# Understanding challenges to the interpretation of disaggregated evaluations of algorithmic fairness

Stephen R. Pfohl*,[1], Natalie Harris[1], Chirag Nagpal[1], David Madras[2], Vishwali Mhasawade[3], Olawale Salaudeen[4], Awa Dieng[2], Shannon Sequeira[1], Santiago Arciniegas[5], Lillian Sung[5], Nnamdi Ezeanochie[1], Heather Cole-Lewis[1], Katherine Heller[1], Sanmi Koyejo[6], Alexander D'Amour[2]

[1]Google Research, Mountain View, CA, USA
[2]Google DeepMind, Mountain View, CA, USA
[3]New York University, New York, NY, USA
[4]Massachussetts Institute of Technology, Cambridge, MA, USA
[5]The Hopsital for Sick Children, Toronto, ON, Canada
[6]Stanford University, Stanford, CA, USA

**Abstract**

Disaggregated evaluation across subgroups is critical for assessing the fairness of machine learning models, but its uncritical use can mislead practitioners. We show that equal performance across subgroups is an unreliable measure of fairness when data are representative of the relevant populations but reflective of real-world disparities. Furthermore, when data are not representative due to selection bias, both disaggregated evaluation and alternative approaches based on conditional independence testing may be invalid without explicit assumptions regarding the bias mechanism. We use causal graphical models to predict metric stability across subgroups under different data generating processes. Our framework suggests complementing disaggregated evaluations with explicit causal assumptions and analysis to control for confounding and distribution shift, including conditional independence testing and weighted performance estimation. These findings have broad implications for how practitioners design and interpret model assessments given the ubiquity of disaggregated evaluation.

## 1 Introduction

A significant body of work uses disaggregated evaluation of machine learning models across subgroups (e.g. by race, ethnicity, or gender) to assess algorithmic fairness properties [1]. In this paradigm, differences in a performance metric (*e.g.*, accuracy, sensitivity, specificity, positive predictive value) or other statistical property (*e.g.* calibration or the distribution of predictions or covariates) across subgroups are taken as evidence of a fairness violation. For example, this paradigm is commonly used for assessment of fairness in machine learning for healthcare [2–6].

We assume a setting where the policy and fairness goal is consistent with accurate prediction for all subgroups of interest based on historical data. This can be justified when optimal risk estimates are monotonically related to the intervention treatment effect and allocation is unconstrained [7–9]. However, we emphasize that fairness is not an inherent property of a model, but rather related to the effect on outcomes that an intervention or policy leveraging the model has in a deployment context [10], and even accurate models can introduce harm and exacerbate disparities [11, 12].

∗ Correspondence to spfohl@google.com

Code to reproduce the experiments is available at `https://github.com/google-research/google-research/tree/master/causal_evaluation`. A preliminary version of this article was published as a non-archival workshop paper titled "Understanding subgroup performance differences of fair predictors using causal models" at the *NeurIPS 2023 Workshop on Distribution Shifts (DistShift): New Frontiers with Foundation Models.*

In this work, we investigate ways that disaggregated evaluation of the performance of predictive models across subgroups can be misleading. For example, one may collect a large, high-quality dataset representative of one or more target population(s) for a clearly-defined use case [13–15], fit a high-quality predictive model to that data, but find that the model performs differently on average across subgroups of the population, ostensibly presenting a fairness or robustness concern. In this case, we argue that models that maximize performance for all subgroups do not generally obtain equal model performance across subgroups because the optimal value of a performance metric typically changes under distribution shift [16], and data distributions tend to differ across subgroups in contexts where disparities are present. For example, differential exposure to social and structural determinants of health contribute to differences in population health across racial and ethnic groups, leading to differences in the distributions of the covariates and outcomes across subgroups in the cohorts used to develop and evaluate models of health outcomes [17, 18]. Alternatively, if the observed data is misrepresentative of the intended target population, a model that predicts outcomes well in the observed data may not generalize, potentially directly introducing fairness-related harms when the structure of the misgeneralization is systematic across subgroups. For example, under various forms of bias affecting problem formulation, data collection, and measurement, accurate prediction in the observed data imply structured forms of misestimation in the target population that cannot be detected in the observed data without knowledge of the structure of the bias [13, 14, 19–23].

We argue that building understanding of the issues discussed is critical to the design and interpretation of disaggregated evaluations. To that end, we make several contributions, which can be summarized as follows:

- We characterize the properties of models learned and evaluated under a variety of data generating processes. We use causal graphical models of distribution shift to encode structured forms of heterogeneity across subgroups and to describe explicit forms of distribution shift through selection bias. This enables anticipation of the expected fairness properties of models under various assumptions on the data generating process. This approach builds on prior works that use causal directed acyclic graphs to study algorithmic fairness and distribution shift [23–30], as well as those that study incompatibilities between different notions of fairness [31–34].

- We present theoretical results that show that in simple, prototypical cases, average performance of the optimal predictor is expected to differ across subgroups in a target population, but that these differences can be directly anticipated based on the causal structure of the data generating process. In some cases, performance differences can be directly explained by differences in the marginal distribution of a confounder across subgroups, motivating the use of evaluation procedures that control for such confounding.

- We investigate the use of weighted evaluation procedures as a means of constructing evaluations that control for confounding due to distributional differences across subgroups, building off of Cai et al. [16]. We show how such procedures can be interpreted as a class of configurable conditional independence tests and provide guidance for the use of such techniques in concert with standard disaggregated evaluations.

- We conduct experiments with synthetic and real-world data to empirically verify the properties suggested by our theoretical analysis.

## 2 Preliminaries

We consider data with covariates $X \in \mathbb{R}^n$, a binary label $Y \in \{0, 1\}$, and a categorical subgroup indicator $A \in \mathcal{A}$. We reason about properties of a model $f$ that takes as inputs $Z \subseteq \{X, \{X, A\}\}$ to produce scores $R = f(Z)$ that can be compared to a threshold $\tau$ to yield binary predictions $\hat{Y} = \mathbb{1}[R \geq \tau]$.

In this work, we primarily reason about the properties of oracle models $f^*$ that can be considered to return $\mathbb{E}[Y \mid Z]$, such that $f^*(Z) = \mathbb{E}[Y \mid Z]$. Following Mhasawade et al. [23], we define $f^*(X)$ as the *population* Bayes-optimal model that returns the conditional expectation of $Y$ given covariates $X$, such that $R^* = f^*(X) = \mathbb{E}[Y \mid X]$ and the *subgroup* Bayes-optimal model as $R^*_A = f^*_A(X, A) = \mathbb{E}[Y \mid X, A]$. The subgroup Bayes-optimal model can also be represented as a set of subgroup-specific Bayes-optimal models ($\{f^*_a\}_{a \in \mathcal{A}}$ for $R^*_a = f^*_A(X, A = a) = f^*_a(X) = \mathbb{E}[Y \mid X, A = a]$). For arbitrary, non-optimal models, we refer to

models that only have access to $X$ as *subgroup-agnostic* and those that have access to $A$ as *subgroup-aware*. We refer to fitting separate models for each subgroup as *stratified* prediction.

Because our scope is limited to modeling binary outcomes, it follows that $\mathbb{E}[Y \mid X] = P(Y = 1 \mid X)$. This implies that the Bayes-optimal predictor $f^*(Z)$ captures all of the information that $Z$ has about $Y$, and thus $Y \perp Z \mid f^*(Z)$ and $\mathbb{E}[Y \mid f^*(Z)] = \mathbb{E}[Y \mid f^*(Z), Z]$. Furthermore, Bayes-optimal predictors are calibrated, meaning that $c(r) = r$ for all $r \in [0, 1]$ for a calibration curve $c(r) = \mathbb{E}[Y \mid R = r]$. We note that calibration is necessary but not sufficient for Bayes-optimality: miscalibration implies lack of Bayes-optimality, but calibration does not imply Bayes-optimality.

We assume that the model of interest is fit and evaluated based on data drawn from a *source* distribution over $\{X, Y, A\}$ given by $P(X, Y, A)$ and evaluated on a *target* distribution indicative of the target population for which it is of interest to deploy the model. Unless stated otherwise, we assume that the source distribution matches the target distribution. When it is of interest to indicate a systematic difference between source and target distributions, we use the convention of *selection* [19], where we consider models fit using data drawn from the selected, source population $P(X, Y, A \mid S = 1)$ and reason about the properties of those models evaluated on samples drawn from the full, target population without selection $P(X, Y, A)$.

## 2.1 Algorithmic Fairness and Robustness

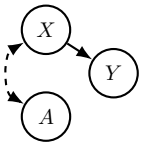
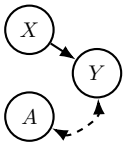
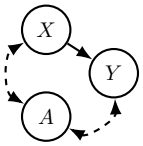
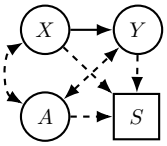
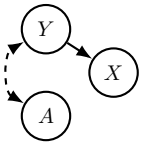
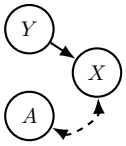
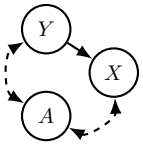
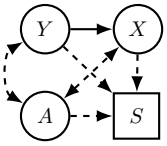
It is common to reason about algorithmic fairness through disaggregated evaluation of model performance over subgroups of the population or otherwise testing for some form of independence or conditional independence involving subgroup membership. For disaggregated evaluation, we consider performance metrics $m : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ that are decomposable (*i.e.*, can be computed at an instance-level and aggregated as a mean over a distribution), where the induced fairness condition is given by $\mathbb{E}[m_j(Y, R) \mid A = a] = \mathbb{E}[m_j(Y, R)]$ for some fixed set of subgroups $\{a\}_{a \in \mathcal{A}}$ and metrics $\{m_j\}_{j=1}^{M}$.

Popular operationalizations of fairness as (conditional) independence include demographic parity ($R \perp A$) [35, 36], separation ($R \perp A \mid Y$) [37], equalized odds ($\hat{Y} \perp A \mid Y$) [37], sufficiency ($Y \perp A \mid R$) [31], and predictive parity ($Y \perp A \mid \hat{Y} = 1$) [33]. Some of these criteria can be directly interpreted as conditions of equal performance across subgroups. For example, separation is a sufficient condition for equalized odds, which corresponds to a condition where both true positive error rates (also known as sensitivity or recall) and false positive error rates (also known as 1-specificity) are equal across subgroups; predictive parity likewise corresponds to a case where the positive predictive value (also known as precision) of the model is equal across subgroups; and demographic parity is sufficient to equalize the classification rate $\mathbb{E}[\mathbb{1}[f(Z) \geq \tau]]$ for any threshold $\tau$. Sufficiency cannot be directly described in terms of equal performance across subgroups, but can instead be described as equivalence of calibration curves across groups.

Other works have framed fairness as a form of model robustness over subgroups [38, 39], where robustness is conceptualized in terms of worst-case performance across subgroups and the optimization objective is formulated in terms of maximizing worst-case subgroup performance. This perspective on robustness is aligned with the perspective of fairness implicitly taken in disaggregated evaluation, in that some deviation in model performance across subgroups is taken as evidence of disparities in model quality across subgroups.

