Participatory Personalization in Classification

Hailey Joren
UC San Diego

Chirag Nagpal
Google Research

Katherine Heller
Google Research

Berk Ustun
UC San Diego

Abstract

Machine learning models are often personalized with 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 personal information to a model, nor tell if they benefit from personalization in the first place. We introduce a family of classification models, called *participatory systems*, that let individuals opt into personalization at prediction time. We present a model-agnostic algorithm to learn participatory systems for personalization with categorical group attributes. We conduct a comprehensive empirical study of participatory systems in clinical prediction tasks, benchmarking them with common approaches for personalization and imputation. Our results demonstrate that participatory systems can facilitate and inform consent while improving performance and data use across all groups who report personal data.

1 Introduction

Machine learning models routinely assign predictions to *people* – be it to screen a patient for a mental illness [35], their risk of mortality in an ICU [44], or their likelihood of responding to treatment [1]. Many models in such applications are designed to target heterogeneous subpopulations using features that explicitly encode personal information. Typically, models are *personalized* with categorical attributes that define groups [i.e., "categorization" as per 27]. In medicine, for example, clinical prediction models use *group attributes* that are *protected* (e.g., `sex` in the ASCVD Score for cardiovascular disease), *sensitive* (e.g., `HIV_status` in the VA COVID-19 Mortality Score), *self-reported* (e.g., `alcohol_use` in the HAS-BLED Score for Major Bleeding Risk), or *costly* to acquire (e.g., `leukocytosis` in the Alvarado Appendicitis Score).

Individuals expect the right to opt out of providing personal data and the ability to understand how it will be used [see, e.g., personal data guidelines in GDPR, OECD privacy guidelines 26, 40]. In many contexts, 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. At the same time, practitioners assume that data available for training will be available at inference time. In practice, this assumption has led to a proliferation of models that use information that individuals may be unwilling or unable to report at prediction time [see e.g., the Denver HIV Risk Score 29, which asks patients to report `age`, `gender`, `race`, and `sexual_practices`]. In tasks where individuals self-report, they may not voluntarily report information that could improve their predictions or may report incorrect information.

The broader need to facilitate and inform consent in personalized prediction tasks stems from the fact that personalization may not improve performance for each group that reports personal data [51]. In practice, a personalized model can perform *worse* or the same as a *generic model* fit 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 42, 51]. However, they would be resolved if individuals could opt out of personalization and understand its expected gains (see Fig. 1).

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

https://doi.org/10.52202/075280-0622

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	$R_{\boldsymbol{g}}(h_0)$		$\boldsymbol{r}$	$\Delta R_{\boldsymbol{g}}(h, h_0)$		$\boldsymbol{r}$	$\Delta R_{\boldsymbol{g}}(h, h_0)$
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	male, old	25
male, young	0	27	−	0	−	0	h	male, young	0	h_0	∅	0
Total	50	51		24		50			26			50

Figure 1: Classification task where participation improves accuracy and minimizes data use. We consider a dataset that has no features, two group attributes $\mathcal{G} = \texttt{sex} \times \texttt{age}$, $n^- = 51$ negative examples and $n^+ = 50$ positive examples. Here, the best personalized linear model $h : \mathcal{X} \times \mathcal{G} \to \mathcal{Y}$ with a one-hot encoding of $\mathcal{G}$ makes 24 mistakes, and the best generic model $h_0 : \mathcal{X} \times \mathcal{Y}$ makes 50 mistakes as it predicts the majority class $(-)$. Under traditional personalization, individuals report group membership to receive personalized predictions from h. As shown, personalization benefits the population as a whole by reducing overall error from 50 to 24 ($\Delta R_{\boldsymbol{g}}(h, h_0) = 26$). However, personalization has a detrimental effect on $[\texttt{female}, \texttt{old}]$, who receive less accurate predictions from the personalized model ($\Delta R_{\boldsymbol{g}}(h, h_0) = -24$), and no effect on $[\texttt{male}, \texttt{young}]$ who receive the same predictions from the personalized and generic models($\Delta R_{\boldsymbol{g}}(h, h_0) = 0$). In a minimal participatory system, individuals *opt in* to personalization, choosing to receive predictions from h or h_0. Here, individuals in groups $[\texttt{female}, \texttt{old}]$ and $[\texttt{male}, \texttt{young}]$ opt out of personalization, leading to an overall error of 0 ($\Delta R_{\boldsymbol{g}}(h, h_0) = 50$) and a reduction in unnecessary data collection ($\varnothing$).

This work introduces a family of classification models that operationalize informed consent called *participatory systems*. Participatory systems *facilitate consent* by allowing individuals to report personal information at prediction time. Moreover, they *inform consent* by showing how reporting personal information will change their predictions. Models that facilitate consent operate as markets in which individuals trade personal information for performance gains. This work seeks to develop systems that perform as well as possible both when individuals opt-in – to incentivize voluntary reporting – and when they opt out – to safeguard against abstention. Our main contributions include:

1. We present a variety of participatory systems that provide opportunities for individuals to make informed decisions about data provision. Each system ensures that individuals who opt into personalization will receive the most accurate possible predictions possible.
2. We develop a model-agnostic algorithm to learn participatory systems. Our approach can produce a variety of systems that promote participation and minimize data use in deployment.
3. We conduct a comprehensive study of participatory systems in real-world clinical prediction tasks. The results show how our approach can facilitate and inform consent in a way that improves performance and minimizes data use.
4. We provide a Python library to build and evaluate participatory systems.

Related Work Participatory systems support modern principles of responsible data use articulated in OECD privacy guidelines [40], the GDPR [26], and the California Consumer Privacy Act [16]. These include: *informed consent*, i.e., that data should be collected with the data subject's consent; and *collection limitation*, i.e., that data collected should be restricted to only what is necessary. These principles stem from extensive work on the right to data privacy [33]. They are motivated, in part, by research showing that individuals care deeply about their ability to control personal data [4, 8, 10] but differ considerably in their desire or capacity to share it [see e.g. 5, 7, 9, 17, 18, 39, 41]. Our proposed systems let decision subjects report personal data in exchange for performance, which is aligned with principles articulated in recent work on data privacy [13, 46] and related to work in designing incentive-compatible prediction functions [24].

We consider models that are personalized with categorical attributes that encode personal characteristics [i.e., "categorization" rather than "individualization" as per 27]. Modern techniques for learning with categorical attributes [see e.g., 2, 48] use them to improve performance at a population level – e.g., by accounting for higher-order interaction effects [14, 38, 58] or recursive partitioning [11, 12, 15, 25]. Our methods can be used to achieve these goals in tasks where models use features that are optional or costly to acquire [see e.g., 6, 7, 52, 61].

Our work is related to algorithmic fairness in that we seek to improve model performance at a group level. Recent work shows that personalization with group attributes does not uniformly improve performance and can reduce accuracy at a group level [see 42, 51, 57]. Our systems can safeguard against such instances of "worsenalization" by informing users of the gains in reporting and allowing

https://doi.org/10.52202/075280-0622

them to opt out of reporting. This line of broad work complements research on preference-based fairness [22, 36, 57, 59, 62], on ensuring fairness across complex group structures [28, 30, 34], and promoting privacy across subpopulations [13, 53].

2 Participatory Systems

We consider a classification 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 m categorical attributes $\boldsymbol{g}_i = [g_{i,1}, \ldots, g_{i,m}] \in \mathcal{G}_1 \times \ldots \times \mathcal{G}_m = \mathcal{G}$. We refer to $\mathcal{G}$ as *group attributes*, and to $\boldsymbol{g}_i$ as the *group membership* of person i. We use $n_{\boldsymbol{g}} := |\{i \,|\boldsymbol{g}_i = \boldsymbol{g}\}|$ denote the size of group $\boldsymbol{g}$, and use $|\mathcal{G}_k|$ to denote the number of categories for group attribute k.

