their personal information. *Pew Research Center: Internet, Science and Tech*, 2019.

[14] Gaurav Bansal, David Gefen, 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.

[15] Naveen Farag Awad and Mayuram S Krishnan. The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS quarterly*, pages 13–28, 2006.

[16] Martin Ortlieb and Ryan Garner. Sensitivity of personal data items in different online contexts. *it-Information Technology*, 58(5):217–228, 2016.

[17] April Moreno Arellano, Wenrui Dai, Shuang Wang, Xiaoqian Jiang, and Lucila Ohno-Machado. Privacy policy and technology in biomedical data science. *Annual review of biomedical data science*, 1:115, 2018.

[18] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[19] Tim S Campbell and William A Kracaw. Information production, market signalling, and the theory of financial intermediation. *the Journal of Finance*, 35(4):863–882, 1980.

[20] Thomas J Chemmanur. The pricing of initial public offerings: A dynamic model with information production. *The Journal of Finance*, 48(1):285–304, 1993.

[21] Ian Lundberg, Arvind Narayanan, Karen Levy, and Matthew J Salganik. Privacy, ethics, and data access: A case study of the fragile families challenge. *Socius*, 5:2378023118813023, 2019.

[22] Kayte Spector-Bagdady, Shengpu Tang, Sarah Jabbour, W Nicholson Price, Ana Bracic, Melissa S Creary, Sachin Kheterpal, Chad M Brummett, and Jenna Wiens. Respecting autonomy and enabling diversity: The effect of eligibility and enrollment on research data demographics: Study examines the effect of eligibility and enrollment on research data demographics. *Health Affairs*, 40(12):1892–1899, 2021.

[23] Alan Agresti. *An introduction to categorical data analysis*. John Wiley & Sons, 2018.

[24] Emilio Carrizosa, Laust Hvas Mortensen, Dolores Romero Morales, and M Remedios Sillero-Denamiel. The tree based linear regression model for hierarchical categorical variables. *Expert Systems with Applications*, 203:117423, 2022.

[25] Ewout W Steyerberg et al. *Clinical prediction models*. Springer, 2019.

[26] Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111, 2013.

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

[28] Gregory Vaughan, Robert Aseltine, Kun Chen, and Jun Yan. Efficient interaction selection for clustered data via stagewise generalized estimating equations. *Statistics in Medicine*, 39(22):2855–2868, 2020.

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

[30] Dimitris Bertsimas, Jack Dunn, and Nishanth Mundru. Optimal prescriptive trees. *INFORMS Journal on Optimization*, 1(2):164–183, 2019.

[31] Max Biggs, Wei Sun, and Markus Ettl. Model distillation for revenue optimization: Interpretable personalized pricing. *arXiv preprint arXiv:2007.01903*, 2020.

[32] Adam N Elmachtoub, Vishal Gupta, and Michael Hamilton. The value of personalized pricing. *Available at SSRN 3127719*, 2018.

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

[34] Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing*, 11(2):200–213, 2017.

[35] David M Kent, Jessica K Paulus, David Van Klaveren, Ralph D'Agostino, Steve Goodman, Rodney Hayward, John PA Ioannidis, Bray Patrick-Lake, Sally Morton, Michael Pencina, et al. The predictive approaches to treatment effect heterogeneity (path) statement. *Annals of internal medicine*, 172(1):35–45, 2020.

[36] Alice B Popejoy and Stephanie M Fullerton. Genomics is failing on diversity. *Nature News*, 538(7624): 161, 2016.

[37] Effy Vayena, Alessandro Blasimme, and I Glenn Cohen. Machine learning in medicine: Addressing ethical challenges. *PLoS medicine*, 15(11):e1002689, 2018.

[38] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 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*, pages 119–133. PMLR, 2018.

[39] Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382, 2019.

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

[41] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic Fairness. In *AEA Papers and Proceedings*, volume 108, pages 22–27, 2018.

[42] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems 31*, pages 8135–8145, 2018.

[43] Hao Wang, Berk Ustun, and Flavio P Calmon. 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.

[44] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 228–238, 2017.

[45] Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. Online certification of preference-based fairness for personalized recommender systems. *arXiv preprint arXiv:2104.14527*, 2021.

[46] Michael P Kim, Aleksandra Korolova, Guy N Rothblum, and Gal Yona. Preference-informed fairness. *arXiv preprint arXiv:1904.01793*, 2019.

[47] Davide Viviano and Jelena Bradic. Fair policy targeting. *arXiv preprint arXiv:2005.12395*, 2020.

[48] Lily Hu and Yiling Chen. Fair Classification and Social Welfare. *arXiv preprint arXiv:1905.00147*, 2019.

[49] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Fairness with minimal harm: A pareto-optimal approach for healthcare. *arXiv preprint arXiv:1911.06935*, 2019.

[50] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764. PMLR, 2020.

[51] Stephen Pfohl, Ben Marafino, Adrien Coulet, Fatima Rodriguez, Latha Palaniappan, and Nigam H Shah. Creating fair models of atherosclerotic cardiovascular disease risk. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 271–278, 2019.