A number of prior works have documented theoretical and empirical incompatibilities and trade-offs between the notions of fairness outlined above [4, 31–34]. Particularly relevant to this work are the findings of Liu et al. [31], where it is shown that, in settings where the prevalence of the label differs across subgroups (*i.e.*, $A \not\perp Y$), fitting a high-quality predictive model for each subgroup implies satisfaction of calibration and sufficiency, but violation of separation and demographic parity, in the sense that empirical risk minimization with a covariate set that encodes $A$ implies some non-zero lower bound on separation and demographic parity error while minimizing an upper bound on calibration and sufficiency error. Other related works [9, 40, 41] show that when it is assumed that the utility of allocation on the basis of the predictor is simple (*e.g.*, monotonically increasing as a function of outcome risk) and unconstrained, maximizing prediction accuracy for each subgroup is consistent with maximizing the utility of the allocation for each subgroup.

**Table 1: Causal graphs encoding assumptions regarding heterogeneity across subgroups**. $X$ indicates covariates, $Y$ a binary label, $A$ subgroup membership, and $S$ selection.

| | Covariate Shift | Outcome Shift | Complex Shift | Selection |
|---|---|---|---|---|
| **Causal Direction** |   |   |   |   |
| | **Label Shift** | **Presentation Shift** | **Complex Shift** | **Selection** |
| **Anticausal Direction** |   |   |   |   |

# 3 A Causal Approach to Understanding Subgroup Fairness and Robustness

## 3.1 Describing subgroup heterogeneity with causal directed acyclic graphs

Our approach relies on the use of causal directed acyclic graphs to describe the causal structure of the data generating processes of interest [42]. These graphs involve $X$, $Y$, and $A$ and are analogous to those used to formalize distribution shift mechanisms in many prior works [24–29]. We consider graphs in both the *causal* direction and *anticausal* direction, that is, when $X$ is a direct parent of $Y$ and when $Y$ is a direct parent of $X$ [43]. While we include $A$ in these graphs to describe the role of heterogeneity across subgroups in these settings, we do not consider $A$ to be a direct cause of either $X$ or $Y$. Rather, we use bidirected edges to describe cases where an unobserved confounder that influences $X$ or $Y$ varies in distribution across subgroups [44].

The causal settings that we study are presented in (Table 1). In brief, for each of the causal and anticausal directions, we consider two simple mechanisms of heterogeneity across subgroups, a complex shift mechanism that combines the two simpler ones, and a configurable graph that incorporates a selection node to describe distribution shift between a source and target distribution.

In the causal direction, the two simple cases are *covariate shift* across subgroups and *outcome shift*. Informally, covariate shift captures the assumption that an outcome $Y$ is generated from covariates $X$ with the same mechanism for all subgroups; outcome shift captures the assumption that the mechanism differs. More formally, under covariate shift, $P(X) \neq P(X \mid A)$ but $P(Y \mid X) = P(Y \mid X, A)$, encoding the invariance $Y \perp A \mid X$. In other words, the distribution of $X$ differs across subgroups but the conditional distribution $Y \mid X$ is unchanged. Under outcome shift, we assume that $P(X) = P(X \mid A)$, but there exists some unobserved confounder that affects $Y$ and is not independent of $A$, such that $P(Y \mid X) \neq P(Y \mid X, A)$. The complex shift case combines the two settings such that $P(X) \neq P(X \mid A)$ and $P(Y \mid X) \neq P(Y \mid X, A)$.

In the anticausal direction, the two simple cases are *label shift* and *presentation shift*. Informally, label shift captures the assumption that conditioned on a class of $Y$, the distribution of covariates $X$ is identical regardless of subgroup membership, while presentation shift allows this relationship to change and holds the prevalence of $Y$ constant. More formally, in the label shift case, the prevalence of $Y$ differs across subgroups, but $P(X \mid Y)$ is stable across subgroups (i.e., $P(X \mid Y) = P(X \mid Y, A)$). In the presentation shift case, the prevalence of $Y$ is the same across groups, but $P(X \mid Y)$ differs across subgroups (*i.e.*, $P(X \mid Y) \neq P(X \mid Y, A)$). As in the causal direction, a complex anticausal shift can be constructed by considering a case where $Y$ is upstream of $X$ but $P(Y) \neq P(Y \mid A)$ and $P(X \mid Y) \neq P(X \mid Y, A)$.

When the causal graph includes the square selection node $S$, direct arrows into $S$ indicate that selection into the observed source distribution depends on the value of a set of parent nodes $Pa(S)$. The distribution $P(S \mid X, Y, A)$ can be simplified to $P(S \mid Pa(S))$.

**Table 2: Conditional independence properties of Bayes-optimal models**. ✓ indicates conditions where Bayes-optimal prediction is a sufficient condition for the listed criteria. ✗ indicates that Bayes-optimal prediction is not a sufficient condition for the property. $f^*(Z)$ corresponds to the Bayes-optimal predictor that depends on $Z$.

| | Sufficiency ($Y \perp A \mid f^*(Z)$) | | Separation ($f^*(Z) \perp A \mid Y$) | |
|---|---|---|---|---|
| Setting | $Z = X$ | $Z = \{X, A\}$ | $Z = X$ | $Z = \{X, A\}$ |
| Covariate Shift | ✓ | ✓ | ✗ | ✗ |
| Outcome Shift | ✗ | ✓ | ✗ | ✗ |
| Causal Complex Shift | ✗ | ✓ | ✗ | ✗ |
| Label Shift | ✗ | ✓ | ✓ | ✗ |
| Presentation Shift | ✗ | ✓ | ✗ | ✗ |
| Anticausal Complex Shift | ✗ | ✓ | ✗ | ✗ |

## 3.2 The effect of causal structure on the properties of Bayes-optimal models

Here, we focus our attention on the conditional independence properties of Bayes-optimal models learned and evaluated on data drawn from each of the causal directed acyclic graphs described in Table 1. We do so by considering the conditional independencies among $X$, $Y$, and $A$ that are directly implied by each of the causal graphical structures. For conditional independence statements involving $R$ where $R$ is not the conditioning variable, we consider $R$ as a deterministic function of $Z \subseteq \{X, \{X, A\}\}$ and use the result that $Z \perp V_1 \mid V_2$ implies $R \perp V_1 \mid V_2$ for any variables $V_1$ and $V_2$ ([45], Lemma 4.2). For Bayes-optimal models, we additionally consider the constraint that $Y \perp Z \mid R$ when $R = f^*(Z)$ is Bayes-optimal.

A central property that we study is the conditional independence $Y \perp A \mid X$. Of the settings we study, this criteria holds only for covariate shift. When the causal graph implies $Y \perp A \mid X$, the population Bayes-optimal predictor is also subgroup Bayes-optimal and there is no expected benefit to subgroup-aware modeling. In contrast, when $Y \not\perp A \mid X$, the population Bayes-optimal predictor is not subgroup Bayes-optimal, and subgroup-aware prediction is expected to improve performance. Furthermore, for graphs without selection, $Y \perp A \mid X$ is satisfied only in the covariate shift case. This suggests that subgroup membership is informative about $Y$ after accounting for $X$ for causal-direction graphs when the covariate shift assumption is violated and in general for anticausal graphs under either label shift or presentation shift.

The conditional independencies directly implied by the causal graph of the data generating process can also be used to understand incompatibilities between fairness criteria [31–34]. We report cases where the structure of the graph implies that the sufficiency and separation criteria is implied by the graph structure in Table 2. We note that we focus on sufficient conditions where Bayes-optimality is sufficient to imply the fairness criteria of interest, which are weaker than impossibility results. For example, sufficiency is implied by subgroup Bayes-optimal prediction, and is thus implied by population Bayes-optimal prediction when $Y \perp A \mid X$, but not when $Y \not\perp A \mid X$. The separation criteria is implied only for the label shift graph and then only when prediction is subgroup-agnostic, regardless of whether the model is Bayes-optimal or arbitrary. This is consistent with prior findings [31] showing that models that fit the data well for subgroups satisfy sufficiency but do not generally satisfy the separation and equalized odds criteria. This serves as a special case for our general findings regarding the stability of performance metrics across subgroups, as it implies that optimal prediction for subgroups does not imply equal true positive rates and false positive rates across subgroups.

In supplementary section A.2, we extend this analysis to settings with selection bias, enumerating the conditions analogous to those described thus far for each of the base graphs augmented with a selection mechanism that depends on combinations of $X$, $Y$, and $A$ (Supplementary Table B1). In brief, a Bayes-optimal model generalizes under selection if the variables used as covariates d-separate the selection node $S$ from $Y$ in the causal graph. For example, when selection depends on only $X$ or $A$, subgroup Bayes-optimal predictors generalize and satisfy sufficiency in the target domain, regardless of other components of the graph; when selection depends only on $Y$, subgroup Bayes-optimality implies sufficiency without calibration in the target domain; and when selection depends on $Y$ and either $X$ or $A$, subgroup Bayes-optimality in the source

**Table 3: Stability of model performance metrics across subgroups**. ✓ indicates cases in which prediction is sufficient to induce $\{R, Y\} \perp A \mid V$; ✗ indicates otherwise. $f$ indicates prediction with an arbitrary model. $f^*$ and $f_A^*$ indicate the population and subgroup Bayes-optimal models.

| | Setting | | $\{R, Y\} \perp A \mid V$ | | | |
|---|---|---|---|---|---|---|
| Setting | $Z \subseteq \{X, \{X, A\}\}$ | Model | $V =$ $\{\}$ | $X$ | $Y$ | $R$ |
| Covariate shift | $X$ | $f^*$ | ✗ | ✓ | ✗ | ✓ |
| | | $f$ | ✗ | ✓ | ✗ | ✗ |
| | $\{X, A\}$ | $f_A^*$ | ✗ | ✓ | ✗ | ✓ |
| | | $f$ | ✗ | ✗ | ✗ | ✗ |
| Label shift | $X$ | $f^*$ | ✗ | ✗ | ✓ | ✗ |
| | | $f$ | ✗ | ✗ | ✓ | ✗ |
| | $\{X, A\}$ | $f_A^*$ | ✗ | ✗ | ✗ | ✓ |
| | | $f$ | ✗ | ✗ | ✗ | ✗ |
| Other $Y \not\perp A \mid X$ | $X$ | $f^*$ | ✗ | ✗ | ✗ | ✗ |
| (Outcome Shift, Presentation Shift, Complex Shift) | | $f$ | ✗ | ✗ | ✗ | ✗ |
| | $\{X, A\}$ | $f_A^*$ | ✗ | ✗ | ✗ | ✓ |
| | | $f$ | ✗ | ✗ | ✗ | ✗ |

domain does not imply sufficiency in the target domain. Furthermore, separation in the target domain is implied by Bayes-optimal prediction only in the label shift graph when the selection mechanism does not depend on $A$, and then only when the model does not depend on $A$.

### 3.3 Stability of model performance across subgroups

In this section, we consider the extent to which the differences in model performance metrics across subgroups can be explained by structured forms of heterogeneity. We do so through characterization of sufficient conditions for model performance to be equal across subgroups for each of the settings discussed, focusing on cases where selection bias is not present. In general, $\{R, Y\} \perp A$ is a sufficient condition for equal average performance across subgroups for any metric that depends only on $R$ and $Y$, as fixing the distribution $P(R, Y)$ fixes the distribution of $m$. By a contrapositive argument, if average performance is unequal, then it follows that $\{R, Y\} \not\perp A$. The converse is not necessarily true, in that it may be that $\{R, Y\} \not\perp A$ but average performance is equal across groups for some metric. Notably, in *none* of the settings discussed thus far is either population or subgroup Bayes-optimal prediction sufficient to induce $\{R, Y\} \perp A$.

