Participatory Systems for Personalized Prediction

Anonymous Authors[1]

Abstract

Machine learning models are often personalized based on information that is protected, sensitive, self-reported, or costly to acquire. These models use information about people, but do not facilitate nor inform their *consent*. Individuals cannot opt out of reporting information that a model needs to personalize their predictions, nor tell if they would benefit from personalization in the first place. In this work, we introduce a new family of prediction models, called *participatory systems*, that allow individuals to opt into personalization at prediction time. We present a model-agnostic algorithm to learn participatory systems for supervised learning tasks where models are personalized with categorical group attributes. We conduct a comprehensive empirical study of participatory systems in clinical prediction tasks, comparing them to common approaches for personalization and imputation. Our results demonstrate that participatory systems can facilitate and inform consent in a way that improves performance and privacy across all groups who report personal data.

1. Introduction

Machine learning models routinely 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 settings are often *personalized*, in that they use personal information to target heterogeneous subpopulations. Typically, models are personalized with categorical attributes that specify groups [i.e., categorization as per the taxonomy of 26]. In medicine, for example, clinical prediction models include *group attributes* that are *protected* (e.g., `sex` in the CHA_2DS_2 Score for Stroke Risk), *sensitive* (e.g., `HIV` status in the VA COVID-19 Mortality Score), *self-reported* (e.g., `first_menstral_period` in the Gail Breast Cancer Risk Scores), or *costly* to acquire (e.g., `lab_values` for Alvarado Acute Appendicitis Score).

Online platforms that solicit personal data from individuals are designed to support *informed consent*: individuals can opt out of providing personal data, and understand how it will be used to support their experience [see e.g., GDPR consent banners 27, 34]. Personalized models do not provide such functionality: individuals cannot opt out of reporting data used to personalize their predictions, nor tell if it would improve their predictions. In effect, models are built under the assumption that data available at training time will also be available at prediction time. In practice, this has led to a proliferation of models that require individuals to report information they may be unwilling or unable to provide – see e.g., Denver HIV Risk Score, which requires individuals to report `age`, `gender`, `sexual practices`, and `ethnicity` [30]. In settings where individuals can input their data directly (e.g., online medical diagnostics), individuals may decline to report optional information that would improve their predictions, or report information that is wrong by reporting untruthfully or pigeonholing themselves into a category they do not identify with.

The broader lack of support for informed consent in personalization is problematic because standard techniques for personalization do not improve performance for all groups who report personal data [see 53, 44]. In practice, a personalized model can perform *worse* or the same as a *generic model* trained without personal information for a group with specific characteristics. Such models violate the implicit promise of personalization – as individuals report personal information without receiving a tailored performance gain in return. These instances of "worsenalization" are prevalent, hard to detect, and hard to resolve [see 53] – but could easily be mitigated by allowing individuals to opt out of personalization, and informing them of its gains (see Fig. 1).

In this paper, we introduce a family of machine learning models called *participatory systems* that facilitate and inform consent. Participatory systems *facilitate* consent by allowing individuals to report additional personal data at prediction time, and *inform* consent by showing them how it will affect their predictions. Models that facilitate consent operate as *markets* in which individuals report personal data in exchange for performance gains, and model developers promote participation by ensuring that reporting will lead to gains for each group. In the context of personalization, incentives are aligned as all parties benefit from more accurate predictions. In turn, the technical challenges stem from designing markets that will operate efficiently – i.e., models that facilitate and inform consent while performing as well

Group	Data		Personalized		Generic		Traditional Personalization: groups receive predictions from h			Minimal Participatory System: groups opt into predictions from h or h_0		
							Model	Data Use	Gain	Model	Data Use	Gain
$\boldsymbol{g}$	$n_{\boldsymbol{g}}^+$	$n_{\boldsymbol{g}}^-$	h	$R_{\boldsymbol{g}}(h)$	h_0	$R_{\boldsymbol{g}}(h_0)$	h	$\boldsymbol{r}$	$\Delta R_{\boldsymbol{g}}(h, h_0)$	h	$\boldsymbol{r}$	$\Delta R_{\boldsymbol{g}}(h, h_0)$
female, old	0	24	+	24	−	0	h	female, old	−24	h_0	∅	0
female, young	25	0	+	0	−	25	h	female, young	25	h	female, young	25
male, old	25	0	+	0	−	25	h	male, old	25	h	female, young	25
male, young	0	27	−	0	−	0	h	male, young	0	h_0	∅	0
Total:	50	51		24		50			26			50

Figure 1: Simple classification task where participation improves performance and limits data collection. We are given $n^+ = 50$ positive 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, and evaluate the gains from personalization with respect to the best linear model without group attributes h_0. In a traditional model (left), individuals must report group membership to h. Here, personalization would reduce error from 50 to 24, but assigns the same predictions to [male, young] and detrimental predictions to [female, old]. In a minimal participatory system (right), individuals who opt into personalization receive predictions from h while those who opt out receive predictions from h_0. In this case, individuals in groups [female, old] and [male, young] would opt out of personalization. The resulting system would achieve an overall error rate of 0 and reduce unnecessary data collection.

as possible. This work addresses these challenges by developing systems that: (i) perform well when individuals opt in (to promote participation) or opt out (to safeguard against abstention); (ii) provide multiple opportunities for individuals to decide what to report and to understand its gain (to facilitate and inform consent). The resulting systems can produce large improvements in performance and privacy across all groups who report personal data, by tailoring and limiting unnecessary data collection when it does not.

The main contributions of this work are:

1. We introduce a family of prediction models to facilitate and inform consent in supervised learning tasks.
2. We develop a model-agnostic algorithm to learn participatory systems. Our approach can produce a variety of systems that promote participation in deployment and that handle constraints on data use and acquisition.
3. We conduct a comprehensive empirical study of participatory systems in clinical prediction tasks. Our results show how our approach can facilitate and inform consent in a way that improves performance and limits unnecessary data collection.
4. We provide a Python library to build and evaluate participatory systems, available at https://anonymous.4open.science/r/psc_public-164C/

Related Work

Data Privacy. Participatory systems support modern principles of responsible data use such as *informed consent* and *collection limitation* (i.e., data should be collected with the consent of a data subject, and restricted to only what is necessary). These principles are articulated in, e.g., OECD guidelines [42], the GDPR [27], and the California Consumer Privacy Act [17]. These principles stem from a long on the right to data privacy [34]. They are motivated - in part - by a line of work showing that individuals care deeply about their ability to control personal data [12, 5, 10] but differ considerably in their desire or capacity to share it [see e.g. 11, 43, 18, 19, 9, 40, 7].

Personalization. We study personalization where models are personalized with categorical attributes that encode personal characteristics. [i.e., "categorization" rather than "individualization" as per the taxonomy of 26]. Modern techniques build on extensive work for learning models with categorical data [see e.g., 3, 49] to improve model performance at a population level using group attributes – e.g., by accounting for higher-order interaction effects [15, 39, 56] and recursive partitioning [25, 16, 14, 13]. Our work provides an alternative approach to personalization in settings where we may wish to facilitate and inform consent – e.g., when we must assign predictions using features that are collected at prediction time [see e.g., 8, 2, 58].

Algorithmic Fairness. Our work is broadly related to algorithmic fairness in that it seeks to improve model performance at a group level. In particular, our goal is to build systems that perform as well as possible for each group that reports personal data [55]. These systems naturally ensure the "fair use" of group attributes for personalization [55, 53, 44] – which are necessary conditions for each group to report personal information voluntarily and truthfully. This line of work broadly complements research on preference-based fairness [59, 55, 36, 57, 24], on ensuring group fairness across complex group structures [35, 31, 28], and on the study of privacy across subpopulations [11].

2. Participatory Systems

We consider a supervised learning task where we personalize a model with categorical attributes. We start with a dataset $\{(\boldsymbol{x}_i, y_i, \boldsymbol{g}_i)\}_{i=1}^n$ where each example consists of a feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$, a label $y_i \in \mathcal{Y}$, and a vector of k cat-

egorical attributes $\boldsymbol{g}_i = [g_{i,1}, \ldots, g_{i,k}] \in \mathcal{G}_1 \times \ldots \times \mathcal{G}_k = \mathcal{G}$ used for personalization. We refer to $\mathcal{G}$ as *group attributes* and to $\boldsymbol{g}_i$ as the *group membership* of person i.

We consider a setting where each person can opt out of personalization by declining to report group attributes at prediction time. We let $\varnothing$ denote the value of a group attribute that a person does not report and let $\boldsymbol{r}_i = [r_{i,1}, \ldots, r_{i,k}] \in \mathcal{R} \subseteq \mathcal{G} \times \boldsymbol{\varnothing}$ denote the *reported group membership* of person i. For example, a person with $\boldsymbol{g}_i = [\texttt{female}, \texttt{HIV} = +]$ could report $\boldsymbol{r}_i = [\texttt{female}, \varnothing]$ by declining to report $\texttt{HIV}$ and $\boldsymbol{r}_i = \boldsymbol{\varnothing} := [\varnothing, \ldots, \varnothing]$ by opting out of personalization entirely. Each model specifies a set of *reporting options* $\mathcal{R}$ that are available to individuals at prediction time. Thus, a model that did not allow individuals to opt out of personalization who have $\mathcal{R} = \mathcal{G}$, and a model that allows individuals to report any subset of group attributes would have $\mathcal{R} = \mathcal{G} \times \boldsymbol{\varnothing}$.