[52] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963, 1993.

[53] Aaron F Struck, Berk Ustun, Andres Rodriguez Ruiz, Jong Woo Lee, Suzette M LaRoche, Lawrence J Hirsch, Emily J Gilmore, Jan Vlachy, Hiba Arif Haider, and Cynthia Rudin. Association of an electroencephalography-based risk score with seizure probability in hospitalized patients. *JAMA neurology*, 74(12):1419–1424, 2017.

[54] Aaron F Struck, Mohammad Tabaeizadeh, Sarah E Schmitt, Andres Rodriguez Ruiz, Christa B Swisher, Thanujaa Subramaniam, Christian Hernandez, Safa Kaleem, Hiba A Haider, and Abbas Fodé Cissé. Assessment of the validity of the 2helps2b score for inpatient seizure risk prediction. *JAMA neurology*, 77 (4):500–507, 2020.

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

[56] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 chi conference on human factors in computing systems*, pages 5092–5103, 2016.

[57] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[58] Valerie F Reyna and Charles J Brainerd. The importance of mathematics in health and human judgment: Numeracy, risk communication, and medical decision making. *Learning and Individual Differences*, 17(2): 147–159, 2007.

[59] Carlos Estrada, Vetta Barnes, Cathy Collins, and James C Byrd. Health literacy and numeracy. *Jama*, 282 (6):527–527, 1999.

[60] David Spiegelhalter. Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, 4(1):31–60, 2017.

[61] Liwei Zhang, Huijie Li, and Kelin Chen. Effective risk communication for public health emergency: reflection on the covid-19 (2019-ncov) outbreak in wuhan, china. In *Healthcare*, page 64. MDPI, 2020.

[62] Adrian GK Edwards, Gurudutt Naik, Harry Ahmed, Glyn J Elwyn, Timothy Pickles, Kerry Hood, and Rebecca Playle. Personalised risk communication for informed decision making about taking screening tests. *Cochrane database of systematic reviews*, Cochrane database of systematic reviews(2), 2013.

[63] Thomas J DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical science*, pages 189–212, 1996.

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

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

[66] Darshali A Vyas, David S Jones, Audra R Meadows, Khady Diouf, Nawal M Nour, and Julianna Schantz-Dunn. Challenging the use of race in the vaginal birth after cesarean section calculator. *Women's Health Issues*, 29(3):201–204, 2019.

[67] Graham Walker and Joe Habboushe. Mdcalc - medical calculators, equations, scores, and guidelines, 2022. URL https://www.mdcalc.com/.

[68] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[69] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[70] Surveillance Research Program NCI, DCCPS. Surveillance, epidemiology, and end results (seer) program research data (1975-2016), 2019. URL www.seer.cancer.gov.

[71] Jérôme Allyn, Cyril Ferdynus, Michel Bohrer, Cécile Dalban, Dorothée Valance, and Nicolas Allou. Simplified acute physiology score ii as predictor of mortality in intensive care units: a decision curve analysis. *PloS one*, 11(10):e0164828, 2016.

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

[73] William A Knaus, Frank E Harrell, Joanne Lynn, Lee Goldman, Russell S Phillips, Alfred F Connors, Neal V Dawson, William J Fulkerson, Robert M Califf, Norman Desbiens, et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3):191–203, 1995.

[74] Alfred F Connors, Neal V Dawson, Norman A Desbiens, William J Fulkerson, Lee Goldman, William A Knaus, Joanne Lynn, Robert K Oye, Marilyn Bergner, Anne Damiano, et al. A controlled trial to improve care for seriously iii hospitalized patients: The study to understand prognoses and preferences for outcomes and risks of treatments (support). *Jama*, 274(20):1591–1598, 1995.

Checklist

1. For all authors...
    1. Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
       [Yes] All authors unanimously believe that the abstract and introduction faithfully represent the intended contributions and scope of the paper.
    2. Did you describe the limitations of your work?
       [Yes] Empirically we found some limitations vis-a-vis approaches that perform imputation of protected group membership. This and other limitations have been thoroughly described through various experiments in Section 5.2 followed by an extensive discussion in Section 6.
    3. Did you discuss any potential negative societal impacts of your work?
       [N/A] Our contribution is around identifying existing sources of harm in Machine Learning models via a set of principled conditions of *'fair-use'*. We then propose approaches that mitigate such harm and improve individual privacy by giving individuals the choice to not report personal or sensitive information. We thus do not believe this question applies to our paper.
    4. Have you read the ethics review guidelines and ensured that your paper conforms to them?
       [Yes]
2. If you are including theoretical results...
    1. Did you state the full set of assumptions of all theoretical results? [N/A]
    2. Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
    1. Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?
       [Yes] The software and the open source datasets have been included in the supplement.
    2. Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
       [Yes] Sufficient details around reproducibility are in the experiments and the supplementary sections.
    3. Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All reported metrics were computed with bootstrapped resampling with replacement and binomial hypothsis test was carried out to compute p-values.
    4. Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] , see Appendix A
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
    1. If your work uses existing assets, did you cite the creators?
       [Yes] All existing assests including data sources were duly cited.
    2. Did you mention the license of the assets? [N/A]
    3. Did you include any new assets either in the supplemental material or as a URL? [N/A]
    4. Did you discuss whether and how consent was obtained from people whose data you're using/curating?
       [N/A] The datasets employed were cleared by the Institution Review Board of the original concerned insitutes.
    5. Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
       [Yes] All of the datasets that were employed in the experiments for this paper were de-identified and do not contain personally identifiable information. For full details on the datasets and their source including the original study please refer to Appendix. A.1