In some cases, performance differences across subgroups can be attributed to specified distributional differences across subgroups. To do so, we introduce a control variable $V$ and consider computation of average performance when the distribution of $V$ has been set to some reference distribution $P_0(V)$. Then, average performance with respect to $P_0(V)$ can be written as $\int m(R, Y)P(R, Y \mid v)P_0(v)dv$. If performance is unequal across subgroups marginally in the observed data, but there exists a control variable $V$ that satisfies $\{R, Y\} \perp A \mid V$, it follows that $P(R, Y \mid V) = P(R, Y \mid V, A)$ and the difference in performance can be explained by the lack of independence of $A$ and $V$. In such cases, we say that $V$ explains the difference in performance across subgroups because fixing the distribution to a reference distribution $P_0(V)$ ensures equal performance across subgroups.

In Table 3, we summarize whether control for either $V = X$, $Y$, or $R$ is sufficient to induce $\{R, Y\} \perp A \mid V$ based on the conditional independencies implied by each of the graphical settings discussed. For models that are subgroup-agnostic, we find that $X$ explains performance differences only for the covariate shift graph and $Y$ explains performance differences only for the label shift graph. However, these relationships no longer hold if the model is subgroup-aware because now $\{R, Y\} \perp A \mid V$ no longer necessarily holds, with the only exception being that, in the covariate shift graph, subgroup Bayes-optimality implies equal performance conditioned on $X$.

In graphs where $Y \not\perp A \mid X$, performance differences are expected in general and neither $X$ nor $Y$ alone

explains those differences. However, if subgroup Bayes-optimality holds, the difference in performance can be explained by differences in the distribution of $R$, because subgroup-Bayes optimality implies sufficiency and thus $\{R, Y\} \perp A \mid R$. In other words, the differences in performance for the set of optimal predictors for each subgroup can be explained by the differences in the distribution of the *optimal risk score* across subgroups.

### 3.4 Controlled evaluation

The observation that performance differences across subgroups may be attributed to specific distributional differences across subgroups suggests that evaluations that control for this source of confounding may be constructed through balancing of the confounding variable $V$ in comparisons of model performance across subgroups or between subgroups and the overall population. Our approach is to map the population distribution $P(V)$ to the subgroup distribution $P(V \mid A = a)$ for each subgroup. We then compare performance in each of the mapped distributions to the corresponding subgroup performance.

Concretely, we compute performance in a mapped distribution with a weighted average of the form $M_a \coloneqq \int w * m(R, Y) P(R, Y \mid v) P(v) \mathrm{d}v$ for weights $w \propto P(A = a \mid V)$, computed separately for each subgroup. This is a computation of a metric over the entire population, weighted such that the distribution of $V$ has been set to $P(V \mid A = a)$ and the distribution of $\{Y, R\}$ to $\int P(R, Y \mid v) P(v \mid A = a) \mathrm{d}v$. With this construction, the relevant comparison is with the analogous unweighted average computed on the subgroup $A = a$, yielding a statistic $T_a \coloneqq \mathbb{E}[m(R, Y) \mid A = a] - M_a$. In supplementary section A.1, we provide additional technical details regarding this weighting approach and describe alternative configurations of the weights, including the approach of Cai et al. [16] that maps the data to a region of shared overlap between a pair of distributions.

We note that a hypothesis test for whether $T_a$ differs from zero corresponds to a conditional independence test against the null hypothesis that $\{R, Y\} \perp A \mid V$. In general, $T_a$ is expected to be zero when $V$ is chosen to be a variable marked with an ✓ in Table 3. However, a few specific configurations of $V$ and $Z$ correspond to notable conditional independence tests. For example, when $V = X$ and $Z = X$, the test corresponds to a test against the null hypothesis $Y \perp A \mid X$. Furthermore, if $V = R$, regardless of $Z$, then the test corresponds to a test for sufficiency, as $\{R, Y\} \perp A \mid R$ simplifies to $Y \perp A \mid R$.

It is important to note that the controlled evaluation procedures that we propose may be misleading for suboptimal models because they do not necessarily distinguish between model errors caused by poor model fit from those that would remain when prediction is subgroup Bayes-optimal. For example, as an arbitrarily poor subgroup-agnostic predictor satisfies $\{R, Y\} \perp A \mid X$ (Table 3) in a covariate shift setting just as the population Bayes-optimal predictor does, a weighted evaluation procedure that controls for $X$ cannot distinguish between differences in performance across subgroups due to poor fit of the model to the data distribution (*e.g.*, due to underrepresentation or model misspecification) from those implied by the structure of the data distribution. Similarly, an evaluation that control for $R$ inherits the same limitations as a test for sufficiency, in that it is possible for poorly-fit models with severe fairness issues to satisfy sufficiency [40, 46].

### 3.5 Subgroup Separability

We highlight the properties of a special case where the distribution of covariates are separable across subgroups [3, 47]. In this setting, there is little to no overlap in the distribution of $X$ for any pair of subgroups $a_i$ and $a_j$, such that the ratio $\frac{P(X|A=a_i)}{P(X|A=a_j)}$ is large for any $X$ where $P(X \mid A = a_i)$ is non-trivial. This implies that each region of $X$ with non-trivial density is associated with exactly one subgroup, and $A$ can be predicted with high accuracy from $X$.

Under subgroup separability, we observe that the population and subgroup Bayes-optimal models behave similarly, as if $Y \perp A \mid X$, regardless of the underlying causal structure that generated the data. To see this, consider that $\mathbb{E}[Y \mid X] = \sum_{a \in \mathcal{A}} \mathbb{E}[Y \mid X, A = a] P(A = a \mid X)$ [47]. As such, for $X$ where $P(X \mid A) \gg 0$, we have that $P(A \mid X) \approx 1$ and $\mathbb{E}[Y \mid X] \approx \mathbb{E}[Y \mid X, A]$. We then have that, as in the case of covariate shift, the population Bayes-optimal predictor satisfies sufficiency and control for $X$ is enough to explain differences in performance for any performance metric.

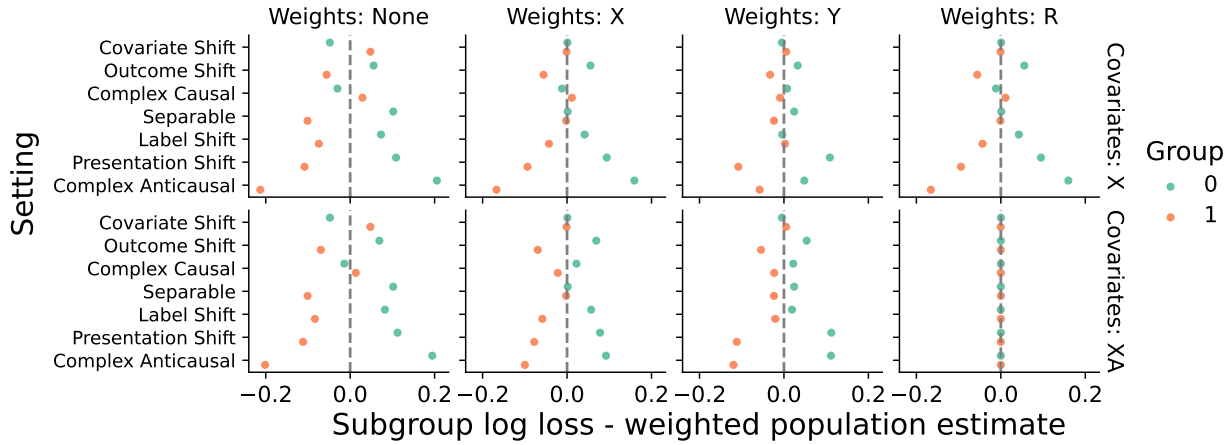



**Figure 1: Controlled evaluation for confounding across subgroups**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between unweighted disaggregated performance with population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction and the second row corresponds to subgroup-aware prediction where $A$ is used as a covariate.

# 4 Experiments

We conduct a simulation study and experiment with real-world tabular data. The purpose of these experiments is to empirically verify the properties discussed in section 3. We briefly describe the design of the experiments here and defer additional details to supplementary sections A.3 and A.4. For the simulation study, we generate data corresponding to each of the settings described in Table 1. For the real-world data experiments, we follow Ding et al. [48] to derive prediction tasks from the American Comunity Survey (ACS) Public Use Microdata Sample (PUMS) provided by the U.S. Census Bureau [49]. We use the 'ACSIncome' (prediction of whether an individual's income exceeds $\$50,000$) and 'ACSPublicCoverage' (prediction of whether an individual is enrolled in a public health insurance plan) task definitions [48], stratifying the data by race and ethnicity (cohort characteristics provided in Supplementary Table B2).

To assess properties with respect to the informativeness of subgroup membership, we assess differences in performance between subgroup-aware and subgroup-agnostic models (as in Liu et al. [50]). In the main text, we report results for subgroup-aware models that leverage $A$ as a covariate and include additional results with subgroup-stratified models in the supplementary material. To assess calibration and sufficiency violation, we visualize calibration curves for each subgroup and conduct evaluations that control for $R$. To verify empirical consistency with the properties presented in Table 3, we compute weighted estimates of performance ($M_a$) and test statistics ($T_a$) to control for $X$, $Y$, or $R$ in the comparison of population and subgroup performance.

**Simulation study results:** The results of the simulation study generally coincide with those that are expected based on the theoretical analysis. We find that subgroup-aware models generally improve predictive performance overall and for subgroups when $Y \not\perp A \mid X$, *i.e.*, in all settings except for covariate shift, with the exception of the case where subgroups are separable based on $X$ (Supplementary Figure B1). Specifically, we observe improvements overall and for subgroups in log-loss and non-negative changes in AUC-ROC. We note that sensitivity and specificity do not strictly improve; rather an increase in one is often paired with a reduction in the other. However, we observe that net benefit [51], a decision-theoretic metric that combines sensitivity and specificity based on an assumed tradeoff between true positives and false negatives, improves like the log-loss.

The effect of subgroup-aware prediction on calibration and sufficiency mirrors the effects on overall model performance (Supplementary Figure B2). In cases where $Y \not\perp A \mid X$ and prediction is subgroup-agnostic, we observe model miscalibration for subgroups and sufficiency violation. In all cases, subgroup-aware prediction results in calibrated models that satisfy sufficiency. As expected, there is no difference in calibration between

subgroup-agnostic and subgroup-aware models when there is minimal overlap in $X$ across subgroups.

We conduct a small experiment to verify a subset of the properties expected under selection. We adapt the complex causal shift setting and introduce three selection mechanisms corresponding to cases where the selection mechanism depends on $X$, $Y$, or $\{Y, A\}$. Further methodological details are provided in supplementary section A.3.1. We visualize calibration curves in Supplementary Figure B3, finding that they correspond to the properties presented in Supplementary Table B1.

