Participatory Personalization in Classification

Supplementary Material

A Supporting Material for Section 2 – Participatory Systems **1**

A.1 Agent Model for Individual Disclosure 1

A.2 Profiling System Performance with Respect to Participation 2

B Supporting Material for Section 3 – Learning Participatory Systems **3**

B.1 Enumeration Routine for Algorithm 1 3

B.2 Assignment Routine for Algorithm 1 3

B.3 Pruning Routine for Algorithm 1 4

B.4 Greedy Induction of Sequential Reporting Interface 4

C Description of Datasets used in Section 4 – Experiments **5**

D Experimental Results for Model Classes and Prediction Tasks **6**

D.1 Logistic Regression for Ranking (AUC) 6

D.2 Random Forests for Decision-Making (Error) 7

D.3 Random Forests for Ranking (AUC) 8

A Supporting Material for Section 2 – Participatory Systems

A.1 Agent Model for Individual Disclosure

The performance of participatory systems will depend on individual reporting decisions. In what follows, we characterize how participatory systems will perform 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:

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

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 $f : \mathcal{X} \times \mathcal{R} \to Y$ 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 f 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), individuals incur a cost of disclosure without receiving a benefit. In such instances, a minimal system $f : \mathcal{X} \times \mathcal{R}^{\min} \to Y$ would allow individuals to opt out of detrimental personalization and receive predictions from a generic model.
- Every participatory system f 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 Profiling System Performance with Respect to Participation

We can use the models for individual disclosure to evaluate how a participatory system will perform once it is deployed. Given a participatory system, we can conduct this evaluation by simulating the parameters in the individual disclosure model shown above. We can then summarize the results from this evaluation for each intersectional group through a performance profile that shows how the system performance will vary across different levels of participation.

We show performance profiles for participatory systems built for the `saps` dataset in Fig. 3. Here, 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 10 times and average over the per group performance error for each sampled cost.

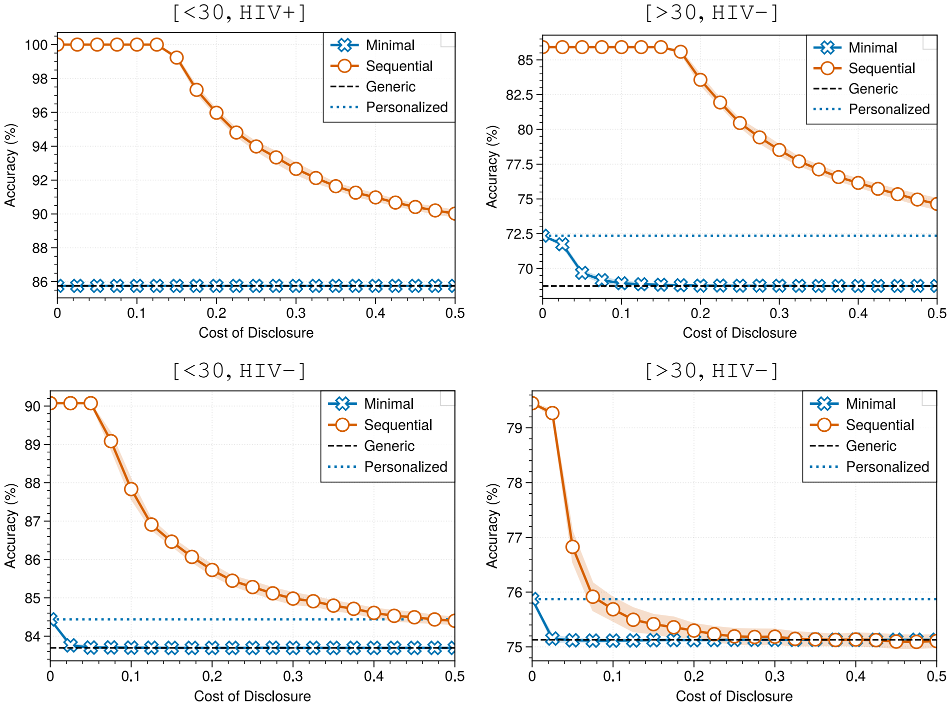

Figure 5: Performance profiles of the simulations performed for each intersectional group in the `saps` dataset. The sequential system outperforms static personalized systems when all group attributes are reported. When the cost of reporting is high, the sequential system still outperforms minimally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

B Supporting Material for Section 3 – Learning Participatory Systems

B.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. 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 6, 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.