We use the dataset to train a model $h : \mathcal{X} \times \mathcal{R} \to \mathcal{Y}$ by empirical risk minimization with a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. We denote the *empirical risk* and *true risk* of a model h as $\hat{R}(h)$ and $R(h)$, respectively. Given a model, we evaluate model performance for each *reporting group* – i.e., over individuals who report specific values $\boldsymbol{r} \in \mathcal{R}$. As part of this evaluation, we consider how model performance changes for group $\boldsymbol{r}$ if they were to report $\boldsymbol{r}'$. Given a model h, we measure its true risk and empirical risk for group $\boldsymbol{r} \in \mathcal{R}$ when they report $\boldsymbol{r}' \in \mathcal{R}$ as:

$$R_{\boldsymbol{r}}(h(\cdot, \boldsymbol{r}')) := \mathbb{E}\left[\ell\left(h(\boldsymbol{x}, \boldsymbol{r}), y\right) \mid \mathcal{R} = \boldsymbol{r}\right]$$
$$\hat{R}_{\boldsymbol{r}}(h(\cdot, \boldsymbol{r}')) := \frac{1}{n_{\boldsymbol{r}}} \sum_{i:\boldsymbol{r}_i = \boldsymbol{r}} \ell\left(h(\boldsymbol{x}_i, \boldsymbol{r}'), y_i\right).$$

We compute the *gains of personalization* for each reporting group by comparing the performance of the personalized model to that of a *generic model* $h_0 : \mathcal{X} \times \mathcal{Y}$. The generic model represents the best model trained on a dataset without group attributes $h_0 \in \operatorname{argmin}_{h \in \mathcal{H}_0} \hat{R}(h)$. We denote the gains of personalization for a reporting group $\boldsymbol{r}$ in terms of true risk and empirical risk as:

$$\Delta_{\boldsymbol{r}}(h, h_0) := \hat{R}_{\boldsymbol{r}}(h_0) - \hat{R}_{\boldsymbol{r}}(h)$$
$$\hat{\Delta}_{\boldsymbol{r}}(h, h_0) := \hat{R}_{\boldsymbol{r}}(h_0) - \hat{R}_{\boldsymbol{r}}(h)$$

We wish to ensure gains in terms of true risk, but can only measure the gains in terms of empirical risk. We assume that individuals prefer to receive more accurate predictions. This assumption holds in personalization tasks where all individuals prefer to receive correct predictions – e.g., when predicting the risk of a serious illness [see e.g., 51, 38, 50]. It does not hold in applications where some individuals may prefer inaccurate predictions – e.g., predicting the risk of organ failure for an organ transplant [and other "polar" applications in 45].

System Architecture In Fig. 2, we show three participatory systems that differ in terms of their reporting options, their ability to inform consent, and their training and implementation requirements:

Minimal systems let individuals opt out of receiving predictions from an existing personalized model h. Individuals who opt out receive predictions from a generic model h_0 trained without group attributes. These systems can be built by training one additional model.

Flat systems let individuals opt into partial personalization by reporting any subset group attributes. This architecture allows individuals to receive personalized predictions without reporting specific characteristics. For example, a person with $\boldsymbol{g}_i = [\texttt{old}, \texttt{female}]$ can report $\boldsymbol{r}_i = [\texttt{old}, \varnothing]$. These systems can improve performance by using a distinct model to assign personalized predictions to each reporting group.

Sequential systems let individuals opt into partial personalization by reporting one group attribute at a time. This architecture is better suited for informing consent as users can make reporting decisions by comparing 1 model compared to 2^k models. They are also well-suited for settings where group attributes encode information that must be acquired at prediction time (e.g., the outcome of a test result).

We represent the interface of each system as an M-ary tree [i.e., a tree with at most M branches 37] whose nodes map personalized models to a reporting group. Each tree starts with a generic model at its root and branches out as individuals make reporting decisions. A minimal system corresponds to a tree of depth 1 with $M = |\mathcal{G}| + 1$ leaves. A flat system corresponds to a tree of depth 1 with $M = |\mathcal{R}|$ leaves. A sequential system corresponds to a M-ary tree of depth k where $M = \max(|\mathcal{G}_1|, \ldots, |\mathcal{G}_k|)$ is the maximum number of categories for any group attribute.

Desiderata We stipulate that a participatory system $h : \mathcal{X} \times \mathcal{R} \to \mathcal{Y}$ should meet three key requirements:

Protect Abstention: The system should ensure that individuals who opt out of personalization are assigned predictions that achieve the performance of a model trained without this information. This requires the system to outperform the performance of a generic model h_0. In practice, this requirement sets a baseline level of performance that model developers can expect when individuals opt out and ensures that the gains used to inform individuals are measured with respect to a model trained in good faith.

Promote Participation: The system should maximize the personalization gains for each reporting group $\boldsymbol{r} \in \mathcal{R}$. Formally, this requires systems that maximize the gains $\Delta_{\boldsymbol{r}}(h_{\boldsymbol{r}}, h_0)$ over a generic model h_0. In practice, this requirement promotes participation across reporting groups by improving the relative benefits of opting in.

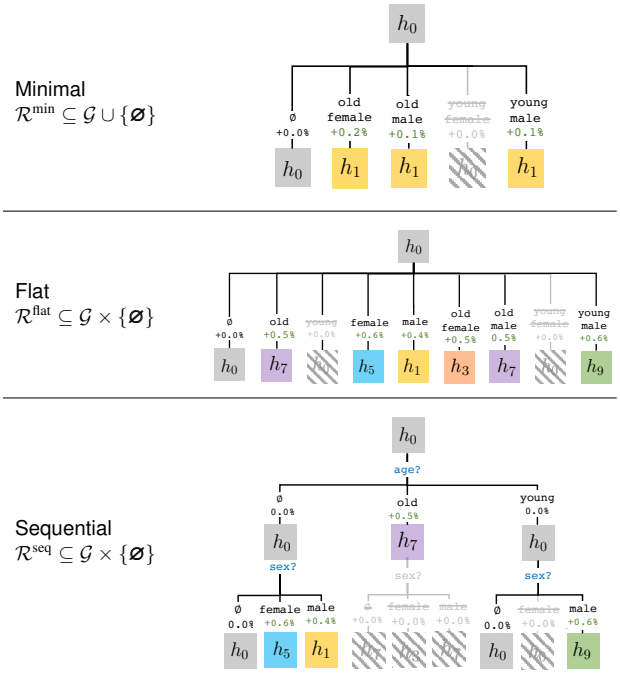

Figure 2: Participatory systems for a task with group attributes $\mathcal{G} = \texttt{sex} \times \texttt{age} = [\texttt{male}, \texttt{female}] \times [\texttt{old}, \texttt{young}]$. Each system allows a person to opt out of personalization informing their choice through comparisons between nested models. Systems limit unnecessary data collection by "pruning" reporting options that do not lead to gains – e.g., $[\texttt{young}, \texttt{female}]$ is pruned in all systems as it leads to a gain $\leq 0.0\%$.

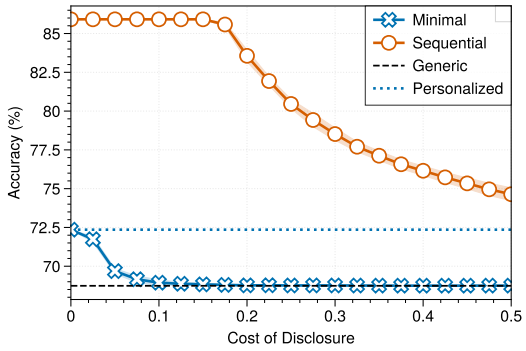

Figure 3: Performance profile of participatory systems for the `saps` dataset for individuals with group membership $\boldsymbol{g}_i = [\texttt{30+}, \texttt{HIV=+}]$. We show test performance with respect to participation in the target population. Here, we control participation by varying the cost of reporting in a simulated model for individual disclosure described in Appendix A. As shown, minimal and sequential systems always outperform the generic model regardless of participation. In regimes where the cost of reporting is low, participation is high. Consequently, a minimal system will achieve the same performance as a personalized model, and a sequential system to achieve the performance of the component model for this subpopulation. We provide additional details and results for other groups in Appendix A.

3. Learning Participatory Systems

In this section, we describe a model-agnostic algorithm to learn participatory systems that meet the requirements. Our procedure takes as input a pool of candidate models $\mathcal{M}$, a training dataset $\mathcal{D}^{\text{train}}$, and a validation $\mathcal{D}^{\text{valid}}$ dataset. It outputs a collection of participatory systems that ensure personalization gains across reporting groups.

By assigning personalized models over a reporting interface, our algorithm can produce the three types of participatory systems in Fig. 2. Our approach combines routines for generating a set of viable reporting interfaces (Line 1); assigning models over the interface (Line 3); and pruning the interface to limit data collection when it does not lead to gains (Line 4). We summarize the procedure in Algorithm 1 and present a detailed description of each routine in Appendix B.

Model Pool Our procedure takes as input a pool of personalized models that can be assigned to nodes in a reporting tree $\mathcal{M}$. At a minimum, $\mathcal{M}$ should contain two models: a personalized model h for individuals who opt into personalization and a generic model h_0 for individuals who opt out of personalization. The pool can contain models trained on different subsets of data (e.g., a model trained on female patients only), and fit from different model classes (e.g., linear classifiers and random forests). As the best personalized model can vary across intersectional groups, using a pool of models allows practitioners to personalize for each group. By default, we include "decoupled models" for each reporting group trained using only data for that group, as such models can perform well on heterogeneous subgroups [47, 55, 53].

Enumerating Viable Interfaces We call the ViableTrees routine in Line 1 to enumerate all viable M-vary trees for sequential systems. We only call this routine for sequential systems because $\mathcal{T}$ contains a single tree that is known a priori. This routine can return trees that obey custom constraints on sample size (e.g., only), as well as on the order of reporting (e.g., users who are `male` should report `age` before `HIV`). This routine will produce at most $|\mathcal{T}| \leq \prod_{i=1}^{k} i^{m^{k-i}}$ [29] trees. In general, this routine scales to tasks with ≤ 8 group attributes. Beyond this limit, one can reduce the size of the enumeration by specifying ordering constraints or a stopping condition based on the number of trees to enumerate before stopping. For a task with 3 binary group attributes $\mathcal{T}$ contains 24 3-ary trees of depth 3. Given a complete ordering of all 3 group attributes, $\mathcal{T}$ would contain 1 tree. The groups at the leaves of the tree should contain at least one positive label, one negative label, and $n_{\boldsymbol{r}} \geq 30$ samples to avoid overfitting. The routine can filter trees during generation to ensure that these criteria are met among the final set of candidates.