We find the results of weighted evaluation are generally consistent with those presented in Table 3. In general, we find that only for the covariate shift graph or when subgroups are separable, control for $X$ is sufficient to control for differences in the average log loss between the population and subgroups for subgroup-agnostic models, consistent with the interpretation of this form of the controlled evaluation as a conditional independence test for $Y \perp A \mid X$ (Figure 1). This same pattern holds for other metrics including AUC-ROC, sensitivity, specificity, net benefit, precision, and classification rate (Supplementary Figures B5–B10). We further find that performance is not generally stable after control for $Y$, except in the label shift setting, and then only when prediction is subgroup-agnostic. Consistent with the interpretation of control for $R$ as corresponding to a test for sufficiency, we note that the test statistic $T_a$ takes on a value not statistically significantly different from zero in cases where calibration and sufficiency are satisfied and a non-zero value otherwise. For completeness, we report the absolute values of the performance of each of the models of interest in each setting for each subgroup as well as the weighted population estimates $T_a$ (Supplementary Figures B11–B17). For comparison, we further apply the approach of Cai et al. [16] (Supplementary Figures B18–B24), finding that the results are largely qualitatively consistent with ours.

**Results on real-world tabular data:** For the experiments with ACS PUMS, we report the change in performance attained through subgroup-aware prediction (Supplementary Figure B25), visualize calibration curves (Supplementary Figure B26), and apply our approach to controlled evaluation based on weighting (Supplementary Figures B27–B33). While the structure of the causal graph underlying these data is not available, our approach can be used for conditional independence testing and attribution of performance differences to structured distribution shifts. We generally observe greater evidence against the hypothesis $Y \perp A \mid X$ for the 'ACSPublicCoverage' task than we do for the 'ACSIncome' task. Specifically, for 'ACSPublicCoverage', we find that subgroup membership improves prediction performance for nearly all race/ethnicity subgroups (Supplementary Figure B25) and that differences in performance between subgroups and the aggregate population are not generally explained by control for $X$, whereas for 'ACSIncome', improvements in performance through subgroup-aware prediction are more minor, and differences in performance are explained to a greater degree by $X$ (Supplementary Figure B27). Furthermore, while the extent of miscalibration appears to be minor for both tasks (Supplementary Figure B26), we do observe evidence of sufficiency violation for subgroup-unaware prediction for both tasks, affecting the "Other" subgroup for 'ACSIncome' and nearly all subgroups for 'ACSPublicCoverage'. Stratified prediction, but not subgroup-aware prediction leveraging a race/ethnicity covariate mitigates the sufficiency violation for the "Other" subgroup for the 'ACSIncome' task, potentially as a result of suboptimality due to relative underrepresentation of this subgroup in the data.

# 5 Discussion

Our work highlights the challenges associated with the interpretation of disaggregated evaluations over subgroups, consistent with findings of prior work studying related challenges for the evaluation of robustness in settings with distribution shift (*e.g.*, Cai et al. [16] and Liu et al. [50]). Given our findings, we recommend that the controlled evaluation procedures we present be used in conjunction with standard disaggregated evaluations with the set of control variables chosen based on domain knowledge regarding the causal structure of the data generating process. However, we note that even in cases where the causal structure of the data is unknown, our approach can be useful for conditional independence testing. Such conditional independence tests can be used to assess hypotheses with respect to the causal structure of the data (*e.g.*, whether the covariate shift or label shift assumptions hold) and to aid in assessment of model fit and fairness violation (*e.g.*, whether the sufficiency fairness criterion holds).

An important aspect of our analysis is that we primarily focus on the properties of Bayes-optimal

models. This highlights that the issues that we study remain even with arbitrarily large datasets, models that fit the data well, and with measures taken to ensure that the data are representative. However, this also introduces limitations and potential pitfalls with respect to interpretation of the results of controlled evaluation procedures in cases where the evaluated model fits the data poorly, as discussed in section 3.4. Extending our approach to allow for finite-sample estimation error to be decomposed from Bayes-optimal error is an important area of future work.

We emphasize that our findings are primarily relevant to cases where it is assumed that modeling the statistical relationship between the covariates and the label of interest as well as possible for each subgroup is aligned with fairness goals, assuming that the observations used for model development and evaluation are observed without bias and are representative of the intended target population. However, we emphasize that the assumptions underpinning this setting should not be taken for granted, and that there are several unresolved normative questions outside of the scope of this work pertaining to the justification for the prediction and control for disparate outcomes across populations. Designing effective evaluation procedures that are grounded in understanding of both the societal context contributing to inequities and the capacity for interventions and policies that incorporate predictive models to promote equity and fairness goals is a critical area of future work.

# References


[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[2] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

[3] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6): e406–e414, 2022.

[4] Stephen R Pfohl, Agata Foryciarz, and Nigam H Shah. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of biomedical informatics*, 113:103621, 2021.

[5] Richard J Chen, Judy J Wang, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6):719–742, 2023.

[6] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual review of biomedical data science*, 4(1):123–144, 2021.

[7] Donna K Arnett, Roger S Blumenthal, Michelle A Albert, Andrew B Buroker, Zachary D Goldberger, Ellen J Hahn, Cheryl Dennison Himmelfarb, Amit Khera, Donald Lloyd-Jones, J William McEvoy, et al. 2019 acc/aha guideline on the primary prevention of cardiovascular disease: a report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Journal of the American College of cardiology*, 74(10):e177–e232, 2019.

[8] Agata Foryciarz, Stephen R Pfohl, Birju Patel, and Nigam Shah. Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation. *BMJ Health & Care Informatics*, 29(1):e100460, 2022.

[9] Stephen Pfohl, Yizhe Xu, Agata Foryciarz, Nikolaos Ignatiadis, Julian Genkins, and Nigam Shah. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1039–1052, 2022.

[10] Melissa D McCradden, Shalmali Joshi, James A Anderson, and Alex John London. A normative framework for artificial intelligence as a sociotechnical system in healthcare. *Patterns*, 4(11), 2023.

[11] Wouter AC van Amsterdam, Nan van Geloven, Jesse H Krijthe, Rajesh Ranganath, and Giovanni Ciná. When accurate prediction models yield harmful self-fulfilling prophecies. *Patterns*, 6(4), 2025.

[12] Melissa D McCradden, Mjaye L Mazwi, and Lauren Oakden-Rayner. Can an accurate model be bad? *Patterns*, 6(4), 2025.

[13] Samir Passi and Solon Barocas. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 39–48, 2019.

[14] Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385, 2021.

[15] Matthew Sperrin, Richard D Riley, Gary S Collins, and Glen P Martin. Targeted validation: validating clinical prediction models in their intended population and setting. *Diagnostic and prognostic research*, 6 (1):24, 2022.

[16] Tiffany Tianhui Cai, Hongseok Namkoong, and Steve Yadlowsky. Diagnosing model performance under distribution shift. *arXiv preprint arXiv:2303.02011*, 2023.

[17] Zinzi D Bailey, Nancy Krieger, Madina Agénor, Jasmine Graves, Natalia Linos, and Mary T Bassett. Structural racism and health inequities in the usa: evidence and interventions. *The lancet*, 389(10077): 1453–1463, 2017.

[18] Zinzi D Bailey, Justin M Feldman, and Mary T Bassett. How structural racism works—racist policies as a root cause of us racial health inequities. *New England Journal of Medicine*, 384(8):768–773, 2021.

[19] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108. PMLR, 2012.

[20] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[21] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International conference on artificial intelligence and statistics*, pages 702–712. PMLR, 2020.

[22] Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. Ground (less) truth: A causal framework for proxy labels in human-algorithm decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 688–704, 2023.

[23] Vishwali Mhasawade, Alexander D'Amour, and Stephen R Pfohl. A causal perspective on label bias. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1282–1294, 2024.

[24] Victor Veitch, Alexander D' Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 16196–16208, 2021.

[25] Jessica Schrouff, Natalie Harris, Sanmi Koyejo, Ibrahim M Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. *Advances in Neural Information Processing Systems*, 35:19304–19318, 2022.

[26] Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2189–2200. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/creager21a.html`.

[27] Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR, 2022.

[28] Maggie Makar and Alexander D'Amour. Fairness and robustness in anti-causal prediction. *arXiv preprint arXiv:2209.09423*, 2022.

[29] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 3–13, 2021.

[30] Charles Jones, Daniel C Castro, Fabio De Sousa Ribeiro, Ozan Oktay, Melissa McCradden, and Ben Glocker. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*, 6(2):138–146, 2024.

[31] Lydia T Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pages 4051–4060. PMLR, 2019.

[32] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[33] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[34] Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. 2017.

[35] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pages 13–18. IEEE, 2009.

[36] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[37] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[38] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ryxGuJrFvS`.

[39] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021.

[40] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[41] Chloé Bakalar, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, et al. Fairness on the ground: Applying algorithmic fairness approaches to production systems. *arXiv preprint arXiv:2103.06172*, 2021.

[42] Judea Pearl. *Causality*. Cambridge university press, 2009.

[43] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 459–466, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

[44] Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

[45] A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(1):1–15, 1979.

[46] Sam Corbett-Davies, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312):1–117, 2023.

[47] Charles Jones, Mélanie Roschewitz, and Ben Glocker. The role of subgroup separability in group-fair medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 179–188. Springer, 2023.

[48] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[49] U.S. Census Bureau. American community survey (ACS) public use microdata sample (PUMS), california 5-year person file, 2018.

[50] Jiashuo Liu, Tianyu Wang, Peng Cui, and Hongseok Namkoong. On the need for a language describing distribution shifts: Illustrations on tabular datasets. In *Advances in Neural Information Processing Systems*, volume 36, pages 51371–51408, 2023.

[51] Andrew J Vickers and Elena B Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006.

[52] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[54] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265, 2017.

[55] Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[56] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

# Supplementary Material

# A Supplementary methods

## A.1 Weighting approaches to controlled evaluation

The approach to weighting model performance metrics to control for distributional differences across subgroups builds on a general approach to estimating model performance under distribution shift. For a source distribution $\mathcal{P}$ and target distribution $\mathcal{Q}$ over $\{R, Y\}$, performance on $\mathcal{Q}$ can be estimated using data from $\mathcal{P}$ with appropriate weights. Formally, this is $\mathbb{E}_{\mathcal{Q}}[m(Y,R)] = \mathbb{E}_{\mathcal{P}}[w * m(Y,R)]$ for $w \propto \frac{P_{\mathcal{Q}}(R,Y)}{P_{\mathcal{P}}(R,Y)}$, assuming positivity ($P_{\mathcal{Q}}(R,Y) > 0 \rightarrow P_{\mathcal{P}}(R,Y) > 0$). From some mixture distribution over $\mathcal{P}$ and $\mathcal{Q}$, where examples from $\mathcal{Q}$ are indexed by $D = 1$ and those from $\mathcal{P}$ by $D = 0$, we note that the weights can be reformulated as $w \propto \frac{P(R,Y|D=1)}{P(R,Y|D=0)} \propto P(D = 1 \mid R, Y)$.

As described in section 3.4, we construct controlled evaluations by setting the distribution of a variable $V$ to a reference distribution. We consider a source distribution $\mathcal{P}$ and target distribution $\mathcal{Q}$ over $V$ such that we are computing the performance in $\mathcal{P}$ after fixing the marginal to $P_{\mathcal{Q}}(V)$. The resulting expectation is given by $\int w * m(R,Y) P_{\mathcal{P}}(R,Y \mid v) P_{\mathcal{P}}(v) dv$ for weights $w \propto \frac{P_{\mathcal{Q}}(V)}{P_{\mathcal{P}}(V)} \propto P(D = 1 \mid V)$, where $D$ indicates the identity of the distribution, as before.