We use the dataset to train a personalized model $h : \mathcal{X} \times \mathcal{G} \to \mathcal{Y}$ via empirical risk minimization with a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. Given a model h, we denote its empirical risk and true risk as $\hat{R}(h)$ and $R(h)$, respectively, and evaluate model performance at the group level. We denote the empirical risk and true risk of a model h on group $\boldsymbol{g} \in \mathcal{G}$ as

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

We consider tasks where every individual prefers more accurate predictions.

Assumption 1. Given models h and h', individuals in group $\boldsymbol{g}$ prefer h to h' when $R_{\boldsymbol{g}}(h) < R_{\boldsymbol{g}}(h')$.

Assumption 1 holds in settings where every individual prefers more accurate predictions – e.g., clinical prediction tasks such as screening or diagnosing illnesses [49, 56]. It does not hold in applications where some individuals prefer predictions that may be inaccurate – e.g., such as predicting the risk of organ failure for a transplant [see e.g., 43, for other "polar" clinical applications].

Operationalizing Consent We consider models where individuals consent to personalization by deciding whether or not to report their group attributes at prediction time. We let $\varnothing$ denote an attribute that was not reported, and let $\boldsymbol{r}_i = [r_{i,1}, \ldots, r_{i,k}] \in \mathcal{R} \subseteq \mathcal{G} \times \boldsymbol{\varnothing}$. For example, a person with $\boldsymbol{g}_i = [\texttt{female}, \texttt{HIV} = +]$ would report $\boldsymbol{r}_i = [\texttt{female}, \varnothing]$ if they only disclose `sex`, and would report $\boldsymbol{r}_i = \boldsymbol{\varnothing} := [\varnothing, \varnothing]$ if they opt out of reporting entirely.

We associate each model with a set of *reporting options* $\mathcal{R}$. A traditional model, which requires each person to report group attributes, has $\mathcal{R} = \mathcal{G}$. A model where each person could report any subset of group attributes has $\mathcal{R} = \mathcal{G} \times \boldsymbol{\varnothing}$. We represent individual decisions to opt into personalization at prediction time through a *reporting interface* defined below.

Definition 1. Given a personalized classification task with group attributes $\mathcal{G}$, a *reporting interface* is a tree T whose nodes represent attributes reported at prediction time. The tree is rooted at $\text{root}(T) = [\varnothing, \ldots, \varnothing]$ and branches as a person reports personal attributes. Given a node $\boldsymbol{r}$, we denote its parent as $\text{pa}(\boldsymbol{r})$. Each parent-child pair represents a *reporting decision*, and the height of the tree represents the maximum number of reporting decisions.

Definition 2. Given a personalized classification task with group attributes $\mathcal{G}$, a *participatory system* with reporting interface T is a prediction model $f_T : \mathcal{X} \times \mathcal{R} \to \mathcal{Y}$ that obeys the following properties:

(P1) *Baseline Performance*: Opting out of personalization entirely guarantees the expected performance from a *generic model* trained without group attributes $h_0 \in \text{argmin}_{h \in \mathcal{H}} R(h)$.

$$R_{\boldsymbol{r}}(f_T(\cdot, \boldsymbol{\varnothing})) = R_{\boldsymbol{r}}(h_0) \text{ for all reporting groups } \boldsymbol{r} \in \mathcal{R}.$$

(P2) *Incentive Compatibility*: Opting into personalization improves expected performance

$$R_{\boldsymbol{r}}(f_T(\cdot, \boldsymbol{r})) < R_{\boldsymbol{r}}(f_T(\cdot, \boldsymbol{r}')) \text{ for all nested reporting groups } \boldsymbol{r}, \boldsymbol{r}' \in \mathcal{G} \times \boldsymbol{\varnothing} \text{ such that } \boldsymbol{r}' = \text{pa}(\boldsymbol{r}).$$

Here, the *Baseline Performance* property ensures that individuals who choose not to share personal information receive the performance of a generic model – i.e., the most accurate model that could be trained without this information. This property also ensures individuals retain the ability to opt out

https://doi.org/10.52202/075280-0622

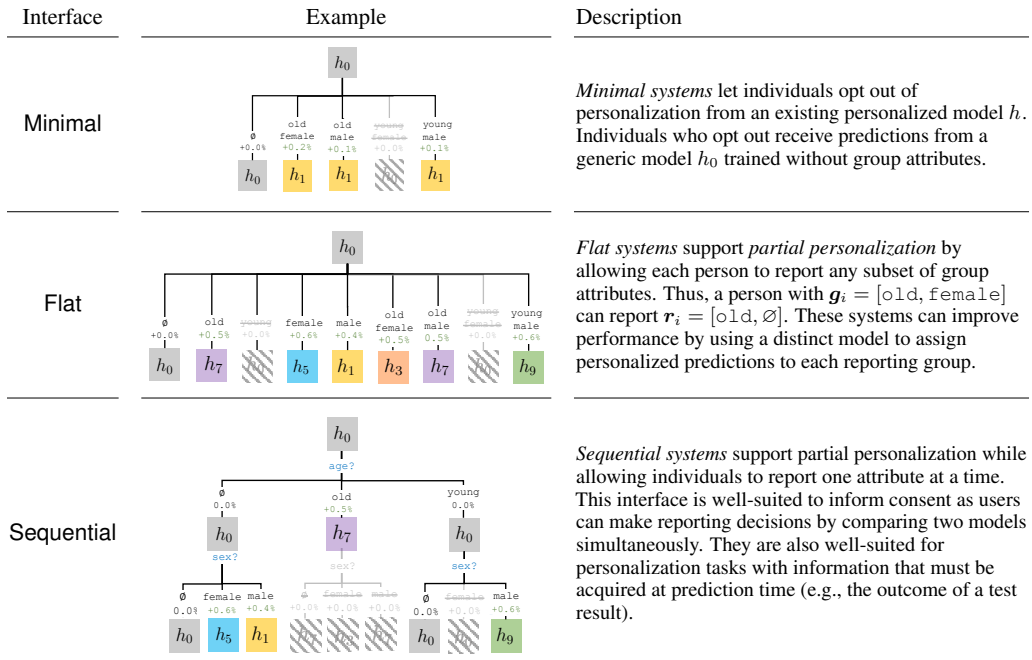

Figure 2: Participatory systems for a personalized classification task with group attributes $\texttt{sex} \times \texttt{age} = [\texttt{male}, \texttt{female}] \times [\texttt{old}, \texttt{young}]$. Each system allows a person to opt out of personalization by reporting $\varnothing$ and informs their choice by showing the expected gains of personalization (e.g., $\texttt{+0.2\%}$ gain in accuracy). Systems minimize data use by removing reporting options that do not improve accuracy (see grey-striped boxes). Here, $[\texttt{young}, \texttt{female}]$ is pruned in all systems as it leads to a gain $\leq 0.0\%$.

of personalization – i.e., $\boldsymbol{\varnothing} \in \mathcal{R}$. The *Incentive Compatibility* property ensures that personalization will improve expected performance – i.e., when individuals report personal data, the system can effectively leverage that data to deliver more accurate predictions in expectation. Together, these properties lead to data minimization, as systems that obey these properties will not request data from a reporting group when it will not lead to an improvement in expected performance.