Algorithm 1 Learning Participatory Systems

Input: $\mathcal{D}^{\text{assign}} = \{(\boldsymbol{x}_i, \boldsymbol{g}_i, y_i)\}_{i=1}^{n^{\text{assign}}}$ *training dataset*
Input: $\mathcal{D}^{\text{prune}} = \{(\boldsymbol{x}_i, \boldsymbol{g}_i, y_i)\}_{i=1}^{n^{\text{prune}}}$ *validation dataset*
Input: $\mathcal{M} : \{h : \mathcal{X} \times \mathcal{G} \to \mathcal{Y}\}$ *pool of candidate models*
1: $\mathcal{T} \leftarrow \mathsf{ViableTrees}(\mathcal{G}, \mathcal{D})$ *$|\mathcal{T}| = 1$ for minimal & flat systems*
2: **for** $T \in \mathcal{T}$ **do**
3: $\quad T \leftarrow \mathsf{AssignModels}(T, \mathcal{M}, \mathcal{D}^{\text{assign}})$ *assign models*
4: $\quad T \leftarrow \mathsf{PruneLeaves}(T, \mathcal{D}^{\text{prune}})$ *prune models*
5: **end for**
Output $\mathcal{T}$ *collection of participatory systems that ensure reporting leads to gain*

Model Assignment We assign each reporting group a model using the AssignModels routine in Line 3. Given a reporting group, we consider all models that could use their group membership. Thus, a group that 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, ensuring that the system performs as well as a generic model in terms of the assignment metric on the assignment dataset. By default, we assign each reporting group a model from $\mathcal{M}$ that optimizes performance on the assignment sample $\mathcal{D}^{\text{assign}}$. 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).

Data Minimization by Pruning Line 1 may output trees where a person reports personal information without receiving a gain in performance. This can happen when we assign the same model to nested reporting groups (see e.g., the Flat system in Fig. 2 that assigns h_0 to $[\texttt{female}, \varnothing]$ and $[\texttt{female}, \texttt{young}]$), or when a model performs just as well as its parent (see e.g., the Sequential system in Fig. 2, where h_7 performs as well as h_3 for [female,old]). We ensure that a participatory system will not solicit data in such cases using the Prune routine in Line 4. This routine takes as inputs a tree T, the candidate models assigned to each node $\mathcal{M}$, and the pruning (validation) sample $\mathcal{D}^{\text{prune}}$. It outputs the pruned tree T. The routine performs a one-sided hypothesis test to check if each group $\boldsymbol{g}$ prefers the parent model h to a leaf model h':

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

Here, H_0 assumes that a group prefers h over h'. Thus, we reject H_0 when there is enough evidence to suggest that h' performs better for $\boldsymbol{r}$ on pruning data. The testing procedure should be chosen based on the performance metric used to evaluate personalization gains. In general, we can use a bootstrap hypothesis test [22]. However, there may exist more powerful tests for salient performance metrics [see e.g., 23, 21, 52, for accuracy and AUC].

On Performance Our procedure allows practitioners to learn systems for prediction tasks by specifying the performance metric used in assignment and pruning. A suitable performance metric should represent the exact gains we would show users (e.g., error for a diagnostic; AUC for triage; ECE for risk assessment). Pruning should be done on a held-out dataset to ensure these gains hold in deployment. Using a pool of models allows practitioners to optimize performance across groups, which translates to gains at the population level. For sequential systems, the procedure outputs all configurations, allowing practitioners to choose between systems on the basis of criteria not known at training time. For example, one can swap the trees to use a system that always requests age before HIV status. By default, we select the configuration that minimizes data collection across groups, such that the ordering of attributes results leads to the greatest number of data requests pruned.

On Computation Our approach provides practitioners with various options to learn participatory systems under a limited computational budget (e.g., one can train only two models and build a minimal system, or train a flat or sequential system with a limited number of models in the pool). The primary bottleneck in building participatory systems is *data* rather than *compute*. Given a finite sample dataset, we are limited in the number of categorical attributes used for personalization. This is because we require a minimum number of samples for each intersectional group to train a personalized model and evaluate its performance. Given that the number of intersectional groups increases exponentially with each additional attribute, we quickly enter a regime where we cannot train models for a given group (e.g., because we lack sufficient labels) or reliably evaluate its gain for assignment and pruning [see 44].

4. Experiments

We benchmark participatory systems and personalized models on real-world clinical prediction tasks. Our goal is to evaluate these approaches in terms of their performance, data usage, and consent in applications where individuals have a low cost of disclosure. We include code to reproduce our results in our anonymized repository.

4.1. Setup

We consider six classification tasks for clinical decision support where we must train a model that is personalized group attributes that are either protected or sensitive (`sex`, `age`, `HIV`). These are tasks where the information used for personalization is readily available, relevant to the prediction task, and is unlikely to be leaked or misused due to laws surrounding the confidentiality, privacy, and use of medical data [1]. Given these conditions, we expect individuals to have a low cost of disclosure – and therefore report personal

information so long as there is any benefit [12, 6].

We list the datasets for each prediction task in Table 2 and describe them in Appendix C. We split each dataset into a test sample (20%, used to evaluate out-of-sample performance), and a training sample (80%, used for training, assignment, pruning, and estimating gains to show users). We train three kinds of personalized models for each dataset:

- *Static*: These models are personalized using a one-hot encoding of group attributes (1Hot), and a one-hot encoding of intersectional groups (mHot)
- *Imputed*: These are variants of personalized models that facilitate consent using imputation. We construct these models by pairing static models with KNN-imputation (KNN-1Hot, KNN-mHot). We report results for these models for an extreme case where all individuals opt out of personalization. In practice, the performance of this imputation will fall between 1Hot (100% opt-in) and KNN-1Hot (100% opt-out).
- *Participatory*: These are participatory systems built using our approach. These include a minimal system built using 1Hot and its generic counterpart (Minimal); and flat and sequential systems built using 1Hot, mHot, and their generic counterparts (Flat, Seq).

We train all models – personalized models and the components of participatory systems – from the same model class. We evaluate all models on all datasets using the metrics in Table 1. We repeat these experiments four times, varying the model class (logistic regression, random forests) and the performance metric of interest (error rate, AUC). These variations are chosen to benchmark our approach in major prediction tasks (decision-making, ranking) and to understand the impact of model capacity on our results.

4.2. Results

We show results for logistic regression models and error rate in Table 2 and for other model classes and prediction tasks in Appendix D. In what follows, we discuss these results.

On Performance Gains Our results show that participatory systems can improve performance for all groups who provide personal data. We find that Flat and Seq achieve the best overall performance on all 6/6 datasets. These gains at a population level are often experienced across all groups who provide personal data, as shown by the fact that Flat and Seq improve both the worst-case and best-case gains of personalization on 5/6 datasets. In contrast, traditional approaches to personalization can improve performance at a group level while reducing performance at a group level. Our results highlight the prevalence of this effect – as we find that static methods exhibit rationality violations on 5/6 datasets Table 2 (c.f. 1/6 datasets for Minimal, Flat, or Seq).

Metric	Definition	Description
Overall Performance	$\sum_{\boldsymbol{g}\in\mathcal{G}} \frac{n_{\boldsymbol{g}}}{n} \hat{R}_{\boldsymbol{g}}(h_{\boldsymbol{g}})$	Population-level performance of a personalized system/model, computed as a weighted average over all groups
Overall Gain	$\sum_{\boldsymbol{g}\in\mathcal{G}} \frac{n_{\boldsymbol{g}}}{n} \hat{\Delta}_{\boldsymbol{g}}(h_{\boldsymbol{g}}, h_0)$	Population-level gain in performance of a personalized system/model over its generic counterpart
Group Gains	$\min_{\boldsymbol{g}\in\mathcal{G}} / \max_{\boldsymbol{g}\in\mathcal{G}} \hat{\Delta}_{\boldsymbol{g}}(h_{\boldsymbol{g}}, h_0)$	Range of gains of a personalized system/model over its generic counterpart across all groups
Rationality Violations	$\sum_{\boldsymbol{g}\in\mathcal{G}} 1[\text{reject } H_0]$	Number of rationality violations detected using a bootstrap hypothesis test with 100 resamples and a significance level of 10%. $H_0 : \Delta_{\boldsymbol{g}}(h, h_0) \geq 0$ and $H_A : \Delta_{\boldsymbol{g}}(h, h_0) < 0$
Imputation Risk	$\min_{\boldsymbol{g}\in\mathcal{G}} \hat{\Delta}_{\boldsymbol{g}}(h_{\boldsymbol{g}}, h_{\boldsymbol{g}'})$	Risk to performance of imputation, or the worst possible performance to a group given they are imputed with the attributes of group $\boldsymbol{g}'$. Relevant for static models only
Options Pruned	$\frac{\lvert\mathcal{R}\rvert - \lvert\mathcal{R}(h)\rvert}{\lvert\mathcal{R}\rvert}$	Number of reporting options pruned out of the total reporting options for a model or system. Here, $\mathcal{R}(h)$ denotes the options that are available after h has been pruned while $\mathcal{R}$ denotes options available before pruning.
Data Use	$\sum_{\boldsymbol{g}\in\mathcal{G}} \frac{n_{\boldsymbol{g}}}{n} \frac{\text{requested}(h, \boldsymbol{g})}{\dim(\mathcal{G})}$	Proportion of total group attributes k requested from h from each group, averaged over all groups in $\mathcal{G}$

Table 1: Overview of metrics used to evaluate performance, data usage, and consent. We report performance on a held-out test sample. We assume that individuals report group membership to static models, never report group membership to imputed models, and only report to participatory systems when disclosure leads to a positive gain. In the latter case, the gain shown to users is estimated using a validation set in the training sample.

In practice, the relative benefits in the performance of a participatory system over a traditional static model stem from (i) allowing users to opt out of instances of detrimental personalization, and (ii) assigning personalized predictions using multiple models (Flat and Seq). For example, on `cardio_eicu`, 1Hot improves performance at a population level but reduces performance at the group level. In particular, we find that 2 groups experience statistically significant rationality violations, meaning they would have been better off with a generic model that did not require them to report personal data. By comparing the performance of 1Hot to Minimal, we can gauge the performance gain that arises from allowing users to opt out of such instances (i.e., a reduction of test error from 22.4% to 21.7%). By comparing the performance on Minimal to Flat and Seq, we can gauge the performance gain that arises from the use of multiple models (i.e., a reduction of test error from 21.7% to 16.1%).