In this work, we primarily consider setting $\mathcal{P} = P(R,Y)$ and $\mathcal{Q} = P(R,Y \mid A = a)$ with weights $w \propto P(A = a \mid V)$. This yields a mapping of the marginal population distribution $P(V)$ onto the subgroup distribution $P(V \mid A = a)$, implicitly fixing the conditional distribution of $P(R,Y \mid V)$ to that of the aggregate population. It follows that if $\{R,Y\} \perp A \mid V$, then $P_{\mathcal{P}}(R,Y \mid v) = P_{\mathcal{Q}}(R,Y \mid v)$ and thus weighted performance $M_a \coloneqq \int w * m(R,Y) P(R,Y \mid v) P(v) \mathrm{d}v$ in $\mathcal{P}$ is equal to the marginal performance $\mathbb{E}[m(R,Y) \mid A = a]$ in $\mathcal{Q}$. If weighted performance in $\mathcal{P}$ is not equal to the marginal performance in $\mathcal{Q}$, then $\{R,Y\} \not\perp A \mid V$.

It is possible to construct alternative weight configurations that can be used for the same purpose as the approach that we present. For example, if the distributions are set such that $\mathcal{P} = P(R,Y \mid A = a)$ and $\mathcal{Q} = P(R,Y)$, then the weights are given by $w \propto \frac{1}{P(A=a|V)}$, corresponding to a mapping of the subgroup $A = a$ onto the population distribution, analogous to approaches to inverse propensity score weighting for treatment effect estimation [52]. Alternatively, if we consider a pair of subgroups $A = a$ and $A = a'$ where $\mathcal{P} = P(R,Y \mid A = a)$ and $\mathcal{Q} = P(R,Y \mid A = a')$, the weights are given by $w \propto \frac{P(A=a'|V)}{P(A=a|V)}$. However, both of these formulations are susceptible to returning unstable and high-variance estimates when small values in the denominator induce extreme weights. We note that our approach is not susceptible to this issue given that $P(A = a \mid V)$ is bounded between 0 and 1.

Cai et al. [16] presented an alternative weighting strategy for pairwise comparisons motivated to address the extreme weights issue by considering mapping the data to a *shared distribution* of sufficient overlap between the distributions of covariates for a pair of subgroups. This approach considers a target distribution $\mathcal{Q} \propto \frac{P(V|A=a)P(V|A=a')}{P(V|A=a)+P(V|A=a')}$ for two source distributions $\mathcal{P}_a$ and $\mathcal{P}_{a'}$. This has the effect of defining a target density that takes on a value of zero in cases where the control variable has zero support in either of the subgroup distributions. The weights for this approach are given by $w \propto \frac{P(A=a'|V)}{P(A=a)P(A=a'|V)+P(A=a')P(A=a|V)}$ for $A = a$ and $w \propto \frac{P(A=a|V)}{P(A=a)P(A=a'|V)+P(A=a')P(A=a|V)}$ for $A = a'$.

## A.2 Properties under selection

In this section, we consider fairness properties under explicit distribution shift through the lens of selection bias [19]. We consider models fit using data drawn from the selected, *source* population $P(X,Y,A \mid S = 1)$ and reason about the properties of those models evaluated on samples drawn from the full, *target* population without selection $P(X,Y,A)$. Graphically, we represent the selection variable $S$ with a square node to indicate conditioning on selection in the observed data, where a directed edge in to $S$ indicates dependence of the selection mechanism on the originating node (Table 1). This formulation allows for us to anticipate the fairness properties of models under selection based on the structure of the causal graph and the connectivity of the selection node (Supplementary Table 2).

Intuitively, the model learned in the selected population generalizes to the full population if $\mathbb{E}[Y \mid Z, S = 1] = \mathbb{E}[Y \mid Z]$, for a set of predictor variables $Z$. To reason about this, we consider the sufficient condition that $S \perp Y \mid Z$, which implies that

$$\frac{P(Y \mid Z)}{P(Y \mid Z, S = 1)} = \frac{P(S = 1 \mid Z)}{P(S = 1 \mid Z, Y)} = 1. \tag{1}$$

The implication is that a Bayes-optimal model will generalize under selection if the variables used as predictors d-separate the selection node $S$ from $Y$. For example, in a graph where the subgroup covariate shift assumption holds and selection depends on $X$ and $A$, a Bayes-optimal model fit in the selected population using only $X$ generalizes to all subgroups in the full population. However, following the results described previously, average performance will not be stable under selection, generally, or across subgroups in the selected or full population.

For cases the graph structure does not imply that a model learned in the selected population is subgroup Bayes-optimal in the full target population, we reason about whether sufficiency is implied by reasoning about whether each of the ratios in the following identity are equal to 1:

$$\underbrace{\frac{P(Y \mid R, A)}{P(Y \mid R)}}_{Y \perp A \mid R} = \underbrace{\frac{P(Y \mid S = 1, R, A)}{P(Y \mid S = 1, R)}}_{Y \perp A \mid R, S=1} * \underbrace{\frac{P(S = 1 \mid R, A)}{P(S = 1 \mid R)}}_{S \perp A \mid R} * \underbrace{\frac{P(S = 1 \mid Y, R)}{P(S = 1 \mid Y, R, A)}}_{S \perp A \mid Y, R} \tag{2}$$

In other words, if the model satisfies sufficiency in the selected population ($Y \perp A \mid R$) and the selection node is d-separated from $A$ given $R$ and $\{R, Y\}$, then the model satisfies sufficiency in the full population. This allows us to, for example, claim that a population Bayes-optimal model satisfies sufficiency in the subgroup covariate shift graph if selection depends only on $Y$, such that the model is miscalibrated to the same extent for all subgroups. We note that equation (2) permits a re-ordering that allows for reasoning about a different set of conditional independencies:

$$\underbrace{\frac{P(Y \mid R, A)}{P(Y \mid R)}}_{Y \perp A \mid R} = \underbrace{\frac{P(Y \mid S = 1, R, A)}{P(Y \mid S = 1, R)}}_{Y \perp A \mid R, S=1} * \underbrace{\frac{P(S = 1 \mid Y, R)}{P(S = 1 \mid R)}}_{S \perp Y \mid R} * \underbrace{\frac{P(S = 1 \mid R, A)}{P(S = 1 \mid Y, R, A)}}_{S \perp Y \mid R, A} \tag{3}$$

For reasoning about separation, we use analogous logic and reason about the terms in the following two identities:

$$\underbrace{\frac{P(R \mid Y, A)}{P(R \mid Y)}}_{R \perp A \mid Y} = \underbrace{\frac{P(R \mid Y, A, S = 1)}{P(R \mid Y, S = 1)}}_{R \perp A \mid Y, S=1} * \underbrace{\frac{P(S = 1 \mid Y, A)}{P(S = 1 \mid Y)}}_{S \perp A \mid Y} * \underbrace{\frac{P(S = 1 \mid Y, R)}{P(S = 1 \mid Y, R, A)}}_{S \perp A \mid Y, R}, \tag{4}$$

and

$$\underbrace{\frac{P(R \mid Y, A)}{P(R \mid Y)}}_{R \perp A \mid Y} = \underbrace{\frac{P(R \mid Y, A, S = 1)}{P(R \mid Y, S = 1)}}_{R \perp A \mid Y, S=1} * \underbrace{\frac{P(S = 1 \mid Y, R)}{P(S = 1 \mid Y)}}_{S \perp R \mid Y} * \underbrace{\frac{P(S = 1 \mid Y, A)}{P(S = 1 \mid Y, R, A)}}_{S \perp R \mid Y, A}. \tag{5}$$

## A.3 Simulation study

We conduct a simulation study to verify the properties studied in this work. We construct data generating processes satisfying each of the settings studied in this work. The data generating processes are provided in supplementary section A.3.1. For each data generating process, we sample 70,000 samples independent samples and use 50,000 for training and 20,000 as a held-out testing dataset for evaluation.

All model fitting and evaluation procedures are repeated and conducted separately for cases where prediction of $Y$ is conducted with (1) $X$ alone, (2) $X$ and an additional categorical covariate indicating subgroup membership $A$, and (3) a set of models using $X$ alone fit separately for each subgroup. For model fitting, we use the scikit-learn version [53] 1.6.1 implementation of gradient boosting classification trees (specifically, `HistGradientBoostingClassifier`) with stratified five-fold cross-validation, with a hyperparameter grid over the maximum number of leaf nodes in $\{10, 25, 50\}$, refitting the model over the training data using the

hyperparameter setting with the minimum average log-loss over the held-out cross-validation folds. The refit model is then used to make predictions on the held-out testing data.

For the experiment involving selection, we modify the complex causal shift graph with three selection mechanisms (see supplementary section A.3.1), and for each data generating process, sample 50,000 samples conditioned on $S = 1$ for training and 20,000 samples from the full population (*i.e.*, without selection). We repeat the same training procedure described above.

To get estimates of $P(A \mid V)$ for use in weighted estimation of model performance, we fit models to predict $A$ from $V$. When $V$ is $X$ or $Y$, the fitting procedure is identical to that used for fitting the models for $Y$, in that we conduct cross-validation with gradient boosting trees on the training data and make predictions with the resulting model on the testing dataset. For cases where $V = R$, we instead conduct a nested cross-validation procedure using only the testing data, similar to cross-fitting [54]. Here, we use an outer stratified five-fold cross-validation partition of the testing data, which, for each outer fold, conducts an inner stratified five-fold cross-validation procedure that returns a model used to make held-out predictions on the corresponding held-out outer fold. Metrics are then computed on the full test set.

For evaluation, we compute unweighted and weighted performance estimates for the log-loss, area under the receiver operating characteristic curve (AUC-ROC), sensitivity (recall), specificity, precision, and net benefit [51]. We compute sensitivity, specificity, and precision using a threshold of 0.5. For net benefit, we use the parametrization presented in Pfohl et al. [9], where the preference trade-off is encoded by a choice of threshold. We use a threshold of 0.5 for both the classifier decision threshold and the preference trade-off threshold. To generate confidence intervals, we use the percentile bootstrap with 10,000 bootstrap samples of the testing data. For weighted metrics, the un-normalized sample weights are treated as fixed based on the result of the procedure described above and sampled alongside the data elements. The resulting samples are then used for weighted computation of metrics.

To generate calibration curves, we quantile-discretize the range of scores $R$ into ten bins and take the empirical mean of $Y$ for the data in each bin. We compute confidence intervals for each bin separately using the Wilson Score Interval Method [55] with the implementation provided by the Statsmodels package version 0.12.1 [56].

The simulation study was conducted on machines with 32 CPUs and 32 GB of RAM. Computing the bootstrap confidence intervals for each of the settings and metrics was the most significant contributor to the overall run time, taking approximately two hours per setting (*i.e.*, approximately 14 hours for the seven settings considered in the primary analyses).