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

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

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).

B.2 Assignment Routine for Algorithm 1

We summarize the routine for AssignModels procedure in Algorithm 3.

Algorithm 3 Assigning Models

```
procedure AssignModels($T, \mathcal{M}, \mathcal{D}$)
    $Q \leftarrow [T.\mathsf{root}]$        ▷ initialize with the root of the tree, reporting group $\varnothing$
    while $Q$ is not empty do
        $\boldsymbol{r} \leftarrow Q.\mathsf{pop}()$
        $\mathcal{M}_{\boldsymbol{r}} \leftarrow \mathsf{ViableModels}(\mathcal{M}, \boldsymbol{r})$        ▷ filter $\mathcal{M}$ to models that can be assigned to $\boldsymbol{r}$
        $h^* \leftarrow \operatorname{argmin}_{h \in \mathcal{M}_{\boldsymbol{r}}} \hat{R}_{\boldsymbol{r}}(h, \mathcal{D})$        ▷ assign the model with the best training performance
        $T.\mathsf{set\_model}(\boldsymbol{r}, h^*)$
        for $\boldsymbol{r}' \in T.\mathsf{get\_subgroups}(\boldsymbol{r})$ do        ▷ iterate through the children reporting groups of $\boldsymbol{r}$
            $Q.\mathsf{enqueue}(\boldsymbol{r}')$
        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.

B.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($T, \mathcal{D}$)
    $Stack \leftarrow [T.\mathsf{leaves}]$                    initialize stack with all leaves
    repeat
        $\boldsymbol{r} \leftarrow Stack.pop()$
        $h \leftarrow T.\mathsf{get\_model}(\boldsymbol{r})$
        $h' \leftarrow T.\mathsf{get\_model}(\mathrm{pa}(\boldsymbol{r}))$
        if not $\mathsf{Test}(\boldsymbol{r}, h, h', \mathcal{D})$ then    test gains to see if parent model is as good as leaf model
            $T.\mathsf{prune}(\boldsymbol{r})$
        end if
        if $T.\mathsf{get\_children}(\mathrm{pa}(\boldsymbol{r}))$ is empty then    consider pruning the parent if the parent has become a leaf
            $Stack.\mathsf{enqueue}(\mathrm{pa}(\boldsymbol{r}))$
        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].

B.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 tree with single leaf
    repeat
        for $\boldsymbol{r} \in \mathrm{leaves}(T)$ do
            $\{\mathcal{A}_{\boldsymbol{r}}\} \leftarrow G_i : \boldsymbol{r}[i] = \varnothing$    $\{\mathcal{A}_{\boldsymbol{r}}\}$ contains all heretofore unused attributes
            $\mathcal{A}^* \leftarrow \mathrm{argmax}_{\mathcal{A} \in \{\mathcal{A}_{\boldsymbol{r}}\}} \min_{\boldsymbol{r}' \in \boldsymbol{r}.\mathsf{split}(\mathcal{A})} \Delta_{\boldsymbol{r}'}(\boldsymbol{r}', \boldsymbol{r})$
            $\boldsymbol{r}.\mathsf{split}(\mathcal{A}^*)$    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.

C 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. [43]	patient with cardiogenic shock dies	1,341	49	8	{age, sex, race}
cardio_mimic	Johnson et al. [30]	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: 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.

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 [29]. 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 [43]. The cardio_eicu dataset contains records for a cohort of patients in the MIMIC-III [30] 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 [35]. 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.

D Experimental Results for 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.

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. [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. [43]	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. [30]	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**.

D.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. [43]	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. [30]	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**.

D.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. [43]	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. [30]	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**.