5. If you used crowdsourcing or conducted research with human subjects...
    1. Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
    2. Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
    3. Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Supporting Material for Experiments

In what follows, we present supporting material for the experiments in Section 5. In Appendix A.1, we include additional information about the datasets. In Appendix A.2, we summarize the performance of component models for the participatory systems in Fig. 3. In Appendix A.3, we include tables showing the performance of models and systems built to minimize error (i.e., for decision-making applications), and expected calibration error (i.e., for risk prediction).

A.1 Additional Information on Datasets

Dataset	Reference	Outcome Variable	n	d	m	$\mathcal{G}$
`cardio_eicu`	Pollard et al. [68]	patient with cardiogenic shock dies	1,341	49	4	{`age`, `sex`}
`cardio_mimic`	Johnson et al. [69]	patient with cardiogenic shock dies	5,289	49	4	{`age`, `sex`}
`lungcancer`	NCI [70]	patient dies within 5 years	120,641	84	6	{`age`, `sex`}
`saps`	Allyn et al. [71]	ICU mortality	7,797	36	4	{`age`, `HIV`}
`sleepapnea`	Ustun et al. [72]	patient has obstructive sleep apnea	1,152	28	6	{`age`, `sex`}
`support`	Connors et al. [74]	mortality within 6 months of discharge	9,105	55	6	{`age`, `sex`}

Table 3: Datasets used in Section 5. n and d denote the number of examples and features in each dataset, respectively. All datasets are de-identified and available to the public. The `cardio_eicu`, `cardio_mimic`, `lungcancer` datasets require access to public data repositories listed under the references. The `saps` and `sleepapnea` datasets must be requested from the authors. The `support` dataset can be downloaded directly from the URL below.

`cardio_eicu` & `cardio_mimic` Cardiogenic shock is an acute condition in which the heart cannot provide sufficient blood to the vital organs. We create a cohort of patients who have cardiogenic shock in an intensive care unit (ICU) stay using data from either the Collaborative Research Database V2.0 [68] or MIMIC-III [69]. Here, the outcome variable indicates whether a patient with cardiogenic shock will while in the ICU. The features reflect an exhaustive set of relevant clinical criteria derived from lab tests and vital signs (e.g. systolic BP, heart rate, hemoglobin count), and reflect measurements obtained up to 24 hours before the onset of cardiogenic shock.

`sleepapnea` We use the obstructive sleep apnea (OSA) dataset outlined in Ustun et al. [72]. This dataset includes a cohort of 1152 patients where 23% have OSA. We use all available features (e.g. BMI, comobordities, age, and sex) and binarize them, resulting in 26 binary features.

`saps` The SAPS II score is an ICU risk score used to predict the mortality of critically ill patients in the ICU [52]. The data contains records of 7,797 patients from 137 medical centers in 12 countries. Here, the outcome variable indicates whether a patient dies in the ICU, with 12.8% patient of patients dying. The features reflect comorbidities, vital signs, and lab measurements.

`support` The `support` Connors et al. [74] dataset is derived from a study of survival risk score of critically-ill patients who were discharged from the ICU. Here, we have records of 9,105 patients. The outcome variable indicates that a patient has died within six months of discharge. The features cover chronic health conditions(e.g., diabetic status, number of comorbidities), vital signs (e.g., mean blood pressure) and results of lab tests (e.g., white blood cell count). The dataset is publically available for research here: `https://biostat.app.vumc.org/wiki/Main/DataSets`.

`lungcancer` We consider a cohort of 120,641 patients who were diagnosed with lung cancer between 2004-2016 and monitored as part of the National Cancer Institute SEER study NCI [70]. Here, the outcome variable indicates if a patient die within five years from any cause, with 16.9% patients died within the first five years from diagnosis. The cohorts only represents patients from Greater California, Georgia, Kentucky, New Jersey and Louisiana, and does not cover patients who were lost to follow up (censored). Age and Sex were considered as group attributes. The features reflect the morphology and histology of the tumor (e.g., size, metastasis, stage, node count and location, number and location of notes) as well as interventions that were administered at the time of diagnosis (e.g., surgery, chemo, radiology).