On Data Minimization via Imputation An alternative approach to allow individuals to opt out of reporting personal information at prediction time is to impute their group membership. Imputation allows individuals to opt out of personalization but does not guarantee the accuracy of their predictions. As a result, individuals who opt out of personalization by reporting $\boldsymbol{r} = \boldsymbol{\varnothing}$ may receive a less accurate prediction than they would receive from a generic model. In the best-case scenario where we could perfectly impute group membership, a group might be assigned better predictions from a generic model (see Fig. 1). In the worst case, imputation may be incorrect, leading to even more inaccurate predictions than those of the generic or personalized model. We highlight these effects on real-world datasets in our experiments in Section 4.

Characterizing System Performance One of the key differences between traditional models and participatory systems is that their performance depends on individual reporting decisions. In what follows, we characterize the performance under a general model of individual disclosure. Given a participatory system f_T, we assume that each individual reports personal information to maximize an individual utility function of the form:

$$u_i(\boldsymbol{r}; f_T) = b_i(\boldsymbol{r}; f_T) - c_i(\boldsymbol{r}) \tag{1}$$

Here, $c_i(\cdot)$ and $b_i(\cdot)$ denote the cost and benefit that individual i receives from reporting $\boldsymbol{r}$ to f_T respectively. We assume that individuals incur no cost when they do not report any attributes such that $c_i(\boldsymbol{\varnothing}) = 0$, and incur costs that increase monotonically with information disclosed such that $c_i(\boldsymbol{r}) \leq c_i(\boldsymbol{r}')$ for $\boldsymbol{r} \subseteq \boldsymbol{r}'$. We assume that benefits increase monotonically with expected gains in true risk so that $R_{\boldsymbol{r}}(f_T(\cdot, \boldsymbol{r})) < R_{\boldsymbol{r}}(f_T(\cdot, \boldsymbol{r}')) \implies b_i(\boldsymbol{r}, f_T) > b_i(\boldsymbol{r}', f_T)$.

In Fig. 3, we show how the system performance for each reporting group can change with respect to participation when we simulate individual disclosure decisions from a model that satisfies the assumptions listed above. When a personalized model h requires individuals to report information that reduces performance as in Fig. 1, individuals incur a cost of disclosure without receiving a benefit in return. In such cases, individuals who interact with a minimal system would opt out of worsenalization and receive more accurate predictions from a generic model, thereby improving the overall performance of the system.

We observe that the maximum utility that each individual can receive from a participatory system can only increase as we add more reporting options. Thus, flat and sequential systems should exhibit better performance than a minimal system.

Given a participatory system f_T with reporting options $\mathcal{R}$, a participatory system $f_{T'}$ with more reporting options $\mathcal{R}' \supseteq \mathcal{R}$ can only improve performance, – i.e., $R(f_{T'}) \leq R(f_T)$. Similarly, the system with more reporting options can only improve utility, – i.e., $u_i(\boldsymbol{r}; f_{T'}) \geq u_i(\boldsymbol{r}; f_T)$ for all individuals i.

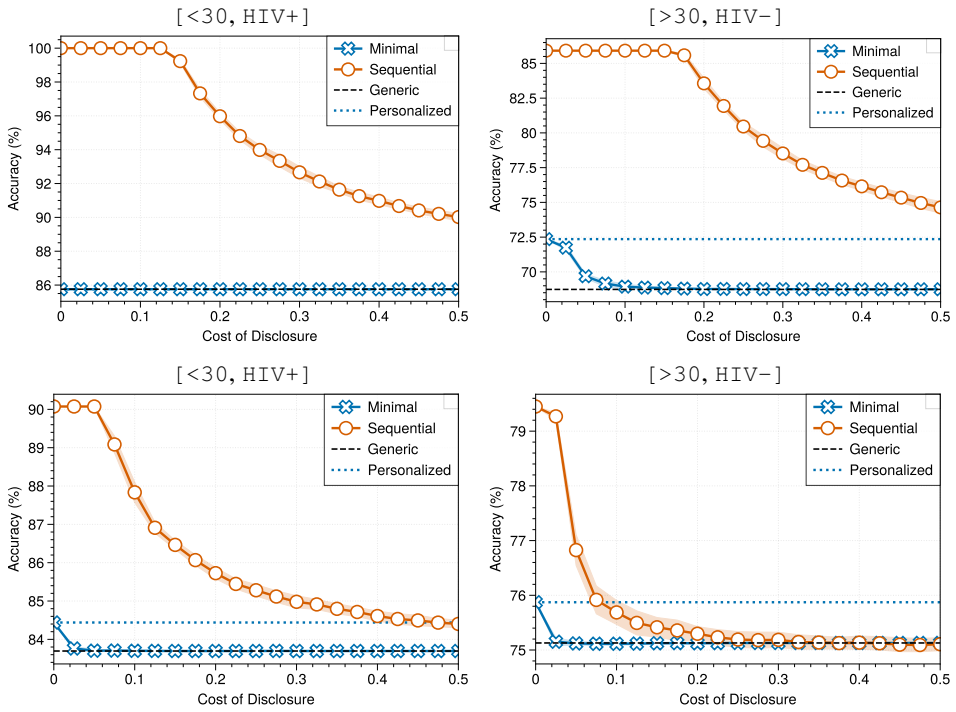

Figure 3: Performance profile of participatory systems for the `saps` dataset for each intersectional group in the `saps` dataset. We plot out-of-sample performance for different levels of participation in the target population. We control participation by varying the reporting cost in a simulated model of individual disclosure. As shown, minimal and sequential systems outperform a generic model at a group level regardless of participation. In regimes where the cost of disclosure is low, participation is high. Consequently, a minimal system will achieve the same performance as a personalized model, and a sequential system will achieve the performance of the component model for each supgroup. We provide details and results in Appendix D.

3 Learning Participatory Systems

This section describes a model-agnostic algorithm to learn participatory systems that ensures incentive compatibility and baseline performance in deployment. We outline our procedure in Algorithm 1 to learn the three kinds of participatory systems in Fig. 2. The procedure takes as input a pool of candidate models $\mathcal{M}$, a dataset for model assignment $\mathcal{D}^{\text{assign}}$, and a dataset for pruning $\mathcal{D}^{\text{prune}}$. It outputs a collection of participatory systems that obey the properties described in Definition 2 on test data. The procedure combines three routines to (1) generate viable reporting interfaces (Line 1); (2) assign models over the interface (Line 3); (3) prune the system to limit unnecessary data collection (Line 4). We present complete procedures for each routine in Appendix A and discuss them below.

https://doi.org/10.52202/075280-0622

Algorithm 1 Learning Participatory Systems

Input: $\mathcal{M} : \{h : \mathcal{X} \times \mathcal{G} \to \mathcal{Y}\}$ *pool of candidate models*
Input: $\mathcal{D}^{\text{assign}} = \{(\boldsymbol{x}_i, \boldsymbol{g}_i, y_i)\}_{i=1}^{n^{\text{assign}}}$ *assignment dataset*
Input: $\mathcal{D}^{\text{prune}} = \{(\boldsymbol{x}_i, \boldsymbol{g}_i, y_i)\}_{i=1}^{n^{\text{prune}}}$ *pruning dataset*
1: $\mathbb{T} \leftarrow \mathsf{ViableTrees}(\mathcal{G}, \mathcal{D}^{\text{assign}})$ $|\mathbb{T}| = 1$ *for minimal & flat systems*
2: **for** $T \in \mathbb{T}$ **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 $\mathbb{T}$, collection of participatory systems

Model Pool Our procedure takes as input a *pool of candidate models* $\mathcal{M}$ to assign over a reporting interface. At a minimum, every pool 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. A single personalized model can perform unreliably across reporting groups due to differences in the data distribution or trade-offs between groups. Using a pool of models safeguards against these effects by drawing on models from different model classes that have been personalized using different techniques for each reporting group. By default, we include models trained specifically on the data for each reporting group, as such models can perform well on heterogeneous subgroups [51, 57].