### A.3.1 Data generating processes

**Causal-direction data generating processes** This description encompasses the covariate shift, outcome shift and complex causal shift settings. We consider $X$ to be univariate, $Y$ to be binary, and $A$ to be binary, taking on a value of 0 or 1. We use a binary latent variable $U$ to encode the relationship between $X$ and $A$. For the covariate shift setting, we set $\mu_0 = -2$, $\mu_1 = 0$, $\gamma_A = 1$, $\beta_{a_0} = \beta_{a_1} = 0.5$, and $\alpha_{a_0} = \alpha_{a_1} = 0$. For the outcome shift setting, we set $\mu_0 = -2$, $\mu_1 = 0$, $\gamma_A = 0$, $\beta_{a_0} = 0.5$, $\beta_{a_1} = -1$, and $\alpha_{a_0} = 0.1$, $\alpha_{a_1} = 0$. For the complex causal shift setting, the settings are identical to the outcome shift case except that $\gamma_A = 1$, which has the effect of introducing a covariate shift. To verify properties in a setting where the subgroup covariates distributions are separable, we further increase the extent of the covariate shift present in the complex causal shift case by setting $\mu_1 = 2$.

$$
\begin{aligned}
U &\sim \text{Bernoulli}\big(0.5\big) \\
X \mid U = 0 &\sim \mathcal{N}(\mu_0, 1) \\
X \mid U = 1 &\sim \mathcal{N}(\mu_1, 1) \\
A \mid U &\sim \gamma_A U + (1 - \gamma_A) * \text{Bernoulli}\big(0.5\big) \\
Y \mid A = 0 &\sim \text{Bernoulli}\Big(\text{logit}^{-1}\big(\beta_{a_0} x + \alpha_{a_0}\big)\Big) \\
Y \mid A = 1 &\sim \text{Bernoulli}\Big(\text{logit}^{-1}\big(\beta_{a_1} x + \alpha_{a_1}\big)\Big)
\end{aligned}
$$

We construct three settings with selection bias through augmentation of the complex causal shift setting. We implement three settings, corresponding to cases where the selection mechanism depends on $X$, $Y$, or $\{Y, A\}$, correspond to $S_X$, $S_Y$, and $S_{YA}$. The selected dataset is constructed by filtering the data to cases where $S = 1$.

$$
\begin{aligned}
S_X &\sim \text{Bernoulli}(-\frac{4}{25}X^2 + 1) \\
S_Y &\sim \text{Bernoulli}(0.8Y + 0.4(1-Y)) \\
S_{YA} \mid A = 0 &\sim \text{Bernoulli}(0.5Y + 0.8(1-Y)) \\
S_{YA} \mid A = 1 &\sim \text{Bernoulli}(0.25Y + 0.8(1-Y))
\end{aligned}
$$

**Anticausal data generating processes** This description encompasses the label shift, presentation shift, and complex anticausal shift settings. We consider $X$ to be univariate, $Y$ to be binary, and $A$ to be binary, taking on a value of 0 or 1. For simplicity, we define this data generating process as having $A$-dependent effects, rather than using a latent variable $U$. For the label shift case, we set $\pi_{Y_0} = 0.5$, $\pi_{Y_1} = 0.1$, $\mu_{A_0Y_0} = -1$, $\mu_{A_0Y_1} = 1$, $\mu_{A_1Y_0} = -1$, $\mu_{A_1Y_1} = 1$. For the presentation shift case, we set $\pi_{Y_0} = 0.5$, $\pi_{Y_1} = 0.5$, $\mu_{A_0Y_0} = 1$, $\mu_{A_0Y_1} = 0$, $\mu_{A_1Y_0} = -1$, $\mu_{A_1Y_1} = 1$. The complex anticausal shift setting uses the same parameters as the presentation shift setting except $\pi_{Y_0} = 0.5$, $\pi_{Y_1} = 0.1$.

$$
\begin{aligned}
A &\sim \text{Bernoulli}\big(0.5\big) \\
Y &\sim \text{Bernoulli}\big(A\pi_{Y_0} + (1-A)\pi_{Y_1}\big) \\
X \mid A, Y &\sim \mathcal{N}(\mu_{AY}, 1)
\end{aligned}
$$

## A.4 Experiments with the American Community Survey (ACS) Public Use Microdata Sample (PUMS)

As described in the main text, we follow Ding et al. [48] to define prediction tasks from ACS PUMS [49]. We use the 'ACSIncome' and 'ACSPublicCoverage' tasks definitions provided by the folktables Python package version 0.0.12 [48]. For all experiments, we use the 5-Year horizon California 'person' files, encompassing census records from 2013-2018. The 'ACSIncome' task definition is to predict whether an individual's income is greater than $\$50,000$ per year for adults of age 18 years or older with an income of at least \$100 that have worked greater than zero hours in the last past twelve months. The 'ACSPublicCoverage' task definition is to predict whether an individual is enrolled in a public health insurance plan, restricted to individuals of age 65 years or younger making less than $\$30,000$ per year. The standard covariates used for the 'ACSIncome' task are age (AGEP), class of worker (COW), educational attainment (SCHL), marital status (MAR), occupation (OCCP), place of birth (POBP), relationship (RELP), usual hours worked per week past 12 month (WKHP), and race (RAC1P). The standard covariates used for the 'ACSPublicCoverage' task are AGEP, SCHL, MAR, sex, disability (DIS), employment status of parents (ESP), citizenship status (CIT), mobility status (MIG), military service (MIL), ancestry (ANC), nativity (NAT), hearing difficulty (DEAR), vision difficulty (DEYE), cognitive difficulty (DREM), income, employment status (ESR), state (ST), gave birth to child within the past 12 months (FER), and RAC1P.

To define subgroups, we create a custom combined race and ethnicity field that combines the RA1CP field with the hispanic origin flag (HISP) and groups rare categories. The combined race/ethnicity field takes on the value "Hispanic" for individuals of Hispanic origin and the value of RA1CP field otherwise. We then combine the American Indian and Alaska Native with the Native Hawaiian and Pacific Islander into a group called "Other". For modeling, we remove RAC1P from the covariate set, and use the combined race/ethnicity field for subgroup-aware prediction.

For modeling and evaluation, we replicate the procedure used in the simulation study, using gradient boosting trees for classification with the same cross-validation and hyperparameter selection procedure. As is standard for the prediction tasks proposed by Ding et al. [48], we do not directly use the person

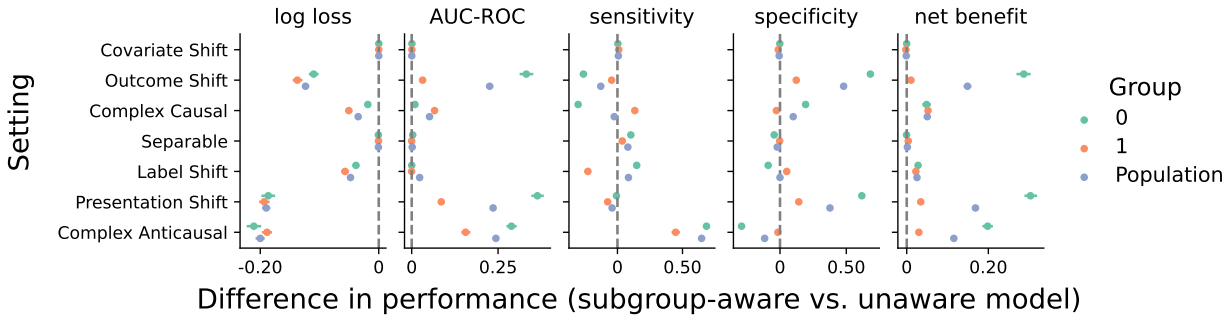
weights provided in the ACS PUMS data, which implies that the derived estimates and models may not be representative of the underlying populations.

As in the case of the simulation study, we conduct these experiments using machines with 32 CPUs and 32 GB of RAM and computation of the bootstrap confidence intervals, was the most significant contributor to the overall run time, taking approximately 14 hours per setting (*i.e.*, approximately 28 hours for the two tasks considered).

# B Supplementary figures and tables

**Supplementary Table B1: Properties of Bayes-optimal models under selection**. ✓ indicates cases where Bayes-optimal prediction in the selected population $P(\cdot \mid S = 1)$ is sufficient to induce the listed property in the full population $P(\cdot)$. "Other $Y \not\perp A \mid X$" indicate the outcome shift, presentation shift, and complex causal and anticausal shift graphs.

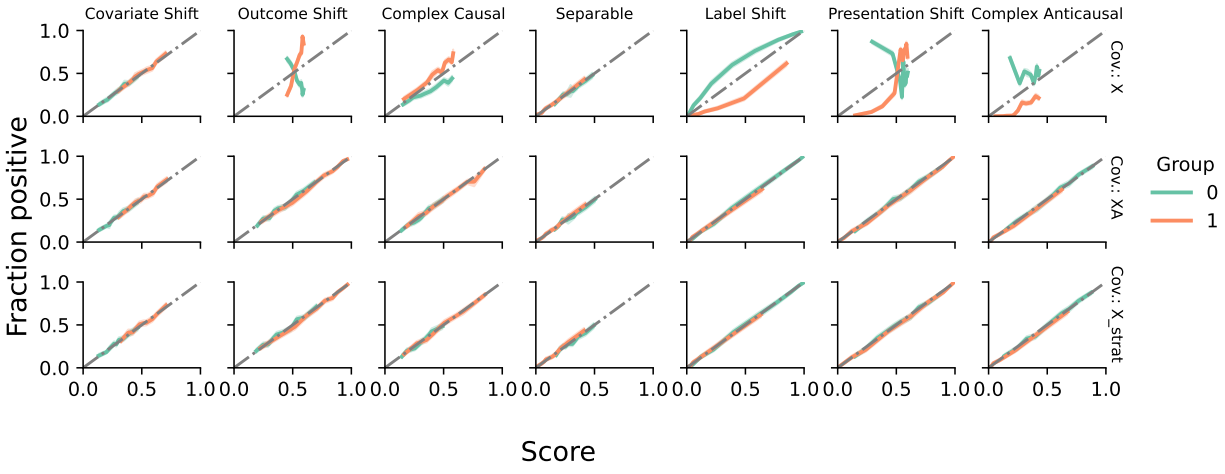
| | Sufficiency | | Subgroup calibration | | Separation | |
|---|---|---|---|---|---|---|
| Setting | $Z = X$ | $Z = \{X, A\}$ | $Z = X$ | $Z = \{X, A\}$ | $Z = X$ | $Z = \{X, A\}$ |
| Covariate shift | | | | | | |
| $X \to S$ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ |
| $A \to S$ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ |
| $\{X, A\} \to S$ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ |
| $Y \to S$ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| $\{X, Y\} \to S$ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| $\{A, Y\} \to S$ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| $\{X, Y, A\} \to S$ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Label shift | | | | | | |
| $X \to S$ | ✗ | ✓ | ✗ | ✓ | ✓ | ✗ |
| $A \to S$ | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ |
| $\{X, A\} \to S$ | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ |
| $Y \to S$ | ✗ | ✓ | ✗ | ✗ | ✓ | ✗ |
| $\{X, Y\} \to S$ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ |
| $\{A, Y\} \to S$ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| $\{X, Y, A\} \to S$ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Other $Y \not\perp A \mid X$ | | | | | | |
| $X \to S$ | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ |
| $A \to S$ | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ |
| $\{X, A\} \to S$ | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ |
| $Y \to S$ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ |
| $\{X, Y\} \to S$ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| $\{A, Y\} \to S$ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| $\{X, Y, A\} \to S$ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |



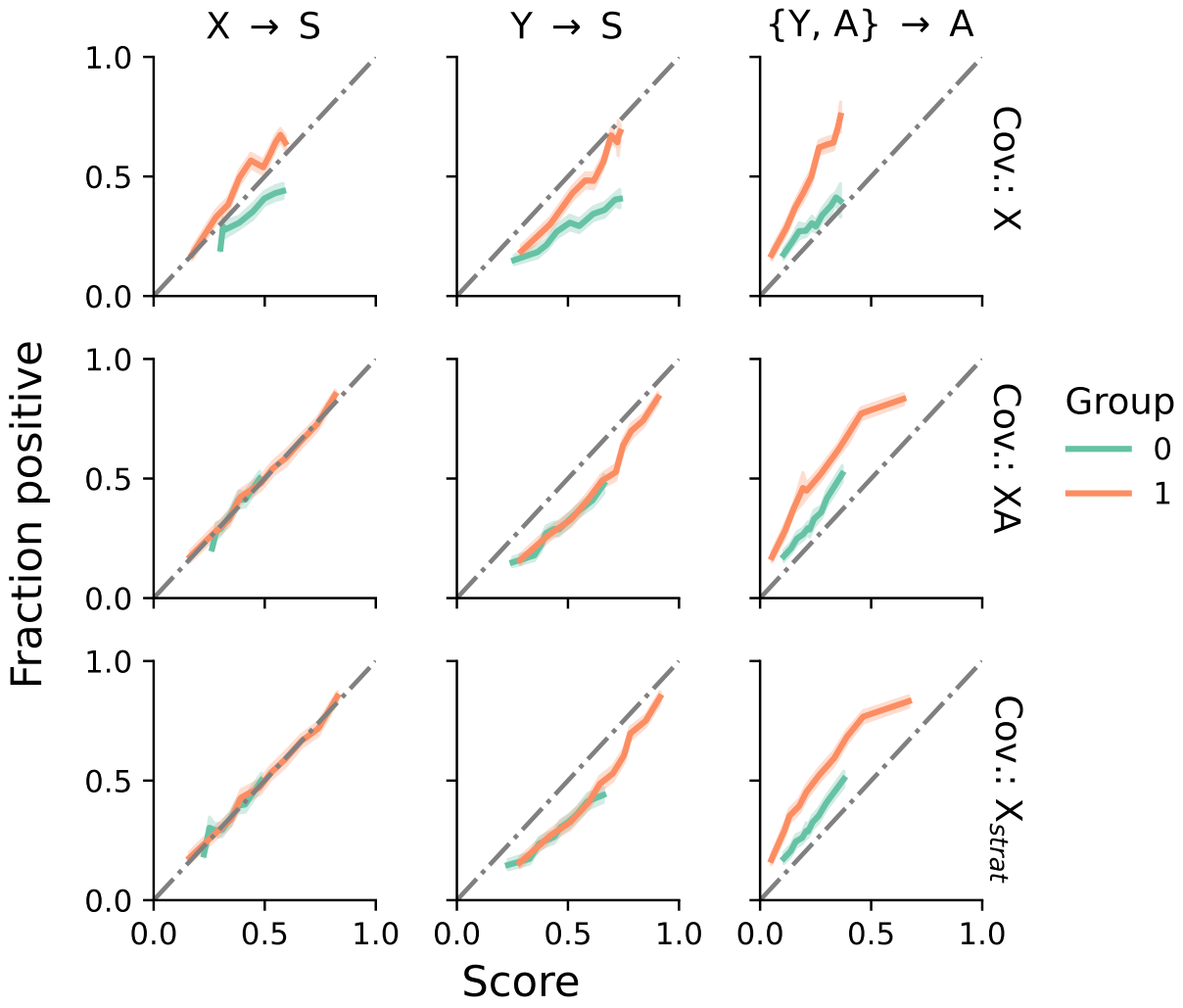

**Supplementary Figure B1: Simulation study: the effect of subgroup-aware prediction on model performance**. We report the difference in performance between models that have access to subgroup membership as an additional covariate as compared to those that do not. Plotted are average differences with 95% confidence intervals for each setting and for several performance metrics.

**Supplementary Table B2: ACS PUMS cohort characteristics.** Shown are the number of individuals and the prevalence of the label for each race/ethnicity subgroup for the two tasks derived from the ACS PUMS data.

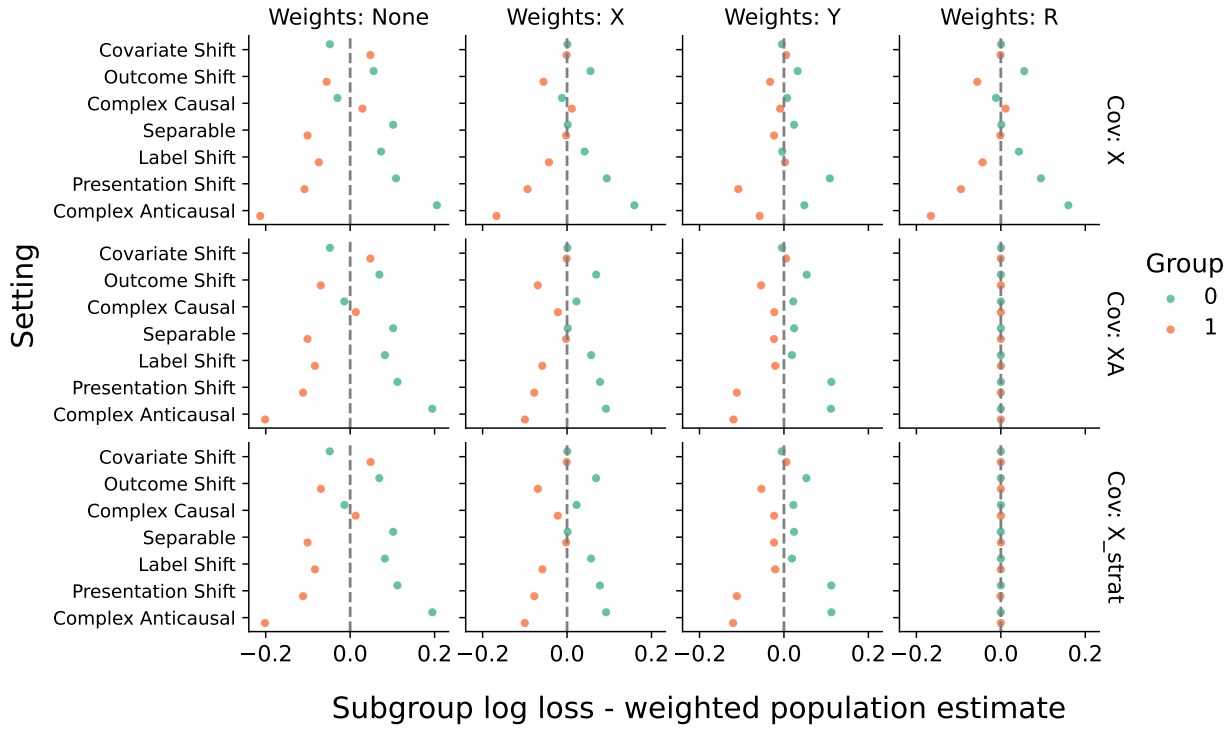
| | ACSIncome | | ACSPublicCoverage | |
|---|---|---|---|---|
| | Count | Prevalence | Count | Prevalence |
| Asian | 151,163 | 0.449 | 104,642 | 0.286 |
| Black | 40,764 | 0.327 | 41,090 | 0.507 |
| Hispanic | 313,007 | 0.200 | 317,534 | 0.393 |
| Multiracial | 23,781 | 0.374 | 20,734 | 0.323 |
| Other | 8,910 | 0.304 | 8,590 | 0.397 |
| White | 412,572 | 0.504 | 236,842 | 0.299 |



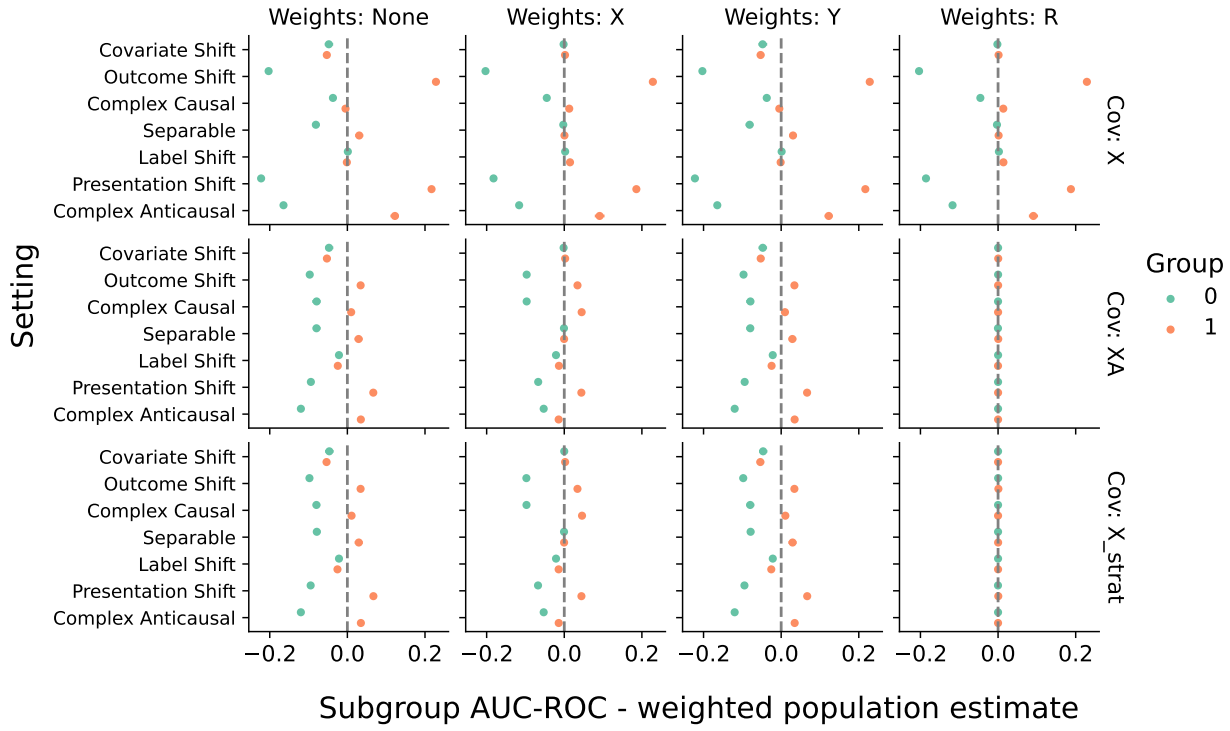

**Supplementary Figure B2: Simulation study: calibration curves**. Plotted are calibration curves for each subgroup with 95% confidence intervals. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.