A.2 Performance of Component Models for the Participatory Systems in Fig. 3

			Training			Validation			Test		
			ERROR			ERROR			ERROR		
Group	Model	Parent	$\Delta_0(h)$	$\Delta_{pa}(h)$	$R(h)$	$\Delta_0(h)$	$\Delta_{pa}(h)$	$R(h)$	$\Delta_0(h)$	$\Delta_{pa}(h)$	$R(h)$
-	h_0	h_0	0.0%	0.0%	20.8%	0.0%	0.0%	21.1%	0.0%	0.0%	21.7%
negative	h_6	h_0	-0.8%	-0.8%	18.8%	-0.4%	-0.4%	19.2%	-0.8%	-0.8%	19.7%
positive	h_0	h_0	0.0%	0.0%	22.0%	0.0%	0.0%	22.6%	0.0%	0.0%	22.8%
<30 & positive	h_3	h_0	-12.3%	-12.3%	0.0%	-13.5%	-13.5%	0.0%	-14.2%	-14.2%	0.0%
>30 & positive	h_{26}	h_0	-3.1%	-3.1%	28.6%	-3.1%	-3.1%	28.9%	-2.7%	-2.7%	28.6%

Table 4: Group-level performance as measured by error on dataset (saps). $\Delta_0(h)$ represents the change in error compared with the generic classifier (negative is a decrease in error). $\Delta_{pa}(h)$ is the change in error compared with the parent classifier in the reporting tree (see column Parent). $R(h)$ is the error rate for the group. Performance is reported across training, validation and test.

			Training			Validation			Test		
			AUC			AUC			AUC		
Group	Model	Parent	$\Delta_0(h)$	$\Delta_{pa}(h)$	$R(h)$	$\Delta_0(h)$	$\Delta_{pa}(h)$	$R(h)$	$\Delta_0(h)$	$\Delta_{pa}(h)$	$R(h)$
-	h_0	h_0	0.000	0.000	0.874	0.000	0.000	0.870	0.000	0.000	0.865
negative	h_9	h_9	0.025	0.000	0.911	0.026	0.000	0.911	0.026	0.000	0.906
positive	h_6	h_6	0.011	0.000	0.881	0.011	0.000	0.876	0.011	0.000	0.871
<30 & negative	h_{27}	h_9	0.033	0.020	0.959	0.030	0.018	0.954	0.035	0.022	0.954
<30 & positive	h_3	h_6	0.082	0.075	1.000	0.092	0.086	1.000	0.101	0.093	1.000
>30 & positive	h_{30}	h_6	0.136	0.121	0.937	0.135	0.121	0.937	0.141	0.123	0.941

Table 5: Group-level performance as measured by AUC on dataset (saps). $\Delta_0(h)$ represents the change in AUC compared with the generic classifier (positive is an increase in AUC). $\Delta_{pa}(h)$ is the change in AUC compared with the parent classifier in the reporting tree (see column Parent). $R(h)$ is the AUC for the group. Performance is reported across training, validation and test.