Enumerating Interfaces We call the ViableTrees routine in Line 1 to enumerate *viable* reporting interfaces. We only call this routine for sequential systems since minimal and flat systems use a single reporting interface that is known a priori. ViableTrees takes as input a group attributes $\mathcal{G}$ and a dataset $\mathcal{D}^{\text{assign}}$. It returns all m-ary trees that obey constraints on sample size and reporting (e.g., users who report `male` should report `age` before `HIV`). By default, we only generate trees so that we have sufficient data to estimate gains at each node of the reporting interface[1]. In general, ViableTrees scales to tasks with ≤ 8 group attributes. Beyond this limit, one can reduce the enumeration size by specifying ordering constraints or a threshold number of trees to enumerate before stopping. For a task with three binary group attributes, $\mathbb{T}$ contains 24 3-ary trees of depth 3. Given a complete ordering of all 3 group attributes, however, $\mathbb{T}$ would have 1 tree. We can also consider a greedy algorithm (see Appendix A.4), which may be practical for large-scale problems.

Model Assignment We assign each reporting group a model using the AssignModels routine in Line 3. Given a reporting group $\boldsymbol{r}$, we consider all models that could use any subset of group attributes in $\boldsymbol{r}$. 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 model at each node performs as well as the generic model on out-of-sample data (i.e., *baseline performance* in Definition 2).

Pruning Reporting Options AssignModels may output trees that violate incentive compatibility by requesting personal information that fails to improve performance. This can happen when the routine assigns a model that performs equally well to nested reporting groups – see, e.g., Fig. 2 where the Flat system assigns h_0 to $[\texttt{female}, \varnothing]$ and $[\texttt{female}, \texttt{young}]$.

We can avoid requesting data from reporting groups in such cases by calling the Prune routine in Line 4. This routine takes as input a participatory system f_T and a pruning dataset $\mathcal{D}^{\text{prune}}$ and outputs a system $f_{T'}$ with a pruned interface $T' \subseteq T$. The routine uses a bottom-up pruning procedure that calls a one-sided hypothesis test at each node:

$$H_0 : \Delta_{\boldsymbol{r}}(\boldsymbol{r}, \text{pa}(\boldsymbol{r})) \leq 0 \qquad H_A : \Delta_{\boldsymbol{r}}(\boldsymbol{r}, \text{pa}(\boldsymbol{r})) > 0$$

The test checks if each reporting group $\boldsymbol{r}$ receives more accurate predictions from the personalized model assigned to its current node or $\boldsymbol{r}$ its parent $\text{pa}(\boldsymbol{r})$. Here, H_0 assumes a reporting group prefers the parent model. Thus, we reject H_0 when we can reliably tell that $f_T(\cdot, \boldsymbol{r})$ performs better for $\boldsymbol{r}$ on the pruning dataset. The exact test should chosen based on the performance metric for the underlying

[1]For example, trees whose leaves contain at least one positive sample, one negative sample, and $n_{\boldsymbol{r}} \geq d + 1$ samples to avoid overfitting

https://doi.org/10.52202/075280-0622

prediction task. In general, we can use a bootstrap hypothesis test [20] and draw on more powerful tests for salient performance metrics [e.g., 19, 21, 50, for accuracy and AUC].

On Computation Our approach provides several options to moderate the computation cost of training a pool of models. For example, we can train only two models and build a minimal system. Alternatively, we can also build a flat or sequential system using a limited number of models in the pool. In practice, the primary bottleneck when 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 attribute, we quickly enter a regime where we cannot reliably evaluate model performance for assignment and pruning [see 42].

On Customization 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 gains we would show users (e.g., error for a diagnosis, AUC for triage, ECE for risk assessment). 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 based on criteria not known at training time. For example, one can swap the trees to use a system that always requests HIV status last. By default, we select the configuration that minimizes data collection across groups, such that the ordering of attributes results leads to the most significant number of data requests pruned.

4 Experiments

We benchmark participatory systems on real-world clinical prediction tasks. Our goal is to evaluate these approaches in terms of performance, data usage, and consent in applications where individuals have a low reporting cost. We include code to reproduce these results in an Python library.

4.1 Setup

We consider six classification tasks for clinical decision support where we personalize a model with group attributes that are protected or sensitive (see Table 2 and Appendix B). Each task pertains to an application where we expect individuals to have a low cost of reporting and to report personal information when there is any expected gain. This is because the information used for personalization is readily available, relevant to the prediction task, and likely to be disclosed given legal protections related to the confidentiality of health data [4, 10, 54]. One exception is `cardio_eicu` and `cardio_mimic`, which are personalized based on race and ethnicity. [2] We split each dataset into a test sample (20% for evaluating out-of-sample performance) and a training sample (80% for training, pruning, assignment, 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 static models where we impute the group membership for each person (KNN-1Hot, KNN-mHot). In practice, personalized systems with imputation will range between the performance for these systems and the performance of 1Hot and mHot.
- *Participatory*: These are participatory systems built using our approach. These include Minimal, a minimal system built from 1Hot and its generic counterpart; and Flat and Seq, flat and sequential systems built from 1Hot, mHot and their generic counterparts.

We train all models – personalized models and the components of participatory systems – from the same model class and evaluate them using the metrics in Table 1. We repeat the experiments four times, varying the model class (logistic regression, random forests) and the target performance metric (error rate for decision-making tasks, AUC for ranking tasks) to evaluate the sensitivity of our findings with respect to model classes and use cases.

[2] The use of race in clinical risk scores should be approached with caution [60]; participatory systems offer one way to safeguard against inappropriate use.

https://doi.org/10.52202/075280-0622

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	$\sum_{\boldsymbol{g} \in \mathcal{G}} \frac{n_{\boldsymbol{g}}}{n} \hat{\Delta}_{\boldsymbol{g}}(\boldsymbol{g}, \boldsymbol{\varnothing})$	Population-level gain in performance of a personalized system/model over its generic counterpart
Group Gains	$\min_{\boldsymbol{g} \in \mathcal{G}}$ / $\max_{\boldsymbol{g} \in \mathcal{G}} \hat{\Delta}_{\boldsymbol{g}}(\boldsymbol{g}, \boldsymbol{\varnothing})$	Range of group-level gains of a personalized system/model over its generic counterpart across all groups
Rationality Violations	$\sum_{\boldsymbol{g} \in \mathcal{G}} \mathbb{I}[\text{reject } H_0]$	Number of rationality violations detected using a bootstrap test with 100 resamples at a significance of 10% where $H_0 : \Delta_{\boldsymbol{g}}(\boldsymbol{g}, \boldsymbol{\varnothing}) \geq 0$.
Imputation Risk	$\min_{\boldsymbol{g} \in \mathcal{G}} \hat{\Delta}_{\boldsymbol{g}}(\boldsymbol{g}, \boldsymbol{g}')$	Worst-case loss in performance due to incorrect imputation. This metric can only be computed for static models
Options Pruned	$\frac{\lvert\mathcal{R}\rvert - \lvert\mathcal{R}(h)\rvert}{\lvert\mathcal{R}\rvert}$	Proportion of reporting options pruned from a system/model. Here, $\mathcal{R}$ denotes all reporting options and $\mathcal{R}(h)$ denotes those after h is pruned
Data Use	$\sum_{\boldsymbol{g} \in \mathcal{G}} \frac{n_{\boldsymbol{g}}}{n} \frac{\text{requested}(h, \boldsymbol{g})}{\dim(\mathcal{G})}$	Proportion of group attributes requested by h from each group, averaged over all groups in $\mathcal{G}$

Table 1: Metrics used to evaluate performance, data use, and consent of personalized models and systems. We report performance on a held-out test sample. We assume that individuals report group membership to static models, do not report group membership to imputed models, and only report to participatory systems when informed that it would lead to a strictly positive gain, as computed on the validation set in the training sample.