On Data Minimization Our results highlight how participatory systems limit data use. On 6/6 datasets, the participatory systems perform better across all groups while requesting less personal data. For example, on `cardio_eicu`, Seq reduces error by 6.3% compared to 1Hot while requesting, on average, 87.5% of the data needed by 1Hot. In general, participatory systems can reduce data where personalization doesn't improve performance, e.g., on `lungcancer`. Even as attributes like `sex` or `age` may be readily reported by patients for any performance benefit, the potential to curb

Dataset	Metrics	Static 1Hot	Static mHot	Imputed KNN-1Hot	Imputed KNN-mHot	Participatory Minimal	Participatory Flat	Participatory Seq
	Overall Performance	29.1%	29.3%	29.0%	27.9%	28.9%	**24.1%**	24.3%
apnea	Overall Gain	0.1%	-0.1%	0.2%	1.3%	0.3%	**5.1%**	4.9%
$n = 1152, d = 26$	Group Gains	-1.1% – 1.2%	-0.8% – 0.4%	-1.1% – 1.2%	-0.8% – 0.4%	0.0% – 1.2%	0.0% – 13.8%	-0.4% – 13.8%
$\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$	Rat. Violations	1	1	1	1	0	0	0
$\lvert\mathcal{G}\rvert = 6$ groups	Imputation Risk	-4.9%	-5.2%					
Ustun et al. [54]	Options Pruned	0/6	0/6	0/12	0/12	4/7	5/12	6/12
	Data Use	100.0%	100.0%	0.0%	0.0%	33.3%	83.3%	58.3%
	Overall Performance	22.4%	21.9%	22.2%	21.4%	21.7%	**16.1%**	**16.1%**
cardio_eicu	Overall Gain	0.2%	0.7%	0.4%	1.2%	0.9%	**6.5%**	**6.5%**
$n = 1341, d = 49$	Group Gains	-2.1% – 3.2%	-1.9% – 5.1%	-2.1% – 3.2%	-1.9% – 5.1%	0.0% – 3.2%	-1.9% – 17.8%	-1.9% – 17.8%
$\mathcal{G} = \{\texttt{sex}, \texttt{age}\}$	Rat. Violations	2	1	2	1	0	1	1
$\lvert\mathcal{G}\rvert = 4$ groups	Imputation Risk	-5.3%	-2.1%					
Pollard et al. [46]	Options Pruned	0/4	0/4	0/9	0/9	2/5	3/9	3/9
	Data Use	100.0%	100.0%	0.0%	0.0%	50.0%	100.0%	87.5%
	Overall Performance	19.5%	19.3%	19.2%	20.7%	19.2%	**18.0%**	**18.0%**
cardio_mimic	Overall Gain	-0.3%	-0.1%	-0.1%	-1.6%	0.0%	**1.2%**	**1.2%**
$n = 5289, d = 49$	Group Gains	-0.8% – 0.3%	-0.5% – 0.3%	-0.8% – 0.3%	-0.5% – 0.3%	0.0% – 0.0%	0.0% – 3.3%	0.0% – 3.3%
$\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$	Rat. Violations	2	2	2	2	0	0	0
$\lvert\mathcal{G}\rvert = 4$ groups	Imputation Risk	-1.5%	-0.8%					
[33]	Options Pruned	0/4	0/4	0/9	0/9	4/5	6/9	5/9
	Data Use	100.0%	100.0%	0.0%	0.0%	0.0%	50.0%	50.0%
	Overall Performance	37.0%	36.7%	37.0%	36.9%	37.0%	36.6%	**36.1%**
coloncancer	Overall Gain	0.1%	0.4%	0.1%	0.2%	0.1%	0.5%	**1.0%**
$n = 29211, d = 72$	Group Gains	-0.4% – 0.3%	-0.1% – 1.1%	-0.4% – 0.3%	-0.1% – 1.1%	0.0% – 0.3%	0.0% – 1.7%	0.2% – 1.7%
$\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$	Rat. Violations	1	0	1	0	0	0	0
$\lvert\mathcal{G}\rvert = 6$ groups	Imputation Risk	-1.4%	-0.9%					
NCI [41]	Options Pruned	0/6	0/6	0/12	0/12	5/7	7/12	5/12
	Data Use	100.0%	100.0%	0.0%	0.0%	16.7%	50.0%	75.0%
	Overall Performance	19.6%	19.6%	19.9%	19.8%	19.5%	18.9%	**18.9%**
lungcancer	Overall Gain	-0.1%	-0.1%	-0.3%	-0.2%	0.0%	0.6%	**0.6%**
$n = 120641, d = 84$	Group Gains	-0.4% – 0.2%	-0.3% – 0.2%	-0.4% – 0.2%	-0.3% – 0.2%	0.0% – 0.0%	0.0% – 0.9%	0.3% – 0.9%
$\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$	Rat. Violations	4	4	4	4	0	0	0
$\lvert\mathcal{G}\rvert = 6$ groups	Imputation Risk	-0.5%	-0.5%					
NCI [41]	Options Pruned	0/6	0/6	0/12	0/12	6/7	3/12	7/12
	Data Use	100.0%	100.0%	0.0%	0.0%	0.0%	83.3%	58.3%
	Overall Performance	20.4%	20.7%	20.4%	29.4%	20.4%	**11.1%**	11.1%
saps	Overall Gain	1.3%	1.0%	1.3%	-7.7%	1.3%	**10.6%**	10.6%
$n = 7797, d = 36$	Group Gains	0.0% – 3.6%	0.0% – 2.7%	0.0% – 3.6%	0.0% – 2.7%	0.0% – 3.6%	4.3% – 17.2%	4.3% – 17.2%
$\mathcal{G} = \{\texttt{HIV}, \texttt{age}\}$	Rat. Violations	0	0	0	0	0	0	0
$\lvert\mathcal{G}\rvert = 4$ groups	Imputation Risk	0.0%	-2.4%					
Allyn et al. [4]	Options Pruned	0/4	0/4	0/9	0/9	1/5	1/9	3/9
	Data Use	100.0%	100.0%	0.0%	0.0%	75.0%	100.0%	75.0%

Table 2: Performance and data use of participatory systems for all datasets. We summarize each metric in Table 1, describe the datasets in Appendix C, and include results for other model classes and prediction tasks in Appendix D.

data use is valuable when there is a tangible cost associated with data collection – e.g., when models make use of the outcome from a diagnostic test or rating scale for a mental disorder that must be administered by a clinician [48]. Our results show that the opportunities for data minimization may vary substantially across prediction tasks. On `apnea` for example, we can prune 6 reporting options for a Seq system for decision making (error) but 4 reporting for Seq when we optimize for ranking (AUC) (see Appendix D).

On Facilitating and Informing Consent Our results highlight the benefits of flat and sequential systems for facilitating and informing consent. These systems provide more opportunities for consent by allowing users to report a subset of group attributes On `saps`, for example, we see that users who are HIV positive can report $[\texttt{30+}, \varnothing]$ or $[\texttt{30+}, \texttt{HIV+}]$. In Table 2, we show these opportunities through total number of reporting options available to users at prediction time.

Although flat and sequential systems provide the same reporting options, they differ in their capacity to inform consent. In a flat system, users may attempt to gauge the marginal benefit of reporting a specific attribute by comparing the gains between reporting options. For example, in Fig. 4), users who are HIV positive would see a gain of 3.7% for reporting $[\varnothing, \texttt{HIV+}]$ and 16.7% for reporting $[\texttt{30+}, \texttt{HIV+}]$, thus concluding that the marginal gain of reporting age is $16.7\% - 3.7\% = 13.0\%$. This estimate presumes that the gains of 3.7% were distributed equally across age groups. Sequential systems naturally overcome such issues by informing users of the exact gains for partial reporting. In the sequential system, group $[\texttt{30+}, \texttt{HIV+}]$

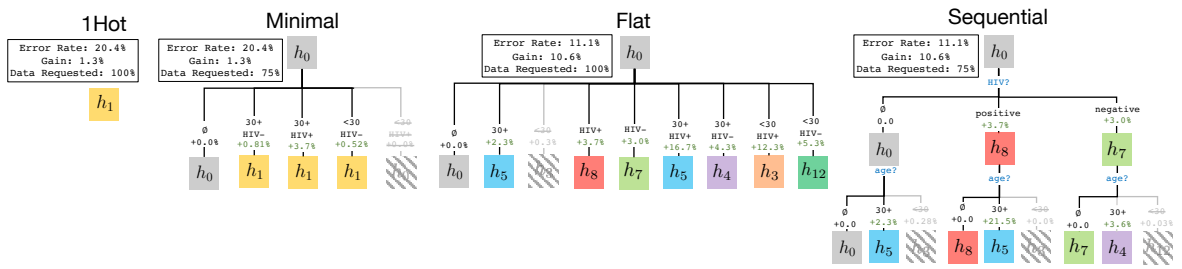

Figure 4: Participatory systems for the `saps` dataset. These models are trained to predict ICU mortality for groups defined by $\mathcal{G} =$ `HIV` $\times$ `age` $=$ `[+,-]` $\times$ `[<30, 30+]`. Here, h_0 denotes the generic model, h_1 denotes a 1Hot model fit with a one-hot encoding of $\mathcal{G}$, and $h_2 \cdots h_n$ are 1Hot and mHot models fit for reporting groups. Grey stripes indicate pruned reporting options. Numbers above each box indicate the gain with reference to the parent node. For example, in the Sequential system, group `(HIV+, 30+)` sees an estimated 21.5% error reduction for `age` after having reported HIV. In contrast, group `(HIV+, <30)` sees no gain from reporting age in addition to HIV status, and this option is pruned.

would see that the marginal gain of reporting age is 21.5%. Likewise, group [`<30, HIV+`] would see that the marginal gain of reporting age is 0.0%.