635 ## A.3 Additional Experimental Results

	Static	Imputed		Participatory			
Dataset	Metrics	1Hot	mHot	Impute	Minimal	Flat	Seq
	Overall Performance	22.4%	21.9%	23.4%	21.7%	**16.1%**	**16.1%**
`cardio_eicu`	Overall Gain	0.2%	0.7%	-0.7%	0.9%	**6.5%**	**6.5%**
$n = 1341, d = 49$	Group Gains	-2.1% – 3.2%	-1.9% – 5.1%	-2.1% – 0.3%	0.0% – 3.2%	-1.9% – 17.8%	-1.9% – 17.8%
$\mathcal{G} = \{age, sex\}$	Max Disparity	5.3%	7.1%	2.4%	3.2%	19.7%	19.7%
$m = 4$	# Violations	2	2	2	**0**	1	1
Pollard et al. [68]	Data Reduction	0.0%	0.0%	NA%	0.0%	50.0%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	50.0%	100.0%
	Overall Performance	19.5%	19.3%	19.1%	19.2%	**18.1%**	**18.1%**
`cardio_mimic`	Overall Gain	-0.3%	-0.1%	0.1%	0.0%	**1.1%**	**1.1%**
$n = 5289, d = 49$	Group Gains	-0.8% – 0.3%	-0.5% – 0.3%	-0.8% – 0.7%	0.0% – 0.0%	-0.6% – 3.3%	-0.6% – 3.3%
$\mathcal{G} = \{age, sex\}$	Max Disparity	1.1%	0.8%	1.5%	0.0%	3.9%	3.9%
$m = 4$	# Violations	2	2	1	**0**	1	1
Johnson et al. [69]	Data Reduction	0.0%	0.0%	NA%	0.0%	62.6%	31.3%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	57.2%	100.0%
	Overall Performance	19.6%	19.6%	19.6%	19.5%	**18.9%**	**18.9%**
`lungcancer`	Overall Gain	-0.1%	-0.1%	-0.1%	-0.0%	**0.6%**	**0.6%**
$n = 120641, d = 84$	Group Gains	-0.4% – 0.1%	-0.3% – 0.1%	-0.4% – 0.0%	-0.1% – 0.0%	0.3% – 0.9%	0.4% – 0.9%
$\mathcal{G} = \{age, sex\}$	Max Disparity	0.6%	0.4%	0.4%	0.1%	0.5%	0.5%
$m = 6$	# Violations	4	3	4	1	**0**	**0**
NCI [70]	Data Reduction	0.0%	0.0%	NA%	0.0%	25.0%	41.6%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	33.3%	100.0%
	Overall Performance	20.4%	20.7%	26.8%	20.4%	**11.1%**	**11.1%**
`saps`	Overall Gain	1.3%	1.0%	-5.1%	1.3%	**10.6%**	**10.6%**
$n = 7797, d = 36$	Group Gains	0.0% – 3.6%	0.0% – 2.7%	-20.8% – 0.7%	0.0% – 3.6%	4.3% – 17.2%	3.9% – 17.2%
$\mathcal{G} = \{HIV, age\}$	Max Disparity	3.6%	2.7%	21.5%	3.6%	12.9%	13.3%
$m = 4$	# Violations	**0**	**0**	2	**0**	**0**	**0**
Allyn et al. [71]	Data Reduction	0.0%	0.0%	NA%	0.0%	37.4%	31.3%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	39.9%	100.0%
	Overall Performance	29.1%	29.3%	30.3%	28.9%	**24.2%**	**24.2%**
`sleepapnea`	Overall Gain	0.1%	-0.1%	-1.1%	0.3%	**4.9%**	**4.9%**
$n = 1152, d = 26$	Group Gains	-1.1% – 1.2%	-0.8% – 0.4%	-2.7% – 0.4%	0.0% – 1.2%	0.0% – 13.8%	0.0% – 13.8%
$\mathcal{G} = \{age, sex\}$	Max Disparity	2.4%	1.2%	3.1%	1.2%	13.8%	13.8%
$m = 6$	# Violations	1	1	3	**0**	**0**	**0**
Ustun et al. [72]	Data Reduction	0.0%	0.0%	NA%	0.0%	58.6%	29.3%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	54.7%	100.0%
	Overall Performance	35.0%	35.0%	35.8%	35.4%	**34.8%**	**34.8%**
`support`	Overall Gain	0.8%	0.8%	0.0%	0.4%	**1.1%**	**1.1%**
$n = 9105, d = 55$	Group Gains	0.0% – 2.3%	-0.5% – 2.6%	-1.8% – 1.9%	0.0% – 1.4%	-0.3% – 2.9%	-0.3% – 2.9%
$\mathcal{G} = \{age, sex\}$	Max Disparity	2.3%	3.0%	3.7%	1.4%	3.1%	3.1%
$m = 6$	# Violations	**0**	**0**	2	**0**	1	**0**
Knaus et al. [73]	Data Reduction	0.0%	0.0%	NA%	0.0%	50.0%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	50.0%	100.0%

Table 6: Overview of performance, data use, and consent for all personalized models on all datasets, as measured by *test error*.