4.2 Discussion

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

On Performance Our results in Table 2 show that participatory systems can improve performance across reporting groups. Here, Flat and Seq achieve the best overall performance on 6/6 datasets and improve the gains from personalization for every reporting group on 5/6 datasets. In contrast, traditional models improve overall performance while reducing performance at a group level (see rationality violations on five datasets for 1Hot, mHot). The performance benefits from participatory systems stem from (i) allowing users to opt out of these instances of "worsenalization" and (ii) assigning personalized predictions with multiple models. Using Table 2, we can measure the impact of (i) by comparing the performance of Minimal vs. 1Hot, and the impact of (ii) by comparing the performance of Minimal to Flat (or Seq). For example, on `apnea`, 1Hot exhibits a significant rationality violation for group $[\texttt{30\_to\_60}, \texttt{male}]$, meaning they would have been better off with a generic model. By comparing the performance of 1Hot to Minimal, we see that allowing users to opt out of worsenalization reduces test error from 29.1% to 28.9%. By comparing the performance on Minimal to Flat and Seq, we see that using multiple models can further reduce test error from 28.9% to 24.1%.

On Informed Consent Our results show how Flat and Seq systems can inform consent by allowing users to report a subset of group attributes (e.g., by including reporting options such as $[\texttt{30+}, \varnothing]$ or $[\varnothing, \texttt{HIV+}]$). Although both Flat and Seq systems allow for partial personalization, their capacity to inform consent differs. In a flat system, users may inaccurately gauge the marginal benefit of reporting an 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 $[\texttt{30+}, \texttt{HIV+}]$ and may mistakenly conclude that the gain of reporting `age` is $16.7\% - 3.7\% = 13.0\%$. This estimate incorrectly presumes that the gains of 3.7% were distributed equally across age groups. Sequential systems directly inform users of the gains for partial reporting. In the sequential system, group $[\texttt{30+}, \texttt{HIV+}]$ is informed that they would see a marginal gain of 21.5% for reporting `age`, while group $[\texttt{<30}, \texttt{HIV+}]$ is informed they would see a marginal gain of reporting `age` of 0.0%.

On Data Minimization Our results show that participatory systems perform better across all groups while requesting less personal data on 6/6 datasets. For example, on `cardio_eicu`, Seq reduces error by 11.3% compared to 1Hot while requesting, on average, 83.3% of the data needed by 1Hot. In general, participatory systems can limit data use where personalization does not improve

https://doi.org/10.52202/075280-0622

		Static		Imputed		Participatory		
Dataset	Metrics	1Hot	mHot	KNN-1Hot	KNN-mHot	Minimal	Flat	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%
$\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$	Worsenalization	1	1	1	1	0	0	0
$\lvert\mathcal{G}\rvert = 6$ groups	Imputation Risk	-4.9%	-5.2%					
Ustun et al. [55]	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	21.4%	21.5%	21.6%	22.1%	21.6%	**10.2%**	**10.2%**
cardio_eicu	Overall Gain	0.4%	0.3%	0.3%	-0.2%	0.3%	**11.7%**	**11.7%**
$n = 1341, d = 49$	Group Gains	-1.3% – 2.6%	-2.7% – 3.0%	-1.3% – 2.6%	-2.7% – 3.0%	0.0% – 2.6%	3.1% – 20.9%	3.1% – 20.9%
$\mathcal{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$	Worsenalization	1	1	1	1	0	0	0
$\lvert\mathcal{G}\rvert = 8$ groups	Imputation Risk	-4.6%	-5.4%					
Pollard et al. [44]	Options Pruned	0/8	0/8	0/27	0/27	6/9	10/27	9/27
	Data Use	100.0%	100.0%	0.0%	0.0%	25.0%	100.0%	83.3%
	Overall Performance	19.4%	19.3%	19.3%	20.1%	19.2%	**15.7%**	**15.7%**
cardio_mimic	Overall Gain	-0.1%	-0.0%	-0.0%	-0.8%	0.1%	**3.5%**	**3.5%**
$n = 5289, d = 49$	Group Gains	-0.9% – 0.4%	-0.9% – 0.5%	-0.9% – 0.4%	-0.9% – 0.5%	0.0% – 0.4%	-1.6% – 9.8%	-1.6% – 9.8%
$\mathcal{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$	Worsenalization	3	2	3	2	0	1	1
$\lvert\mathcal{G}\rvert = 8$ groups	Imputation Risk	-1.1%	-1.1%					
Johnson et al. [32]	Options Pruned	0/8	0/8	0/27	0/27	6/9	6/27	8/27
	Data Use	100.0%	100.0%	0.0%	0.0%	25.0%	100.0%	91.7%
	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%
$\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$	Worsenalization	1	0	1	0	0	0	0
$\lvert\mathcal{G}\rvert = 6$ groups	Imputation Risk	-1.4%	-0.9%					
Scosyrev et al. [45]	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%
$\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$	Worsenalization	4	4	4	4	0	0	0
$\lvert\mathcal{G}\rvert = 6$ groups	Imputation Risk	-0.5%	-0.5%					
Scosyrev et al. [45]	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%
$\mathcal{G} = \{\texttt{HIV}, \texttt{age}\}$	Worsenalization	0	0	0	0	0	0	0
$\lvert\mathcal{G}\rvert = 4$ groups	Imputation Risk	0.0%	-2.4%					
Allyn et al. [3]	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: Participatory systems and personalized models for all datasets. We summarize metrics in Table 1 and present results for other model classes and prediction tasks in Appendix C. The best performance across each system is highlighted in green with bold text, and instances of worsenalization are highlighted in red.

performance, e.g., on `lungcancer`. Even as attributes like `sex` or `age` may be readily reported by patients for any performance benefit, limiting data use is valuable when there is a tangible cost associated with data collection – e.g., when models make use of rating scale for a mental disorder that must be administered by a clinician [47]. The potential for data minimization varies substantially across prediction tasks. On `apnea`, for example, we can prune six reporting options when building a Seq for decision making (which optimizes error) but four options for Seq for ranking (which optimizes AUC; see Appendix C.1). Overall, participatory systems satisfy "global data minimization" as proposed in [13], in that they minimize the amount of per-user data requested while achieving the quality of a system with access to the full data on average.

On the Benefits of a Model-Agnostic Approach Our findings highlight some of the benefits of a model-agnostic approach, in which we can draw on a rich set of models to achieve better performance while mitigating harm. The resulting system can balance training costs with performance benefits. We can also ensure generalization across reporting groups – e.g., by including a generic model fit from a complex model class and personalized models fit from a simpler model class. As expected, fitting for a complex model class can lead to 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 fitting a random forest rather than a logistic regression model (see Appendix C). However, a gain in overall performance does not always translate to gains at the group level. On `saps`, for example, using a random forest also introduces a rationality violation for one group.