On the Pitfalls of Imputation One of the simplest approaches to allow individuals to opt out of personalization is to pair a personalized model with an imputation technique. Although this approach can facilitate consent, it does not meet the requirements in 2. Consider a personalized model that exhibits "worsenalization" in Fig. 1. Even if one could correctly impute the group membership for every person, individuals would still receive more accurate predictions from a generic model h_0. In practice, imputation can perform unreliably – as individuals who opt out of reporting their group membership to a personalized model may be imputed the group membership for a group that is assigned considerably different predictions. In such cases, opting out may be beneficial, making it difficult for model developers to promote participation while informing consent. Our results highlight the relative prevalence of this effect across model classes and prediction tasks. For example, on `cardio_eicu` our estimate of the "risk of imputation" is -5.3%, indicating that groups can experience an error rate up to 5.3% greater if their values are incorrectly imputed at the level of intersectional groups. Our results for KNN-1Hot show that this predicted loss in performance can be realized in practice using KNN-imputation as we find that the imputed system leads to rationality violations on 5/6 datasets.

On the Value of a Model Agnostic Approach As expected, we find that complex model class can produce considerable changes in overall accuracy – e.g., we can reduce overall test error for a personalized model from 20.4% to 14.1% on `saps` by training a random forest rather than a logistic regression model (see Appendix D). However, a gain in overall performance does not always translate to gains at the group level. On `saps`, using a random forest also introduces a rationality violation for one group. These findings highlight the value of a model-agnostic approach, in which we can use a variety of models to achieve better performance while mitigating harm. For example, we can ensure generalization across reporting groups – e.g., by a generic model fit from a complex model class, and personalized models fit from a simpler model class.

5. Concluding Remarks

In this work, we introduced a new family of prediction models that allow individuals to report personal data at prediction time. Our systems can facilitate and inform consent in a way that can produce large improvements in performance and privacy for each group that reports personal data.

The systems in this work should be seen as foundational machinery for informing consent. In practice, the viability of reaping these benefits will hinge on individual preferences for disclosure, which can change based on the information solicited, the outcome predicted, and the ability to inform users effectively of these impacts [6]. Implementing these systems will require developing tailored approaches to communicate the gains of personalization (e.g., communicating risk and uncertainty).

One common concern is that allowing individuals to opt out of personalization could prevent us from collecting data that could be used to monitor or improve its performance. While this is a real possibility, the core issue stems from a lack of transparency surrounding the *purpose* of data collection [42]. If the purpose of data collection is to monitor or improve a model, then individuals could be given the ability to report this information voluntarily for the sake of auditing or training. If the purpose of data collection is personalization, then individuals are within their right to opt out.

References

[1] Health insurance portability and accountability act of 1996, Aug 1996. URL https://www.govinfo.gov/app/details/PLAW-104publ191.

[2] *Advances in Neural Information Processing Systems 19*, 2007. doi: 10.7551/mitpress/7503.003.0011. URL http://dx.doi.org/10.7551/mitpress/7503.003.0011.

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

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

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

[6] Anderson, C. L. and Agarwal, R. The digitization of healthcare: Boundary risks, emotion, and consumer willingness to disclose personal health information. *Information Systems Research*, 22(3):469–490, 2011. ISSN 10477047, 15265536. URL http://www.jstor.org/stable/23015590.

[7] Arellano, A. M., Dai, W., Wang, S., Jiang, X., and Ohno-Machado, L. Privacy policy and technology in biomedical data science. *Annual review of biomedical data science*, 1: 115, 2018.

[8] Atan, O., Whoiles, W., and Schaar, M. Data-driven online decision making with costly information acquisition. *Arxiv*, 02 2016.

[9] Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[10] Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., and Turner, E. Americans and privacy: Concerned, confused and feeling lack of control over their personal information. *Pew Research Center: Internet, Science and Tech*, 2019.

[11] Awad, N. F. and Krishnan, M. S. The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS quarterly*, pp. 13–28, 2006.

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

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

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

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

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

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

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

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

[20] Connors, A. F., Dawson, N. V., Desbiens, N. A., Fulkerson, W. J., Goldman, L., Knaus, W. A., Lynn, J., Oye, R. K., Bergner, M., Damiano, A., 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.

[21] DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pp. 837–845, 1988.

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

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

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

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

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

[27] 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.

[28] Globus-Harris, I., Kearns, M., and Roth, A. An algorithmic framework for bias bounties. *2022 ACM Conference on Fairness, Accountability, and Transparency*, Jun 2022. doi: 10.1145/3531146.3533172. URL http://dx.doi.org/10.1145/3531146.3533172.

[29] Graham, R. L., Knuth, D. E., Patashnik, O., and Liu, S. Concrete mathematics: a foundation for computer science. *Computers in Physics*, 3(5):106–107, 1989.

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

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

[32] Hollenberg, S. Cardiogenic shock. In *Intensive Care Medicine*, pp. 447–458. Springer, 2003.

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

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

[35] Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pp. 2564–2572, 2018.

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

[37] Korsh, J. F. Loopless generation of k-ary tree sequences. *Inf. Process. Lett.*, 52(5):243–247, dec 1994. ISSN 0020-0190. doi: 10.1016/0020-0190(94)00149-9. URL https://doi.org/10.1016/0020-0190(94)00149-9.

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

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

[40] Lundberg, I., Narayanan, A., Levy, K., and Salganik, M. J. Privacy, ethics, and data access: A case study of the fragile families challenge. *Socius*, 5:2378023118813023, 2019.

[41] NCI, DCCPS, S. R. P. Surveillance, epidemiology, and end results (seer) program research data (1975-2016), 2019. URL www.seer.cancer.gov.

[42] 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.

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

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

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

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

[47] Saar-Tsechansky, M. and Provost, F. Handling missing values when applying classification models. 2007.

[48] Sharp, R. The hamilton rating scale for depression. *Occupational Medicine*, 65(4):340–340, 2015.

[49] Steyerberg, E. W. et al. *Clinical prediction models*. Springer, 2019.

[50] Struck, A. F., Ustun, B., Ruiz, A. R., Lee, J. W., LaRoche, S. M., Hirsch, L. J., Gilmore, E. J., Vlachy, J., Haider, H. A., and Rudin, C. Association of an electroencephalography-based risk score with seizure probability in hospitalized patients. *JAMA neurology*, 74(12):1419–1424, 2017.

[51] Struck, A. F., Tabaeizadeh, M., Schmitt, S. E., Ruiz, A. R., Swisher, C. B., Subramaniam, T., Hernandez, C., Kaleem, S., Haider, H. A., and Cissé, A. F. Assessment of the validity of the 2helps2b score for inpatient seizure risk prediction. *JAMA neurology*, 77(4):500–507, 2020.

[52] Sun, X. and Xu, W. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, 2014. doi: 10.1109/LSP.2014.2337313.

[53] Suriyakumar, V. M., Ghassemi, M., and Ustun, B. When personalization harms: Reconsidering the use of group attributes in prediction. In *arXiv preprint*, 2022.

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

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

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

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

[58] Yu, S., Krishnapuram, B., Rosales, R., and Rao, R. B. Active sensing. In *AISTATS*, 2009.

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

Table of Contents

A. Supporting Material for Performance Profile in Section 2

A.1. Disclosure Model

The performance of participatory systems will depend on individual reporting decisions. In what follows, we characterize performance under a generalized model of individual disclosure. Given a participatory system $h : \mathcal{X} \times \mathcal{R} \to \mathcal{Y}$, we assume that each person will report group membership as:

$$\boldsymbol{r}_i \in \underset{\boldsymbol{r} \in \mathcal{R}}{\operatorname{argmax}}\, u_i(\boldsymbol{r}; h)$$

Here, the utility function can be

$$u_i(\boldsymbol{r}; h) = b_i(\boldsymbol{r}; h) - c_i(\boldsymbol{r}),$$

where $c_i(\cdot)$ and $b_i(\cdot)$ denote their cost and benefit of disclosure, respectively. We assume that costs increase monotonically with information that is disclosed so that $c_i(\boldsymbol{r}) \geq 0$ for all $\boldsymbol{r} \in \mathcal{R}$ and $c_i(\boldsymbol{r}) \leq c_i(\boldsymbol{r}')$ for $\boldsymbol{r} \subseteq \boldsymbol{r}'$. We assume that benefits increase monotonically with true risk so that $b_i(\boldsymbol{r}, h) > b_i(\boldsymbol{r}', h)$ when $R_{\boldsymbol{r}}(h(\boldsymbol{x}_i, \boldsymbol{r})) < R_{\boldsymbol{r}}(h(\boldsymbol{x}_i, \boldsymbol{r}'))$.

The following remarks apply to any participatory system $h : \mathcal{X} \times \mathcal{R} \to Y$ that include a personalized model $h : \mathcal{X} \times \mathcal{G} \to \mathcal{Y}$ and a generic model $h_0 : \mathcal{X} \to \mathcal{Y}$ as its components.

- Every participatory system h will perform as well as a generic model h_0. When a personalized model h requires users to report information detrimental to performance (see Fig. 1), users incur a cost of the disclosure without receiving a benefit. In such instances, a minimal system $h : \mathcal{X} \times \mathcal{R}^{\min} \to Y$ would allow users to opt out of detrimental personalization and receive predictions from a generic model.
- Every participatory system h with more reporting options will perform better. Given that utility can only increase with the number of reporting options, the maximum utility for each person will exceed that of a minimal system. Thus, flat and sequential systems will perform better than a minimal system.
- The best-case performance of any participatory system will exceed the performance of any of its components. Thus, we are guaranteed that any participatory system will outperform a traditional personalized model so long as it is considered a component.

A.2. Simulation Results

We simulate the result of increased reporting costs for each intersectional group by sampling an individual's reporting cost from a uniform distribution. Thus for each individual i, we sample c_i as.

$$c_i \sim \text{Uniform}(0, \gamma), \text{ where } \gamma \in [0, 0.2]$$

For each value of γ, we sample reporting costs 10 times and average over the per group performance error for each sampled cost. Figure 3 presents the participation profiles for various levels of cost of disclosure.