		STATIC		IMPUTED	PARTICIPATORY		
Dataset	Metrics	1Hot	mHot	Impute	Minimal	Flat	Seq
	Overall Performance	0.858	0.857	0.858	0.858	**0.923**	**0.923**
cardio_eicu	Overall Gain	0.001	-0.000	0.001	0.001	**0.067**	**0.067**
$n = 1341, d = 49$	Group Gains	-0.001 – 0.002	-0.001 – 0.002	-0.001 – 0.002	-0.001 – 0.002	0.008 – 0.094	0.008 – 0.094
$\mathcal{G} = \{age, sex\}$	Max Disparity	0.003	0.003	0.003	0.003	0.087	0.087
$m = 4$	# Violations	2	1	3	1	**0**	**0**
Pollard et al. [68]	Data Reduction	0.0%	0.0%	NA%	0.0%	50.0%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	50.0%	100.0%
	Overall Performance	0.876	0.876	0.876	0.877	**0.896**	**0.896**
cardio_mimic	Overall Gain	-0.000	-0.000	-0.000	0.000	**0.020**	**0.020**
$n = 5289, d = 49$	Group Gains	-0.000 – 0.001	-0.000 – 0.001	-0.000 – 0.001	-0.000 – 0.001	0.005 – 0.034	0.005 – 0.034
$\mathcal{G} = \{age, sex\}$	Max Disparity	0.001	0.001	0.001	0.001	0.028	0.028
$m = 4$	# Violations	**0**	2	**0**	**0**	**0**	**0**
Johnson et al. [69]	Data Reduction	0.0%	0.0%	NA%	0.0%	37.5%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	40.0%	100.0%
	Overall Performance	0.855	0.855	0.855	0.855	**0.861**	**0.861**
lungcancer	Overall Gain	0.001	0.001	0.001	0.001	**0.007**	**0.007**
$n = 120641, d = 84$	Group Gains	-0.000 – 0.000	-0.000 – 0.000	-0.000 – 0.000	-0.000 – 0.000	0.001 – 0.012	0.001 – 0.012
$\mathcal{G} = \{age, sex\}$	Max Disparity	0.001	0.000	0.001	0.001	0.011	0.011
$m = 6$	# Violations	2	2	2	1	**0**	**0**
NCI [70]	Data Reduction	0.0%	0.0%	NA%	0.0%	29.2%	16.7%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	35.3%	100.0%
	Overall Performance	0.875	0.877	0.875	0.875	**0.960**	**0.960**
saps	Overall Gain	0.010	0.011	0.010	0.009	**0.095**	**0.095**
$n = 7797, d = 36$	Group Gains	-0.000 – 0.015	-0.002 – 0.019	-0.000 – 0.015	0.000 – 0.015	0.035 – 0.139	0.026 – 0.139
$\mathcal{G} = \{HIV, age\}$	Max Disparity	0.015	0.020	0.015	0.015	0.105	0.114
$m = 4$	# Violations	**0**	1	**0**	**0**	**0**	**0**
Allyn et al. [71]	Data Reduction	0.0%	0.0%	NA%	0.0%	25.0%	31.3%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	33.3%	100.0%
	Overall Performance	0.774	0.774	0.774	0.775	**0.850**	**0.850**
sleepapnea	Overall Gain	-0.002	-0.002	-0.002	-0.001	**0.074**	**0.074**
$n = 1152, d = 26$	Group Gains	-0.002 – 0.002	-0.002 – 0.003	-0.002 – 0.002	-0.002 – 0.002	0.004 – 0.115	0.004 – 0.115
$\mathcal{G} = \{age, sex\}$	Max Disparity	0.004	0.005	0.004	0.003	0.111	0.111
$m = 6$	# Violations	2	3	2	1	**0**	**0**
Ustun et al. [72]	Data Reduction	0.0%	0.0%	NA%	0.0%	50.0%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	50.0%	100.0%
	Overall Performance	0.707	0.706	0.707	0.706	**0.712**	**0.712**
support	Overall Gain	0.002	0.001	0.002	0.001	**0.007**	**0.007**
$n = 9105, d = 55$	Group Gains	-0.000 – 0.003	-0.000 – 0.003	-0.000 – 0.003	0.000 – 0.003	-0.000 – 0.023	-0.000 – 0.023
$\mathcal{G} = \{age, sex\}$	Max Disparity	0.003	0.003	0.003	0.003	0.023	0.023
$m = 6$	# Violations	**0**	**0**	**0**	**0**	**0**	**0**
Knaus et al. [73]	Data Reduction	0.0%	0.0%	NA%	0.0%	66.7%	33.3%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	60.0%	100.0%

Table 7: Overview of performance, data use, and consent for all personalized models on all datasets, as measured by *test AUC*.

		STATIC		IMPUTED	PARTICIPATORY		
Dataset	Metrics	1Hot	mHot	Impute	Minimal	Flat	Seq
	Overall Performance	0.893	0.893	0.893	0.893	**0.949**	**0.949**
`cardio_eicu`	Overall Gain	0.003	0.002	0.003	0.003	**0.059**	**0.059**
$n = 1341, d = 49$	Group Gains	-0.006 – 0.012	-0.008 – 0.010	-0.006 – 0.012	-0.006 – 0.012	0.017 – 0.070	0.017 – 0.070
$\mathcal{G} = \{age, sex\}$	Max Disparity	0.018	0.018	0.018	0.018	0.053	0.053
$m = 4$	# Violations	2	2	2	2	**0**	**0**
Pollard et al. [68]	Data Reduction	0.0%	0.0%	NA%	0.0%	12.6%	12.6%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	28.6%	100.0%
	Overall Performance	0.880	0.881	0.880	0.880	**0.920**	**0.920**
`cardio_mimic`	Overall Gain	-0.000	0.001	-0.000	0.000	**0.039**	**0.039**
$n = 5289, d = 49$	Group Gains	-0.002 – 0.001	-0.000 – 0.002	-0.002 – 0.001	0.000 – 0.000	0.016 – 0.048	0.016 – 0.048
$\mathcal{G} = \{age, sex\}$	Max Disparity	0.003	0.002	0.003	0.000	0.032	0.032
$m = 4$	# Violations	2	**0**	1	**0**	**0**	**0**
Johnson et al. [69]	Data Reduction	0.0%	0.0%	NA%	0.0%	50.0%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	50.0%	100.0%
	Overall Performance	0.849	0.849	0.849	0.848	**0.856**	**0.856**
`lungcancer`	Overall Gain	0.002	0.001	0.002	0.000	**0.008**	**0.008**
$n = 120641, d = 84$	Group Gains	-0.001 – 0.003	-0.001 – 0.002	-0.001 – 0.003	0.000 – 0.003	0.002 – 0.020	0.002 – 0.020
$\mathcal{G} = \{age, sex\}$	Max Disparity	0.004	0.003	0.004	0.003	0.018	0.018
$m = 6$	# Violations	1	1	**0**	**0**	**0**	**0**
NCI [70]	Data Reduction	0.0%	0.0%	NA%	0.0%	29.2%	20.8%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	35.3%	100.0%
	Overall Performance	0.921	0.922	0.921	0.922	**0.966**	**0.966**
`saps`	Overall Gain	0.003	0.004	0.003	0.004	**0.048**	**0.048**
$n = 7797, d = 36$	Group Gains	-0.002 – 0.010	-0.002 – 0.013	-0.002 – 0.010	-0.000 – 0.010	0.009 – 0.109	0.009 – 0.109
$\mathcal{G} = \{HIV, age\}$	Max Disparity	0.012	0.015	0.012	0.011	0.100	0.100
$m = 4$	# Violations	2	1	2	1	**0**	**0**
Allyn et al. [71]	Data Reduction	0.0%	0.0%	NA%	0.0%	50.0%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	50.0%	100.0%
	Overall Performance	0.825	0.824	0.825	0.824	**0.944**	**0.944**
`sleepapnea`	Overall Gain	0.008	0.006	0.008	0.006	**0.126**	**0.126**
$n = 1152, d = 26$	Group Gains	-0.004 – 0.009	-0.005 – 0.012	-0.004 – 0.009	-0.003 – 0.009	0.059 – 0.159	0.059 – 0.159
$\mathcal{G} = \{age, sex\}$	Max Disparity	0.012	0.017	0.012	0.012	0.100	0.100
$m = 6$	# Violations	2	2	**0**	1	**0**	**0**
Ustun et al. [72]	Data Reduction	0.0%	0.0%	NA%	0.0%	41.7%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	42.9%	100.0%
	Overall Performance	0.695	0.698	0.695	0.695	**0.722**	**0.722**
`support`	Overall Gain	0.001	0.003	0.001	0.001	**0.027**	**0.027**
$n = 9105, d = 55$	Group Gains	-0.004 – 0.007	0.001 – 0.007	-0.004 – 0.007	0.000 – 0.007	0.008 – 0.052	0.008 – 0.052
$\mathcal{G} = \{age, sex\}$	Max Disparity	0.011	0.006	0.011	0.007	0.044	0.044
$m = 6$	# Violations	2	**0**	1	**0**	**0**	**0**
Knaus et al. [73]	Data Reduction	0.0%	0.0%	NA%	0.0%	41.6%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	42.8%	100.0%