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 may violate the requirements in 2. Consider a personalized

https://doi.org/10.52202/075280-0622

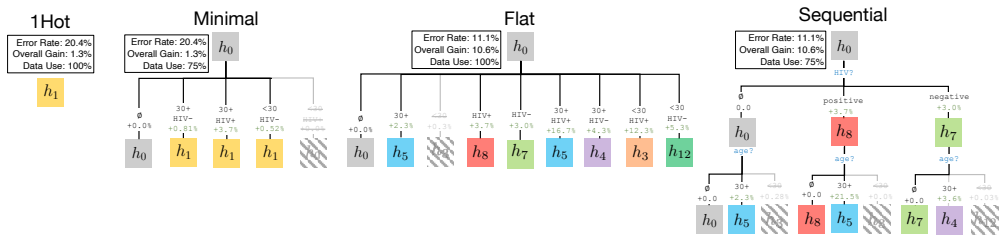

Figure 4: Participatory systems for the `saps` dataset. These models predict ICU mortality for groups defined by $\mathcal{G} =$ `HIV` $\times$ `age` $=$ `[+,-]` $\times$ `[<30, 30+]` using logistic regression component models. Here, h_0 is a generic model, h_1 is a 1Hot model fit with a one-hot encoding of $\mathcal{G}$, and $h_2 \cdots h_m$ are 1Hot and mHot models fit for each reporting group. We show the gains of each reporting option above each box and highlight pruned options in grey. For example, in Seq, the group `(HIV+, 30+)` sees an estimated 21.5% error reduction after reporting `HIV` if they report `age`. In contrast, the group `(HIV+, <30)` sees no gain from reporting `age` in addition to `HIV` status, so this option is pruned.

model that exhibits "worsenalization" in Fig. 1. Even if one could correctly impute the group membership for every person, individuals may receive more accurate predictions from a generic model h_0. In practice, imputation is imperfect – as individuals who opt out of reporting their group membership to a personalized model may be assigned "worse" predictions because they are imputed the group membership of a different group. In such cases, opting out may be beneficial, making it difficult for model developers to promote participation while informing consent. Our results highlight the prevalence of these effects in practice. For example, on `cardio_eicu` the estimated "risk of imputation" is -4.6% , indicating that every intersectional group can experience an increase of 4.6% in the error rate as a result of incorrect imputation. The results for KNN-1Hot show that this potential harm can be realized in practice using KNN-imputation, as we find that the imputed system leads to rationality violations on 5/6 datasets.

5 Concluding Remarks

We introduced a new family of classification models that allow individuals to report personal data at prediction time. Our work focuses on personalization with group attributes; our approach could be used to facilitate and inform consent in a broader class of prediction tasks. In such cases, the key requirement for building a participatory system is that we can reliably estimate the gains of personalization for each person who reports personal data.

Our results show that participatory systems can inform consent while improving performance and reducing data use across groups. Reaping these benefits in practice will hinge on the ability to effectively inform decision subjects on the impact of their reporting decisions. [4]. Even as there may be good "default practices" for what kind of information we should show decision subjects, practitioners should tailor this information to the application and target audience [23].

One common concern in using a participatory system arises when practitioners wish to collect data from a model in deployment to improve its performance in the future. In practice, a participatory system can thwart data collection in such settings by allowing individuals to opt out. In such cases, we would note that this issue should be resolved in a way that is aligned with the principle of *purpose specification* [40]. If the goal of data collection is to improve a model, then individuals could always be asked to report information voluntarily for this purpose. If the goal of data collection is to personalize predictions, then individuals should be able to opt out, especially when it may lead to worse performance.

https://doi.org/10.52202/075280-0622

Acknowledgements

We thank the following individuals for helpful discussions: Taylor Joren, Sanmi Koyejo, Charlie Marx, Julian McAuley, and Nisarg Shah. This work was supported by funding from the National Science Foundation IIS 2040880, the NIH Bridge2AI Center Grant U54HG012510, and an Amazon Research Award.

References

[1] Abajian, Aaron, Nikitha Murali, Lynn Jeanette Savic, Fabian Max Laage-Gaupp, Nariman Nezami, James S Duncan, Todd Schlachter, MingDe Lin, Jean-Francois Geschwind, and Julius Chapiro. Predicting treatment response to intra-arterial therapies for hepatocellular carcinoma with the use of supervised machine learning—an artificial intelligence concept. *Journal of Vascular and Interventional Radiology*, 29 (6):850–857, 2018.

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

[3] Allyn, Jérôme, 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.

[4] Anderson, Catherine L 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.

[5] Arellano, April Moreno, 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.

[6] Atan, Onur, William Whoiles, and Mihaela Schaar. Data-driven online decision making with costly information acquisition. *Arxiv*, 02 2016.

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

[8] Auxier, Brooke, 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.

[9] Awad, Naveen Farag 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.

[10] Bansal, Gaurav, 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.

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

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

[13] Biega, Asia J, Peter Potash, Hal Daumé, Fernando Diaz, and Michèle Finck. Operationalizing the legal principle of data minimization for personalization. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 399–408, 2020.

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

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

[16] 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.

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

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

[19] DeLong, Elizabeth R, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

https://doi.org/10.52202/075280-0622

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

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

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

[23] Edwards, Adrian GK, 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.

[24] Eliaz, Kfir and Ran Spiegler. On incentive-compatible estimators. *Games and Economic Behavior*, 132: 204–220, 2022.

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

[26] European Parliament and of the Council. Regulation 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj. Official Journal of the European Union.

[27] Fan, Haiyan 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.

[28] Globus-Harris, Ira, Michael Kearns, and Aaron Roth. 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] Haukoos, Jason S, Michael S Lyons, Christopher J Lindsell, Emily Hopkins, Brooke Bender, Richard E Rothman, Yu-Hsiang Hsieh, Lynsay A MacLaren, Mark W Thrun, Comilla Sasson, 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.

[30] Hébert-Johnson, Úrsula, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the International Conference on Machine Learning*, pages 1944–1953, 2018.

[31] Hollenberg, SM. Cardiogenic shock. In *Intensive Care Medicine*, pages 447–458. Springer, 2003.

[32] Johnson, Alistair EW, 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.

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

[34] Kearns, Michael, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.

[35] Kessler, Ronald C, Lenard Adler, Minnie Ames, Olga Demler, Steve Faraone, EVA Hiripi, Mary J Howes, Robert Jin, Kristina Secnik, Thomas Spencer, et al. The world health organization adult adhd self-report scale (asrs): a short screening scale for use in the general population. *Psychological medicine*, 35(2): 245–256, 2005.

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

[37] Le Gall, Jean-Roger, 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.

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

[39] Lundberg, Ian, 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.

[40] 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.

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

[42] Paes, Lucas Monteiro, Carol Xuan Long, Berk Ustun, and Flavio Calmon. On the epistemic limits of personalized prediction. In Oh, Alice H., Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Snp3iEj7NJ.

[43] Paulus, Jessica K 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.

[44] Pollard, Tom J, 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.

[45] Scosyrev, Emil, James Messing, Katia Noyes, Peter Veazie, and Edward Messing. Surveillance epidemiology and end results (seer) program and population-based research in urologic oncology: an overview. In *Urologic Oncology: Seminars and Original Investigations*, volume 30, pages 126–132. Elsevier, 2012.

[46] Shanmugam, Divya, Fernando Diaz, Samira Shabanian, Michèle Finck, and Asia Biega. Learning to limit data collection via scaling laws: A computational interpretation for the legal principle of data minimization. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 839–849, 2022.

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

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

[49] Struck, Aaron F, 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.

[50] Sun, Xu and Weichao Xu. 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.