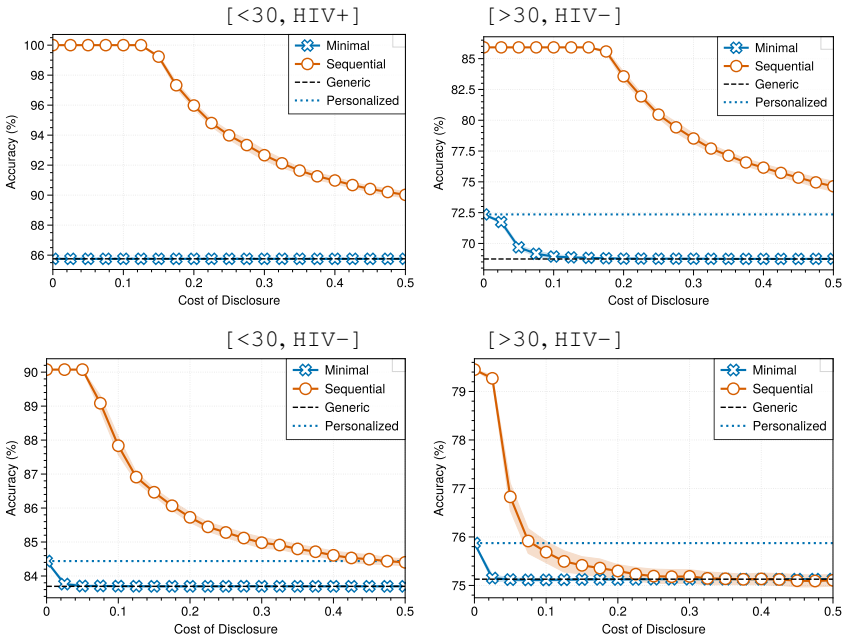

Figure 5: Performance profiles of the simulations performed for each intersectional group in the `saps` dataset. The sequential system outperforms static personalized systems when all group attributes are reported. When the cost of reporting is high, the sequential system still outperforms minimally personalized systems as evidenced by higher accuracy at varying reporting cost thresholds.

B. Description of Routines in Algorithm 1

In what follows, we describe the routines called in Algorithm 1.

B.1. Enumeration Routine

We summarize the Enumeration routine in Algorithm 2. Algorithm 2 takes as input a set of group attributes $\mathcal{G}$ and a dataset $\mathcal{D}$ and outputs a collection of reporting trees $\mathcal{T}$ that obey ordering and plausibility constraints. The procedure uses the attributes in $\mathcal{G}$ to enumerate all possible reporting trees through a recursive branching process. There exists only one possible tree for Minimal and Flat systems, so the full routine is only called for building Sequential systems. Given a set of group attributes, the routine is called for each attribute that has yet been considered in the tree Line 5, ensuring a complete enumeration.

Enumerating all possible trees ensures we can recover the best tree given the selection criteria and allows practitioners to choose between models based on other criteria. We generate trees that meet plausibility constraints based on the dataset, such as having at least one negative and one positive sample and at least s total samples at each leaf. In settings constrained by computational resources, we can impose additional stopping criteria and modify the ordering to enumerate more plausible trees first or exclusively (e.g., by changing the ordering of $\mathcal{G}$ or imposing constraints in VALIDASSIGNMENTS).

Algorithm 2 Enumerate All Possible Reporting Trees for Reporting Options $\mathcal{G}$

1: **procedure** VIABLETREES($\mathcal{G}, \mathcal{D}$)
2: **if** Minimal or Flat **return** $[T_{\{\varnothing\}\times\mathcal{G}}]$ *return a single tree with all reporting options for minimal and flat systems*
3: **if** $\dim(\mathcal{G}) = 1$ **return** $[T_{\mathcal{G}}]$ *base case: we are left with only a single attribute on which to branch*
4: $\mathcal{T} \leftarrow [\,]$
5: **for** each group attribute $\mathcal{A} \in [\mathcal{G}_1, \ldots, \mathcal{G}_k]$ **do**
6: $T_{\mathcal{A}} \leftarrow$ reporting tree of depth 1 with $|\mathcal{A}|$ leaves
7: $\mathcal{S} \leftarrow$ ViableTrees$(\mathcal{G} \setminus \mathcal{A}, \mathcal{D})$ *all subtrees using all attributes except $\mathcal{A}$*
8: **for** Π in ValidAssignments$(\mathcal{S}, \mathcal{A}, \mathcal{D})$ **do**: *each assignment is a permutation of $|\mathcal{A}|$ to leaves of $T_{\mathcal{A}}$*
9: $\mathcal{T} \leftarrow \mathcal{T} \cup T_{\mathcal{A}}$.assign$(\Pi)$ *extends the tree by assigning subtrees to each leaf*
10: **end for**
11: **end for**
12: **return** $\mathcal{T}$, collection of all reporting trees for group attributes $\mathcal{G}$ that obey plausibility and ordering constraints
13: **end procedure**

B.2. Assignment Routine

We summarize the routine for AssignModels procedure in Algorithm 3. Algorithm 3 takes as inputs a reporting tree T, a pool candidate models $\mathcal{M}$, and an assignment (training) dataset $\mathcal{D}$ and outputs a tree T that maximizes the gains of reporting group information. The pool of candidate models is filtered to viable models for each reporting group. Since the pool of candidate models includes the generic model h_0, each reporting group will have at least one viable model. We assign each reporting group the best-performing model on the training set, and default to the generic model h_0 when a better-performing personalized model is not found. We assign performance on the training set and then prune using performance on the validation set to avoid biased gain estimations.

Algorithm 3 Assigning Models

procedure ASSIGNMODELS($T, \mathcal{M}, \mathcal{D}$)
  $Q \leftarrow [T.root]$ *initialize with the root of the tree, reporting group* $\varnothing$
  while Q is not empty **do**
    $\boldsymbol{r} \leftarrow Q.pop()$
    $\mathcal{M}_{\boldsymbol{r}} \leftarrow$ ViableModels($\mathcal{M}, \boldsymbol{r}$) *filter* $\mathcal{M}$ *to models that can be assigned to* $\boldsymbol{r}$
    $h^* \leftarrow \underset{h \in \mathcal{M}_{\boldsymbol{r}}}{\operatorname{argmin}} \hat{R}_{\boldsymbol{r}}(h, \mathcal{D})$ *assign the model with the best training performance*
    T.set_model($\boldsymbol{r}, h^*$)
    for $\boldsymbol{r}' \in T$.get_subgroups($\boldsymbol{r}$) **do** *iterate through the children reporting groups of* $\boldsymbol{r}$
      $Q.enqueue(\boldsymbol{r}')$
    end for
  end while
  return T that maximizes gain for each reporting group
end procedure

B.3. Pruning Routine

We summarize the routine used for the PruneLeaves procedure in Algorithm 1 below. The inputs to PruneLeaves are the tree T and the pruning sample $\mathcal{D}$ (validation) and the output is a pruned tree T that ensures data collection leads to gain. In PruneLeaves, we consider each leaf individually and prune it based on a hypothesis test comparing the model at the leaf with the model at the parent on the validation set for the group at the leaf. We perform a hypothesis test and calculate the probability of the observations under the assumption that the parent model performs as well as the child model for the group. For example, we would test if model h_7 assigned to group (`female`, `old`) outperforms the model h_0 assigned to the parent (`female`, $\varnothing$) over the validation set $\mathcal{D}$ by performing Test((`female`, `old`),$h_7, h_0, \mathcal{D}$). We use the McNemar test for accuracy [23] and the Delong test for AUC [21, 52]. In general, we can use a bootstrap hypothesis test [22]. Here the null hypothesis is that the parent model is as good as the leaf model and the alternative hypothesis is that the leaf model outperforms the parent model. In our experiments, we prune the node when we cannot reject the null hypothesis with a p-value less than 0.05.

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

Algorithm 4 Pruning Participatory Systems

procedure PRUNELEAVES($T, \mathcal{D}$)
  $Stack \leftarrow [T.\text{leaves}]$ *initialize stack with all leaves*
  while $Stack$ is not empty **do**
    $\boldsymbol{r} \leftarrow Stack.pop()$
    $h \leftarrow T.\text{get\_model}(\boldsymbol{r})$
    $h' \leftarrow T.\text{get\_model}(\boldsymbol{r}_{\text{parent}})$
    if not Test($\boldsymbol{r}, h, h', \mathcal{D}$) **then** *test gains to see if parent model is as good as leaf model*
      $T.\text{prune}(leaf)$
    end if
    if $T.\text{get\_children}(\boldsymbol{r}_{\text{parent}})$ is empty **then** *consider pruning the parent if the parent has become a leaf*
      $Stack.enqueue(\boldsymbol{r}_{\text{parent}})$
    end if
  end while
  return T that ensures data collection leads to gain
end procedure

C. Description of Datasets

We include additional information about the datasets used in Section 4.

Dataset	Reference	Outcome Variable	n	d	m	$\mathcal{G}$
cardio_eicu	Pollard et al. [46]	patient with cardiogenic shock dies	1,341	49	4	{age, sex}
cardio_mimic	Johnson et al. [33]	patient with cardiogenic shock dies	5,289	49	4	{age, sex}
coloncancer	NCI [41]	patient dies within 5 years	29,211	72	6	{age, sex}
lungcancer	NCI [41]	patient dies within 5 years	120,641	84	6	{age, sex}
saps	Allyn et al. [4]	ICU mortality	7,797	36	4	{age, HIV}
apnea	Ustun et al. [54]	patient has obstructive sleep apnea	1,152	28	6	{age, sex}

Table 3: Overview of datasets used to fit clinical prediction models in Section 4. Here: n denotes the number of examples in each dataset; d denotes the number of features; $\mathcal{G}$ denotes the group attributes that are used for personalization; and $m = |\mathcal{G}|$ denotes the number of intersectional groups. Each dataset is de-identified and available to the public. The cardio_eicu, cardio_mimic, lungcancer datasets require access to public repositories listed under the references. The saps and apnea datasets must be requested from the authors. The support dataset can be downloaded directly from the URL below.

cardio_eicu and cardio_mimic Cardiogenic shock is an acute condition in which the heart cannot provide sufficient blood to the vital organs [32]. These datasets are designed to predict cardiogenic shock for patients in intensive care. Each dataset contains the same features, group attributes, and outcome variables for patients in different cohorts. The cardio_eicu dataset contains records for a cohort of patients in the Collaborative Research Database V2.0 [46]. The cardio_eicu dataset contains records for a cohort of patients in the MIMIC-III [33] database. Here, the outcome variable indicates whether a patient in the ICU with cardiogenic shock will die while in the ICU. The features encode the results of vital signs and routine lab tests (e.g. systolic BP, heart rate, hemoglobin count) that were collected up to 24 hours before the onset of cardiogenic shock.