Table 8: Performance and Data Use of personalized models for all datasets, as measured by **test AUC** using random forest component classifiers.

		Static		Imputed	Participatory		
Dataset	Metrics	1Hot	mHot	Impute	Minimal	Flat	Seq
	Overall Performance	17.9%	17.5%	19.2%	17.7%	**12.9%**	**12.9%**
`cardio_eicu`	Overall Gain	0.9%	1.2%	-0.4%	1.1%	**5.9%**	**5.9%**
$n = 1341, d = 49$	Group Gains	-0.4% – 3.2%	-0.7% – 2.9%	-1.8% – 0.3%	0.0% – 3.2%	2.6% – 8.1%	2.6% – 8.1%
$\mathcal{G} = \{age, sex\}$	Max Disparity	3.5%	3.6%	2.1%	3.2%	5.5%	5.5%
$m = 4$	# Violations	**0**	1	1	**0**	**0**	**0**
Pollard et al. [68]	Data Reduction	0.0%	0.0%	NA%	0.0%	50.0%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	50.0%	100.0%
	Overall Performance	21.3%	20.9%	21.3%	20.3%	**16.8%**	**16.8%**
`cardio_mimic`	Overall Gain	-1.2%	-0.7%	-1.2%	-0.2%	**3.4%**	**3.4%**
$n = 5289, d = 49$	Group Gains	-1.9% – -0.6%	-1.1% – -0.3%	-1.8% – -0.7%	-0.7% – 0.0%	0.5% – 5.0%	0.5% – 5.0%
$\mathcal{G} = \{age, sex\}$	Max Disparity	1.3%	0.8%	1.1%	0.7%	4.5%	4.5%
$m = 4$	# Violations	4	4	4	1	**0**	**0**
Johnson et al. [69]	Data Reduction	0.0%	0.0%	NA%	0.0%	50.0%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	50.0%	100.0%
	Overall Performance	20.0%	20.2%	20.0%	20.0%	**19.3%**	**19.3%**
`lungcancer`	Overall Gain	0.1%	-0.1%	0.1%	0.1%	**0.8%**	**0.8%**
$n = 120641, d = 84$	Group Gains	-0.3% – 0.2%	-0.5% – 0.0%	-0.3% – 0.3%	0.0% – 0.2%	0.0% – 2.3%	0.0% – 2.3%
$\mathcal{G} = \{age, sex\}$	Max Disparity	0.6%	0.5%	0.6%	0.2%	2.3%	2.3%
$m = 6$	# Violations	1	4	1	**0**	**0**	**0**
NCI [70]	Data Reduction	0.0%	0.0%	NA%	0.0%	33.3%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	37.5%	100.0%
	Overall Performance	14.1%	15.0%	17.0%	13.9%	**9.8%**	**9.8%**
`saps`	Overall Gain	0.9%	-0.0%	-1.9%	1.1%	**5.2%**	**5.2%**
$n = 7797, d = 36$	Group Gains	-0.8% – 3.4%	-0.5% – 0.3%	-5.1% – 0.8%	0.0% – 3.4%	0.0% – 16.4%	0.0% – 16.4%
$\mathcal{G} = \{HIV, age\}$	Max Disparity	4.2%	0.8%	5.9%	3.4%	16.4%	16.4%
$m = 4$	# Violations	1	1	3	**0**	**0**	**0**
Allyn et al. [71]	Data Reduction	0.0%	0.0%	NA%	0.0%	37.3%	18.6%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	36.3%	100.0%
	Overall Performance	26.3%	26.0%	26.9%	26.2%	**12.5%**	**12.5%**
`sleepapnea`	Overall Gain	1.5%	1.8%	0.9%	1.6%	**15.3%**	**15.3%**
$n = 1152, d = 26$	Group Gains	-0.8% – 4.2%	0.4% – 3.8%	-2.2% – 4.2%	0.0% – 4.2%	3.3% – 22.2%	3.3% – 22.2%
$\mathcal{G} = \{age, sex\}$	Max Disparity	5.0%	3.4%	6.5%	4.2%	18.9%	18.9%
$m = 6$	# Violations	1	**0**	1	**0**	**0**	**0**
Ustun et al. [72]	Data Reduction	0.0%	0.0%	NA%	0.0%	33.5%	25.0%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	37.6%	100.0%
	Overall Performance	36.0%	35.9%	35.9%	35.8%	**35.6%**	**35.6%**
`support`	Overall Gain	-0.3%	-0.2%	-0.2%	-0.0%	**0.1%**	**0.1%**
$n = 9105, d = 55$	Group Gains	-0.9% – 0.2%	-1.2% – 1.3%	-1.0% – 0.9%	-0.8% – 0.2%	-1.6% – 1.4%	-1.6% – 1.1%
$\mathcal{G} = \{age, sex\}$	Max Disparity	1.2%	2.5%	1.9%	1.0%	3.1%	2.7%
$m = 6$	# Violations	3	3	4	**1**	**1**	**1**
Knaus et al. [73]	Data Reduction	0.0%	0.0%	NA%	0.0%	33.4%	33.3%
	Opportunity for Consent	0.0%	0.0%	NA%	0.0%	37.5%	100.0%