[51] Suriyakumar, Vinith M, Marzyeh Ghassemi, and Berk Ustun. When personalization harms: Reconsidering the use of group attributes in prediction. In *International Conference on Machine Learning*, 2023.

[52] Tran, Cuong and Ferdinando Fioretto. Personalized privacy auditing and optimization at test time. *arXiv preprint arXiv:2302.00077*, 2023.

[53] Tran, Cuong, My Dinh, and Ferdinando Fioretto. Differentially private empirical risk minimization under the fairness lens. *Advances in Neural Information Processing Systems*, 34:27555–27565, 2021.

[54] U.S. Congress. Health insurance portability and accountability act of 1996, 1996. URL https://www.hhs.gov/hipaa/for-professionals/privacy/index.html. Public Law 104-191.

[55] Ustun, Berk, 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.

[56] Ustun, Berk, Lenard A Adler, Cynthia Rudin, Stephen V Faraone, Thomas J Spencer, Patricia Berglund, Michael J Gruber, and Ronald C Kessler. The world health organization adult attention-deficit/hyperactivity disorder self-report screening scale for dsm-5. *Jama psychiatry*, 74(5):520–526, 2017.

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

[58] Vaughan, Gregory, 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.

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

[60] Vyas, Darshali A, Leo G Eisenstein, and David S Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.

[61] Yu, Shipeng, Balaji Krishnapuram, Rómer Rosales, and R. Bharat Rao. Active sensing. In *AISTATS*, 2009.

[62] Zafar, Muhammad Bilal, 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.

https://doi.org/10.52202/075280-0622

Supplementary Material

https://doi.org/10.52202/075280-0622

A Supporting Material for Section 3

A.1 Enumeration Routine for Algorithm 1

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 interfaces $\mathbb{T}$ that obey ordering and plausibility constraints.

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

procedure VIABLETREES($\mathcal{G}, \mathcal{D}$)
  if dim($\mathcal{G}$) = 1 **return** $[T_{\mathcal{G}}]$ *base case: we are left with only a single attribute on which to branch*
  $\mathbb{T} \leftarrow [\,]$
  for each group attribute $\mathcal{A} \in [\mathcal{G}_1, \ldots, \mathcal{G}_k]$ **do**
    $T_{\mathcal{A}} \leftarrow$ reporting tree of depth 1 with $|\mathcal{A}|$ leaves
    $\mathcal{S} \leftarrow$ ViableTrees($\mathcal{G} \setminus \mathcal{A}, \mathcal{D}$) *all subtrees using all attributes except $\mathcal{A}$*
    for Π in ValidAssignments($\mathcal{S}, \mathcal{A}, \mathcal{D}$) **do**: *each assignment is a permutation of $|\mathcal{A}|$ to leaves of $T_{\mathcal{A}}$*
      $\mathbb{T} \leftarrow \mathbb{T} \cup T_{\mathcal{A}}$.assign($\Pi$) *extends the tree by assigning subtrees to each leaf*
    end for
  end for
  return $\mathbb{T}$, reporting interfaces for group attributes $\mathcal{G}$ that obey plausibility and ordering constraints
end procedure

The routine enumerates all possible reporting interfaces for a given set of group attributes $\mathcal{G}$ through a recursive branching process. Given a set of group attributes, the routine is called for each attribute that has yet to be considered in the tree Line 4, ensuring a complete enumeration. We note that the routine is only called for building Sequential systems since there is only one possible reporting interface for Minimal and Flat systems.

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).

A.2 Assignment Routine for Algorithm 1

We summarize the routine for AssignModels procedure in Algorithm 3.

Algorithm 3 Assigning Models

procedure ASSIGNMODELS($T, \mathcal{M}, \mathcal{D}$)
  $Q \leftarrow [T.\text{root}]$ *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

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.

https://doi.org/10.52202/075280-0622

A.3 Pruning Routine for Algorithm 1

We summarize the routine used for the PruneLeaves procedure in Algorithm 1. The PruneLeaves routine

Algorithm 4 Pruning Participatory Systems

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

Algorithm 1 takes as input a reporting interface T and a validation sample $\mathcal{D}$, and performs a bottom-up pruning to output a reporting interface T that asks individuals to report attributes that are expected to lead to a gain. The pruning decision at each leaf is based on a hypothesis test that evaluates the gains of reporting for a reporting group on a validation dataset. This test has the form:

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

This procedure evaluates the gains of reporting by comparing the performance of a model assigned at a leaf node h and a model assigned at a parent node h' which does not use the reported information. Here, the null hypothesis H_0 assumes that the parent model performs as well as the leaf model – and thus, we reject the null hypothesis when there is sufficient evidence to suggest that reporting will improve performance in deployment. Our routine allows practitioners to specify the hypothesis test to compute the gains. By default, we use the McNemar test for accuracy [21] and the Delong test for AUC [19, 50]. In general, we can use a bootstrap hypothesis test [20].

A.4 Greedy Induction of Sequential Reporting Interface

We present an additional routine to construct reporting interfaces for sequential systems in Algorithm 5. We include this routine as an alternative option that can be used to construct a reporting interface in settings where it may be impractical or undesirable to enumerate all possible reporting interfaces. The procedure results in a valid reporting interface that ensures gains. However, it does not guarantee an optimal tree in terms of maximizing the overall gain and does not allow to practitioners to choose between reporting interfaces after training.

Algorithm 5 Greedy Induction Routine for Sequential Reporting Interfaces

procedure GREEDYTREE($\mathcal{R}$)
  $T \leftarrow$ empty reporting interface
  repeat
    for $\boldsymbol{r} \in \text{leaves}(T)$ **do**
      $\{\mathcal{A}_{\boldsymbol{r}}\} \leftarrow G_i : \boldsymbol{r}[i] = \varnothing$ *$\{\mathcal{A}_{\boldsymbol{r}}\}$ contains all heretofore unused attributes*
      $\mathcal{A}^* \leftarrow \text{argmax}_{\mathcal{A} \in \{\mathcal{A}_{\boldsymbol{r}}\}} \min_{\boldsymbol{r}' \in \boldsymbol{r}.\text{split}(\mathcal{A})} \Delta_{\boldsymbol{r}'}(\boldsymbol{r}', \boldsymbol{r})$
      $\boldsymbol{r}.\text{split}(\mathcal{A}^*)$ *Split on attribute that maximizes worse-case gain*
    end for
  until no splits are added
  return T, reporting interface that ensures gains for reporting each $\mathcal{R}$.
end procedure

Algorithm 5 takes as input a collection of reporting options $\mathcal{R}$ and outputs a single reporting interface using a greedy tree induction routine that chooses the attribute to report to maximize the minimum gain at each step. The procedure uses the reporting options to iteratively construct a reporting tree that branches on all of the attributes in $\mathcal{R}$. The procedure considers each unused attribute for each splitting point and splits on the attribute that provides the greatest minimum gain for the groups contained at that node.

https://doi.org/10.52202/075280-0622
14128

B Description of Datasets used in Section 4 – Experiments

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

Dataset	Reference	Outcome Variable	n	d	m	$\mathcal{G}$
`apnea`	Ustun et al. [55]	patient has obstructive sleep apnea	1,152	28	6	{`age`, `sex`}
`cardio_eicu`	Pollard et al. [44]	patient with cardiogenic shock dies	1,341	49	8	{`age`, `sex`, `race`}
`cardio_mimic`	Johnson et al. [32]	patient with cardiogenic shock dies	5,289	49	8	{`age`, `sex`, `race`}
`coloncancer`	Scosyrev et al. [45]	patient dies within 5 years	29,211	72	6	{`age`, `sex`}
`lungcancer`	Scosyrev et al. [45]	patient dies within 5 years	120,641	84	6	{`age`, `sex`}
`saps`	Allyn et al. [3]	ICU mortality	7,797	36	4	{`age`, `HIV`}

Table 3: 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.