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 [41]. Here, the outcome variable indicates if a patient dies within five years from any cause, and 16.9% of patients died within the first five years from diagnosis. The cohort includes 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).

coloncancer We consider a cohort of 120,641 patients who were diagnosed with colorectal cancer between 2004-2016 and monitored as part of the National Cancer Institute SEER study [41]. Here, the outcome variable indicates if a patient dies within five years from any cause, and 42.1% of patients die within the first five years from diagnosis. The cohort includes patients from Greater California. 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).

saps The Simplified Acute Physiology Score II (SAPS II) score predicts the risk of mortality of critically-ill patients in intensive care [38]. 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.

apnea We use the obstructive sleep apnea (OSA) dataset outlined in Ustun et al. [54]. This dataset includes a cohort of 1,152 patients where 23% have OSA. We use all available features (e.g. BMI, comorbidities, age, and sex) and binarize them, resulting in 26 binary features.

support The support dataset is derived from a study of the survival risk of critically-ill patients who were discharged from the ICU conducted by Connors et al. [20]. 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://hbiostat.org/data/.

D. Experimental Results for Model Classes and Prediction Tasks

In this Appendix, we present the results of our experiments for additional model classes and prediction tasks. We produce these results using the setup described in Section 4.1, and summarize them in the same way as Table 2. We refer to them in our discussion in Section 4.2.

D.1. Logistic Regression for Ranking (AUC)

		Static		Imputed		Participatory		
Dataset	Metrics	1Hot	mHot	KNN-1Hot	KNN-mHot	Minimal	Flat	Seq
apnea $n = 1152, d = 26$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathcal{G}\rvert = 6$ groups Ustun et al. [54]	Overall Performance	0.774	0.774	0.776	0.776	0.776	**0.851**	**0.851**
	Overall Gain	-0.002	-0.002	0.000	-0.000	0.000	**0.074**	**0.074**
	Group Gains	-0.002 – 0.002	-0.002 – 0.003	-0.002 – 0.002	-0.002 – 0.003	0.000 – 0.002	0.004 – 0.115	0.004 – 0.115
	Max Disparity	0.004	0.005	0.004	0.005	0.002	0.111	0.111
	Rat. Violations	2	2	2	2	0	0	0
	Imputation Risk	-0.002	-0.002					
	Options Pruned	0/6	0/6	0/12	0/12	5/7	4/12	4/12
	Data Use	100.0%	100.0%	0.0%	0.0%	16.7%	100.0%	83.3%
cardio_eicu $n = 1341, d = 49$ $\mathcal{G} = \{\texttt{sex}, \texttt{age}\}$ $\lvert\mathcal{G}\rvert = 4$ groups Pollard et al. [46]	Overall Performance	0.858	0.857	0.859	0.857	0.857	**0.923**	**0.923**
	Overall Gain	0.001	-0.000	0.002	0.000	0.000	**0.067**	**0.067**
	Group Gains	-0.001 – 0.002	-0.001 – 0.002	-0.001 – 0.002	-0.001 – 0.002	0.000 – 0.000	0.008 – 0.094	0.008 – 0.094
	Max Disparity	0.003	0.003	0.003	0.003	0.000	0.087	0.087
	Rat. Violations	2	1	2	1	0	0	0
	Imputation Risk	-0.001	-0.001					
	Options Pruned	0/4	0/4	0/9	0/9	4/5	3/9	2/9
	Data Use	100.0%	100.0%	0.0%	0.0%	0.0%	100.0%	100.0%
cardio_mimic $n = 5289, d = 49$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathcal{G}\rvert = 4$ groups [33]	Overall Performance	0.876	0.876	0.878	0.876	0.876	**0.896**	**0.896**
	Overall Gain	-0.000	-0.000	0.002	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	0.000 – 0.000	0.005 – 0.034	0.005 – 0.034
	Max Disparity	0.001	0.001	0.001	0.001	0.000	0.028	0.028
	Rat. Violations	0	1	0	1	0	0	0
	Imputation Risk	-0.000	-0.000					
	Options Pruned	0/4	0/4	0/9	0/9	4/5	3/9	2/9
	Data Use	100.0%	100.0%	0.0%	0.0%	0.0%	100.0%	100.0%
coloncancer $n = 29211, d = 72$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathcal{G}\rvert = 6$ groups NCI [41]	Overall Performance	0.685	0.685	0.683	0.683	0.685	**0.700**	0.700
	Overall Gain	0.001	0.002	-0.000	-0.000	0.001	**0.016**	0.016
	Group Gains	-0.001 – 0.002	-0.001 – 0.001	-0.001 – 0.002	-0.001 – 0.001	0.000 – 0.001	0.001 – 0.021	0.001 – 0.021
	Max Disparity	0.003	0.002	0.003	0.002	0.001	0.020	0.020
	Rat. Violations	3	2	3	2	0	0	0
	Imputation Risk	-0.001	-0.002					
	Options Pruned	0/6	0/6	0/12	0/12	5/7	2/12	5/12
	Data Use	100.0%	100.0%	0.0%	0.0%	16.7%	100.0%	75.0%
lungcancer $n = 120641, d = 84$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathcal{G}\rvert = 6$ groups NCI [41]	Overall Performance	0.855	0.855	0.852	0.854	0.855	**0.861**	0.861
	Overall Gain	0.001	0.001	-0.002	0.000	0.001	**0.006**	0.006
	Group Gains	-0.000 – 0.000	-0.000 – 0.000	-0.000 – 0.000	-0.000 – 0.000	0.000 – 0.000	0.001 – 0.012	0.001 – 0.012
	Max Disparity	0.001	0.001	0.001	0.001	0.000	0.011	0.011
	Rat. Violations	2	2	2	2	1	0	0
	Imputation Risk	-0.000	-0.000					
	Options Pruned	0/6	0/6	0/12	0/12	4/7	2/12	2/12
	Data Use	100.0%	100.0%	0.0%	0.0%	33.3%	100.0%	91.7%
saps $n = 7797, d = 36$ $\mathcal{G} = \{\texttt{HIV}, \texttt{age}\}$ $\lvert\mathcal{G}\rvert = 4$ groups Allyn et al. [4]	Overall Performance	0.875	0.877	0.875	0.857	0.875	**0.960**	0.960
	Overall Gain	0.010	0.011	0.010	-0.008	0.009	**0.095**	0.095
	Group Gains	-0.000 – 0.016	-0.002 – 0.019	-0.000 – 0.016	-0.002 – 0.019	0.000 – 0.016	0.035 – 0.141	0.035 – 0.141
	Max Disparity	0.017	0.021	0.017	0.021	0.016	0.106	0.106
	Rat. Violations	1	1	1	1	0	0	0
	Imputation Risk	-0.000	-0.002					
	Options Pruned	0/4	0/4	0/9	0/9	1/5	2/9	3/9
	Data Use	100.0%	100.0%	0.0%	0.0%	75.0%	100.0%	87.5%

Table 4: Overview of performance, data use, and consent for all personalized models and systems on all datasets as measured by **test auc**. We show the performance of models and systems built using **logistic regression**.

D.2. Random Forests for Decision-Making (Error)