Table 9: Performance and Data Use of personalized models for all datasets, as measured by **test error** using random forest component classifiers.

B Supporting Material for Section 4

In what follows, we provide details on the routine used for the EnumerateTrees procedure in Algorithm 1. We summarize the routine in Algorithm 2, and discuss it below. The input to Algorithm 2 is an ordered collection of reporting options $\mathcal{R}$. The algorithm uses the reporting options to construct the set of all possible reporting trees, each of which branches on all of the attributes in $\mathcal{R}$. At a high level, Algorithm 2 recurses through the attributes one at a time, building trees that begin with each attribute sequentially. Enumerating all possible trees ensures we can recover the best tree given the selection criteria and allows for flexible post-hoc selection criteria (e.g., let a developer choose among the top k trees). In settings constrained by computational resources, we can impose additional stopping criteria and modify the ordering such that we enumerate more plausible trees first or exclusively (e.g., by changing the ordering of $\mathcal{R}$ or imposing constraints in ALLPERMUTATIONS).

Algorithm 2 Routine to Enumerate All Possible Reporting Trees for Reporting Options $\mathcal{R}$

```
procedure ENUMERATETREES($\mathcal{R}$)
    if dim($\mathcal{R}$) = 1 return [$T_\mathcal{R}$]        base case: we are left with only a single attribute on which to branch
    AllTrees ← [ ]
    for $\mathcal{A}$ in $\mathcal{R}$ do        Each attribute in list of attributes $\mathcal{R}$
        $T_\mathcal{A}$ ← reporting tree with $n_\mathcal{A} := |\mathcal{A}|$ leaves
        $\mathcal{U}$ ← unsolicited attributes $\mathcal{R} \setminus \mathcal{A}$
        AllSubtrees ← ENUMERATETREES($\mathcal{U}$)        All subtrees using all attributes except $\mathcal{A}$
        for $\mathcal{P}$ in ALLPERMUTATIONS(AllSubTrees, $n_\mathcal{A}$) do:        Each permutation of $n_\mathcal{A}$ subtrees
            $T_{a,\mathcal{P}}$ ← $T_a$.copy()
            $T_{a,\mathcal{P}}$ ← $T_{a,\mathcal{P}}$.assign_to_leaves($\mathcal{P}$)        assign_to_leaves extends the tree by assigning subtrees to each leaf
            AllTrees ← AllTrees ∪ $T_{a,s}$
        end for
    end for
    return AllTrees, set of all distinct reporting trees for reporting options $\mathcal{R}$
end procedure
```