`apnea` We use the obstructive sleep apnea (OSA) dataset outlined in Ustun et al. [55]. 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.

`cardio_eicu` & `cardio_mimic` Cardiogenic shock is an acute condition in which the heart cannot provide sufficient blood to the vital organs [31]. 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 [44]. The `cardio_eicu` dataset contains records for a cohort of patients in the MIMIC-III [32] 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 [45]. 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 [45]. 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 [37]. 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.

https://doi.org/10.52202/075280-0622

C Results for Different Model Classes and Prediction Tasks

In this Appendix, we present experimental results for additional model classes and prediction tasks. We produce these results using the setup 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.

C.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. [55]	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 $n = 1341, d = 49$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$ $\lvert\mathcal{G}\rvert = 8$ groups Pollard et al. [44]	Overall Performance	0.864	0.863	0.863	0.862	0.865	0.966	**0.966**
	Overall Gain	0.002	0.001	0.000	-0.001	0.002	0.103	**0.103**
	Group Gains	-0.005 – 0.003	-0.010 – 0.010	-0.005 – 0.003	-0.010 – 0.010	0.000 – 0.003	0.010 – 0.180	0.010 – 0.180
	Max Disparity	0.009	0.019	0.009	0.019	0.003	0.170	0.170
	Rat. Violations	3	3	3	3	0	0	0
	Imputation Risk	-0.005	-0.010					
	Options Pruned	0/8	0/8	0/27	0/27	6/9	13/27	11/27
	Data Use	100.0%	100.0%	0.0%	0.0%	25.0%	100.0%	95.8%
cardio_mimic $n = 5289, d = 49$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$ $\lvert\mathcal{G}\rvert = 8$ groups Johnson et al. [32]	Overall Performance	0.881	0.881	0.882	0.880	0.881	**0.914**	**0.914**
	Overall Gain	0.000	0.000	0.002	-0.000	0.000	**0.034**	**0.034**
	Group Gains	-0.001 – 0.001	-0.001 – 0.001	-0.001 – 0.001	-0.001 – 0.001	0.000 – 0.001	0.008 – 0.057	0.008 – 0.057
	Max Disparity	0.002	0.002	0.002	0.002	0.001	0.049	0.049
	Rat. Violations	3	3	3	3	0	0	0
	Imputation Risk	-0.001	-0.001					
	Options Pruned	0/8	0/8	0/27	0/27	6/9	9/27	8/27
	Data Use	100.0%	100.0%	0.0%	0.0%	25.0%	100.0%	91.7%
coloncancer $n = 29211, d = 72$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathcal{G}\rvert = 6$ groups Scosyrev et al. [45]	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 Scosyrev et al. [45]	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. [3]	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**.

https://doi.org/10.52202/075280-0622
14130

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

		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. [55]	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 $n = 1341, d = 49$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$ $\lvert\mathcal{G}\rvert = 8$ groups Pollard et al. [44]	Overall Performance	18.6%	17.8%	18.2%	18.6%	18.4%	**5.7%**	6.0%
	Overall Gain	-0.2%	0.6%	0.2%	-0.2%	0.0%	**12.7%**	12.4%
	Group Gains	-3.5% – 1.4%	-2.2% – 3.0%	-3.5% – 1.4%	-2.2% – 3.0%	0.0% – 0.0%	6.0% – 14.9%	6.0% – 14.9%
	Max Disparity	4.9%	5.3%	4.9%	5.3%	0.0%	8.9%	8.9%
	Rat. Violations	2	2	2	2	0	0	0
	Imputation Risk	-3.5%	-2.2%					
	Options Pruned	0/8	0/8	0/27	0/27	8/9	11/27	8/27
	Data Use	100.0%	100.0%	0.0%	0.0%	0.0%	100.0%	91.7%
cardio_mimic $n = 5289, d = 49$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$ $\lvert\mathcal{G}\rvert = 8$ groups Johnson et al. [32]	Overall Performance	19.9%	20.1%	19.9%	20.2%	19.6%	11.5%	**11.4%**
	Overall Gain	-0.3%	-0.5%	-0.3%	-0.6%	0.0%	8.1%	**8.1%**
	Group Gains	-1.1% – 1.3%	-1.3% – 0.5%	-1.1% – 1.3%	-1.3% – 0.5%	0.0% – 0.0%	1.0% – 14.9%	1.0% – 14.9%
	Max Disparity	2.4%	1.7%	2.4%	1.7%	0.0%	13.8%	13.8%
	Rat. Violations	5	6	5	6	0	0	0
	Imputation Risk	-1.1%	-1.3%					
	Options Pruned	0/8	0/8	0/27	0/27	8/9	6/27	5/27
	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 Scosyrev et al. [45]	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 Scosyrev et al. [45]	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. [3]	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**.

https://doi.org/10.52202/075280-0622

C.3 Random Forests 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. [55]	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 $n = 1341, d = 49$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$ $\lvert\mathcal{G}\rvert = 8$ groups Pollard et al. [44]	Overall Performance	0.896	0.896	0.897	0.886	0.894	**0.987**	0.987
	Overall Gain	0.003	0.003	0.004	-0.007	0.001	**0.094**	0.094
	Group Gains	-0.008 – 0.011	-0.005 – 0.011	-0.008 – 0.011	-0.005 – 0.011	0.000 – 0.004	0.010 – 0.132	0.010 – 0.130
	Max Disparity	0.020	0.016	0.020	0.016	0.004	0.122	0.120
	Rat. Violations	3	4	3	4	0	0	0
	Imputation Risk	-0.008	-0.005					
	Options Pruned	0/8	0/8	0/27	0/27	7/9	10/27	10/27
	Data Use	100.0%	100.0%	0.0%	0.0%	12.5%	100.0%	87.5%
cardio_mimic $n = 5289, d = 49$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$ $\lvert\mathcal{G}\rvert = 8$ groups Johnson et al. [32]	Overall Performance	0.884	0.883	0.884	0.881	0.885	**0.955**	0.954
	Overall Gain	0.000	-0.001	0.001	-0.002	0.001	**0.071**	0.071
	Group Gains	-0.005 – 0.006	-0.006 – 0.013	-0.005 – 0.006	-0.006 – 0.013	0.000 – 0.006	0.016 – 0.108	0.016 – 0.107
	Max Disparity	0.011	0.019	0.011	0.019	0.006	0.092	0.090
	Rat. Violations	3	7	3	7	0	0	0
	Imputation Risk	-0.005	-0.006					
	Options Pruned	0/8	0/8	0/27	0/27	5/9	6/27	6/27
	Data Use	100.0%	100.0%	0.0%	0.0%	37.5%	100.0%	83.3%
coloncancer $n = 29211, d = 72$ $\mathcal{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathcal{G}\rvert = 6$ groups Scosyrev et al. [45]	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 Scosyrev et al. [45]	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. [3]	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**.

https://doi.org/10.52202/075280-0622
14132

D Supporting Material for Performance Profiles

In the performance profiles, we measure the benefit of disclosure in terms of their expected performance gain and simulate the cost of reporting for each individual by sampling their reporting cost from a uniform distribution – i.e., for each individual i, we sample c_i as $c_i \sim \text{Uniform}(0, \gamma),$ where $\gamma \in [0, 0.2]$. For each value of γ, we sample reporting costs ten times and average over the per group performance error for each sampled cost.

https://doi.org/10.52202/075280-0622