Dataset	Metrics	Static 1Hot	Static mHot	Imputed KNN-1Hot	Imputed KNN-mHot	Participatory Minimal	Participatory Flat	Participatory Seq
apnea $n = 1152, d = 26$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathcal{G}\rvert = 6$ groups Ustun et al. [54]	Overall Performance	26.3%	26.0%	25.9%	27.4%	26.3%	**12.2%**	**12.2%**
	Overall Gain	1.5%	1.8%	1.9%	0.4%	1.5%	**15.6%**	**15.6%**
	Group Gains	-0.8% – 4.2%	0.4% – 3.8%	-0.8% – 4.2%	0.4% – 3.8%	0.0% – 4.2%	5.3% – 22.2%	5.3% – 22.2%
	Max Disparity	5.0%	3.4%	5.0%	3.4%	4.2%	16.9%	16.9%
	Rat. Violations	1	0	1	0	0	0	0
	Imputation Risk	-1.2%	-1.2%					
	Options Pruned	0/6	0/6	0/12	0/12	2/7	1/12	2/12
	Data Use	100.0%	100.0%	0.0%	0.0%	66.7%	100.0%	91.7%
cardio_eicu $n = 1341, d = 49$ $\mathcal{G} = \{\texttt{sex}, \texttt{age}\}$ $\lvert\mathcal{G}\rvert = 4$ groups Pollard et al. [46]	Overall Performance	17.9%	17.5%	18.0%	18.5%	18.0%	13.1%	**13.0%**
	Overall Gain	0.9%	1.2%	0.8%	0.3%	0.8%	5.7%	**5.8%**
	Group Gains	-0.4% – 3.2%	-0.7% – 2.9%	-0.4% – 3.2%	-0.7% – 2.9%	0.0% – 3.2%	1.9% – 8.4%	2.6% – 8.4%
	Max Disparity	3.5%	3.6%	3.5%	3.6%	3.2%	6.4%	5.8%
	Rat. Violations	0	0	0	0	0	0	0
	Imputation Risk	-2.3%	-1.0%					
	Options Pruned	0/4	0/4	0/9	0/9	3/5	2/9	3/9
	Data Use	100.0%	100.0%	0.0%	0.0%	25.0%	100.0%	87.5%
cardio_mimic $n = 5289, d = 49$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathcal{G}\rvert = 4$ groups [33]	Overall Performance	21.3%	20.9%	21.3%	21.2%	20.2%	**16.8%**	**16.8%**
	Overall Gain	-1.2%	-0.7%	-1.1%	-1.0%	0.0%	**3.4%**	**3.4%**
	Group Gains	-1.9% – -0.6%	-1.1% – -0.3%	-1.9% – -0.6%	-1.1% – -0.3%	0.0% – 0.0%	0.5% – 5.0%	0.5% – 5.0%
	Max Disparity	1.3%	0.8%	1.3%	0.8%	0.0%	4.5%	4.5%
	Rat. Violations	4	4	4	4	0	0	0
	Imputation Risk	-1.9%	-1.1%					
	Options Pruned	0/4	0/4	0/9	0/9	4/5	1/9	2/9
	Data Use	100.0%	100.0%	0.0%	0.0%	0.0%	100.0%	87.5%
coloncancer $n = 29211, d = 72$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathcal{G}\rvert = 6$ groups NCI [41]	Overall Performance	37.2%	37.0%	37.2%	37.0%	37.0%	**35.9%**	35.9%
	Overall Gain	-0.2%	0.0%	-0.2%	-0.0%	0.0%	**1.0%**	1.0%
	Group Gains	-0.7% – 0.1%	-0.3% – 0.2%	-0.7% – 0.1%	-0.3% – 0.2%	0.0% – 0.0%	0.1% – 3.2%	0.1% – 3.2%
	Max Disparity	0.7%	0.5%	0.7%	0.5%	0.0%	3.1%	3.1%
	Rat. Violations	4	1	4	1	0	0	0
	Imputation Risk	-0.7%	-0.3%					
	Options Pruned	0/6	0/6	0/12	0/12	6/7	3/12	5/12
	Data Use	100.0%	100.0%	0.0%	0.0%	0.0%	100.0%	75.0%
lungcancer $n = 120641, d = 84$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathcal{G}\rvert = 6$ groups NCI [41]	Overall Performance	20.0%	20.2%	20.0%	20.3%	20.0%	**19.3%**	19.3%
	Overall Gain	0.1%	-0.1%	0.1%	-0.2%	0.1%	**0.8%**	0.7%
	Group Gains	-0.3% – 0.2%	-0.5% – 0.0%	-0.3% – 0.2%	-0.5% – 0.0%	0.0% – 0.2%	0.0% – 2.3%	0.0% – 2.2%
	Max Disparity	0.6%	0.5%	0.6%	0.5%	0.2%	2.3%	2.1%
	Rat. Violations	1	4	1	4	0	0	0
	Imputation Risk	-0.3%	-0.5%					
	Options Pruned	0/6	0/6	0/12	0/12	3/7	1/12	3/12
	Data Use	100.0%	100.0%	0.0%	0.0%	50.0%	100.0%	83.3%
saps $n = 7797, d = 36$ $\mathcal{G} = \{\texttt{HIV}, \texttt{age}\}$ $\lvert\mathcal{G}\rvert = 4$ groups Allyn et al. [4]	Overall Performance	14.1%	15.0%	14.1%	15.7%	13.9%	**9.8%**	**9.8%**
	Overall Gain	0.9%	-0.0%	0.9%	-0.7%	1.1%	**5.2%**	**5.2%**
	Group Gains	-0.8% – 3.4%	-0.5% – 0.3%	-0.8% – 3.4%	-0.5% – 0.3%	0.0% – 3.4%	0.0% – 16.4%	0.0% – 16.4%
	Max Disparity	4.2%	0.8%	4.2%	0.8%	3.4%	16.4%	16.4%
	Rat. Violations	1	1	1	1	0	0	0
	Imputation Risk	-0.8%	-0.7%					
	Options Pruned	0/4	0/4	0/9	0/9	2/5	1/9	1/9
	Data Use	100.0%	100.0%	0.0%	0.0%	50.0%	75.0%	87.5%

Table 5: Overview of performance, data use, and consent for all personalized models and systems on all datasets as measured by **test error**. We show the performance of models and systems built using **random forests**.

D.3. Random Forests for Ranking (AUC)

Dataset	Metrics	Static		Imputed		Participatory		
		1Hot	mHot	KNN-1Hot	KNN-mHot	Minimal	Flat	Seq
apnea $n = 1152, d = 26$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathcal{G}\rvert = 6$ groups Ustun et al. [54]	Overall Performance	0.825	0.824	0.822	0.806	0.823	**0.944**	0.942
	Overall Gain	0.008	0.006	0.004	-0.012	0.005	**0.126**	0.124
	Group Gains	-0.004 – 0.009	-0.005 – 0.012	-0.004 – 0.009	-0.005 – 0.012	0.000 – 0.009	0.058 – 0.157	0.058 – 0.157
	Max Disparity	0.012	0.017	0.012	0.017	0.009	0.098	0.098
	Rat. Violations	2	3	2	3	0	0	0
	Imputation Risk	-0.004	-0.005					
	Options Pruned	0/6	0/6	0/12	0/12	3/7	2/12	4/12
	Data Use	100.0%	100.0%	0.0%	0.0%	50.0%	100.0%	75.0%
cardio_eicu $n = 1341, d = 49$ $\mathcal{G} = \{\texttt{sex}, \texttt{age}\}$ $\lvert\mathcal{G}\rvert = 4$ groups Pollard et al. [46]	Overall Performance	0.893	0.893	0.894	0.888	0.893	0.949	**0.949**
	Overall Gain	0.003	0.002	0.004	-0.003	0.003	0.058	**0.059**
	Group Gains	-0.006 – 0.012	-0.008 – 0.010	-0.006 – 0.012	-0.008 – 0.010	0.000 – 0.012	0.017 – 0.070	0.017 – 0.072
	Max Disparity	0.018	0.018	0.018	0.018	0.012	0.053	0.055
	Rat. Violations	2	2	2	2	0	0	0
	Imputation Risk	-0.006	-0.008					
	Options Pruned	0/4	0/4	0/9	0/9	3/5	2/9	3/9
	Data Use	100.0%	100.0%	0.0%	0.0%	25.0%	100.0%	87.5%
cardio_mimic $n = 5289, d = 49$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathcal{G}\rvert = 4$ groups [33]	Overall Performance	0.880	0.881	0.880	0.879	0.880	**0.920**	**0.920**
	Overall Gain	-0.000	0.001	0.000	-0.001	0.000	**0.039**	**0.039**
	Group Gains	-0.002 – 0.001	-0.000 – 0.002	-0.002 – 0.001	-0.000 – 0.002	0.000 – 0.000	0.018 – 0.048	0.018 – 0.048
	Max Disparity	0.003	0.002	0.003	0.002	0.000	0.030	0.030
	Rat. Violations	1	0	1	0	0	0	0
	Imputation Risk	-0.002	-0.000					
	Options Pruned	0/4	0/4	0/9	0/9	4/5	0/9	2/9
	Data Use	100.0%	100.0%	0.0%	0.0%	0.0%	100.0%	75.0%
coloncancer $n = 29211, d = 72$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathcal{G}\rvert = 6$ groups NCI [41]	Overall Performance	0.684	0.682	0.681	0.680	0.683	**0.696**	0.696
	Overall Gain	0.002	0.000	-0.001	-0.002	0.001	**0.014**	0.014
	Group Gains	-0.002 – 0.004	-0.004 – 0.002	-0.002 – 0.004	-0.004 – 0.002	0.000 – 0.004	0.004 – 0.035	0.004 – 0.031
	Max Disparity	0.006	0.007	0.006	0.007	0.004	0.030	0.026
	Rat. Violations	0	0	0	0	0	0	0
	Imputation Risk	-0.002	-0.004					
	Options Pruned	0/6	0/6	0/12	0/12	3/7	2/12	5/12
	Data Use	100.0%	100.0%	0.0%	0.0%	50.0%	100.0%	75.0%
lungcancer $n = 120641, d = 84$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathcal{G}\rvert = 6$ groups NCI [41]	Overall Performance	0.849	0.849	0.848	0.849	0.848	**0.856**	**0.856**
	Overall Gain	0.002	0.001	0.001	0.001	0.000	**0.008**	**0.008**
	Group Gains	-0.001 – 0.003	-0.001 – 0.002	-0.001 – 0.003	-0.001 – 0.002	0.000 – 0.003	0.002 – 0.020	0.002 – 0.020
	Max Disparity	0.004	0.003	0.004	0.003	0.003	0.018	0.018
	Rat. Violations	1	1	1	1	0	0	0
	Imputation Risk	-0.001	-0.001					
	Options Pruned	0/6	0/6	0/12	0/12	2/7	1/12	2/12
	Data Use	100.0%	100.0%	0.0%	0.0%	66.7%	100.0%	91.7%
saps $n = 7797, d = 36$ $\mathcal{G} = \{\texttt{HIV}, \texttt{age}\}$ $\lvert\mathcal{G}\rvert = 4$ groups Allyn et al. [4]	Overall Performance	0.921	0.922	0.922	0.906	0.921	**0.966**	**0.966**
	Overall Gain	0.003	0.004	0.003	-0.012	0.002	**0.048**	**0.048**
	Group Gains	-0.002 – 0.010	-0.002 – 0.013	-0.002 – 0.010	-0.002 – 0.013	0.000 – 0.010	0.009 – 0.109	0.009 – 0.109
	Max Disparity	0.012	0.015	0.012	0.015	0.010	0.100	0.100
	Rat. Violations	2	2	2	2	0	0	0
	Imputation Risk	-0.002	-0.002					
	Options Pruned	0/4	0/4	0/9	0/9	2/5	2/9	2/9
	Data Use	100.0%	100.0%	0.0%	0.0%	50.0%	100.0%	87.5%

Table 6: Overview of performance, data use, and consent for all personalized models and systems on all datasets as measured by **test auc**. We show the performance of models and systems built using **random forests**.