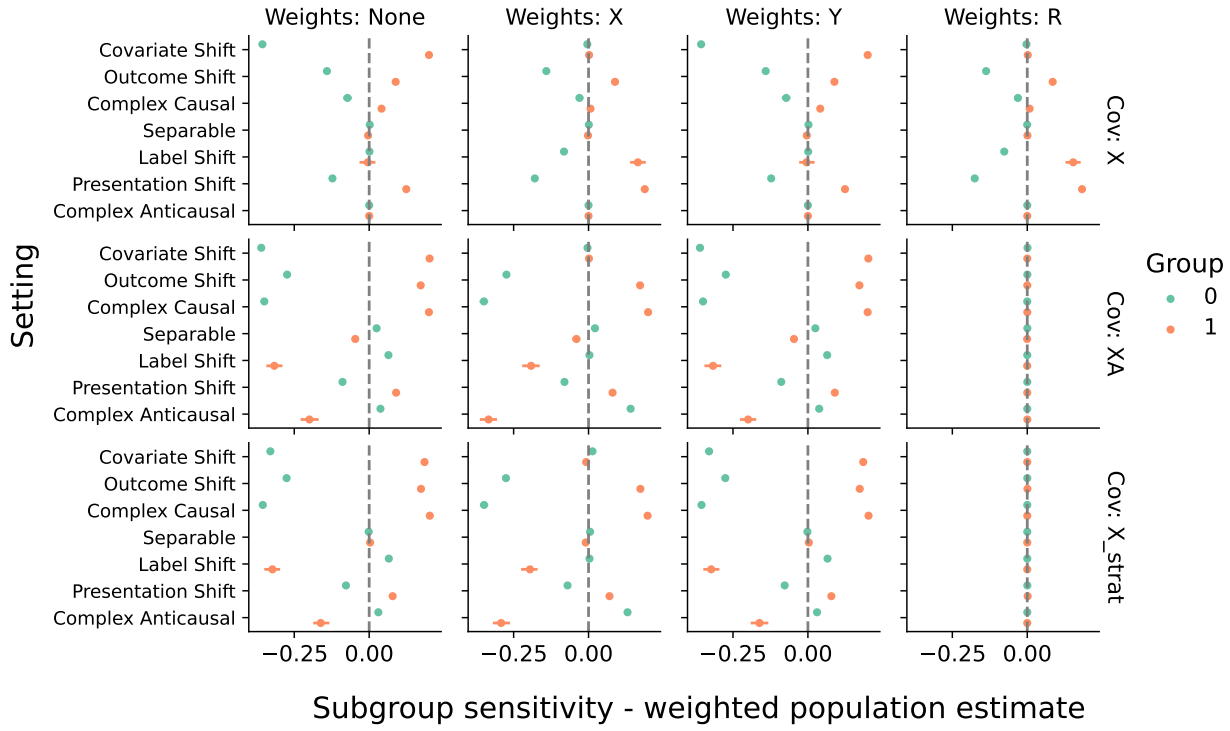

**Supplementary Figure B3: Simulation study: calibration with selection bias**. Plotted are calibration curves for each subgroup with 95% confidence intervals. Models are fit in the selected population and evaluated in the full population without selection. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.



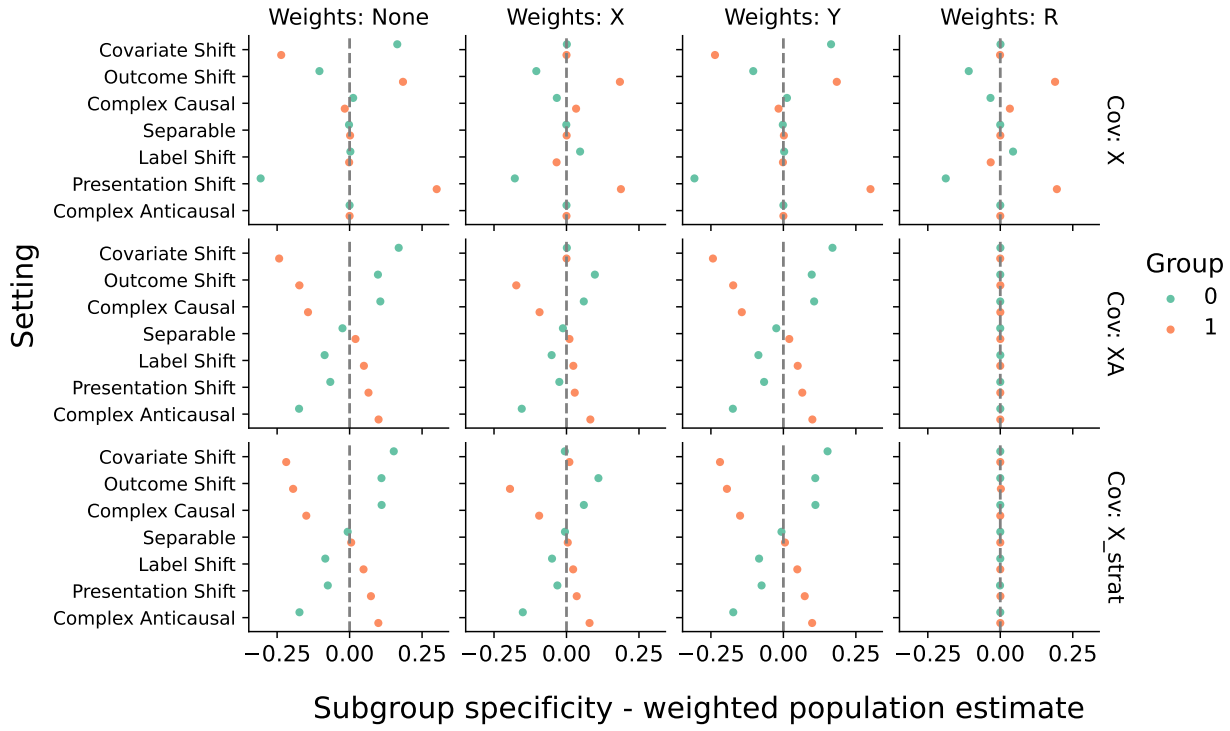

**Supplementary Figure B4: Simulation study: controlled evaluation of log loss**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between the unweighted disaggregated performance with the population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.



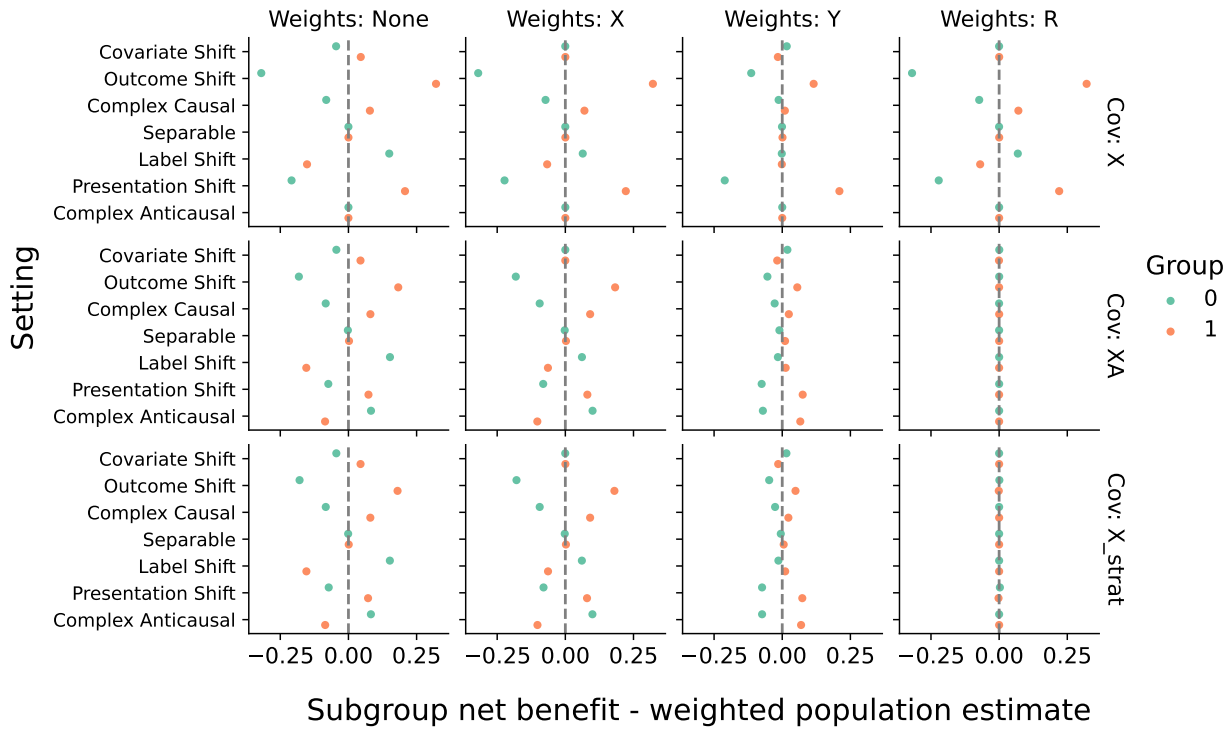

**Supplementary Figure B5: Simulation study: controlled evaluation of AUC-ROC**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between the unweighted disaggregated performance with the population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.



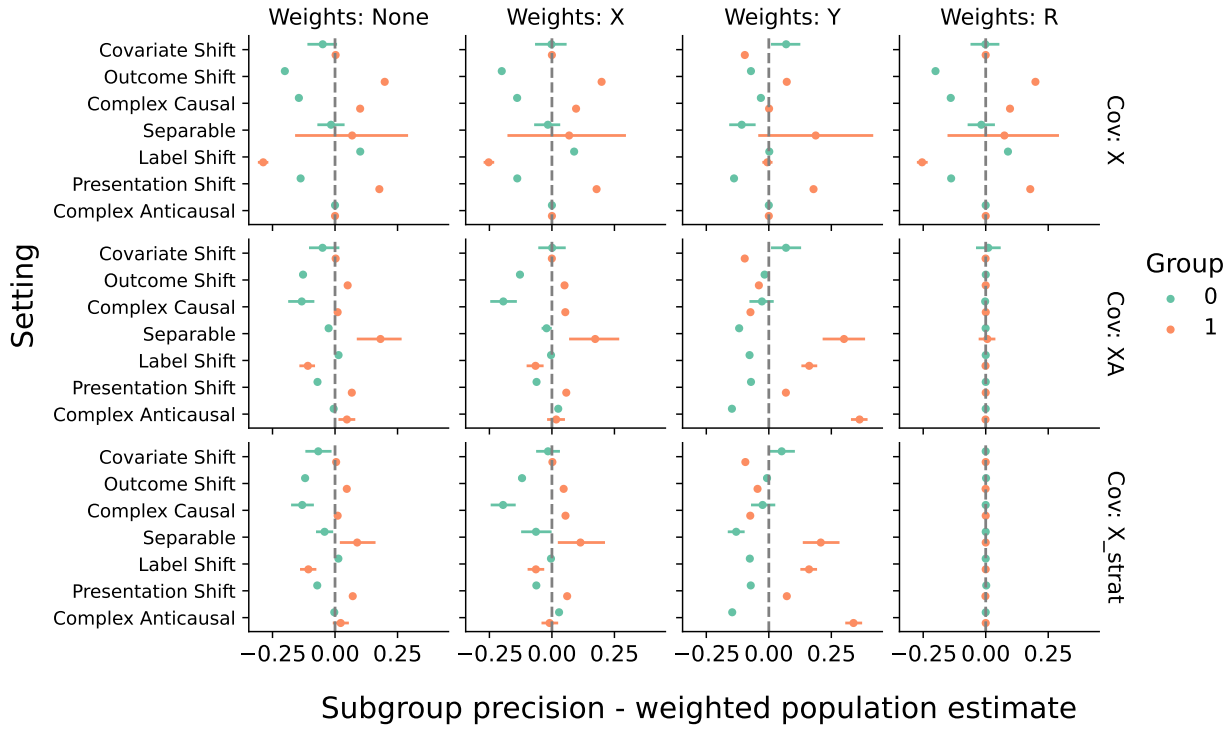

**Supplementary Figure B6: Simulation study: controlled evaluation of sensitivity**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between the unweighted disaggregated performance with the population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.



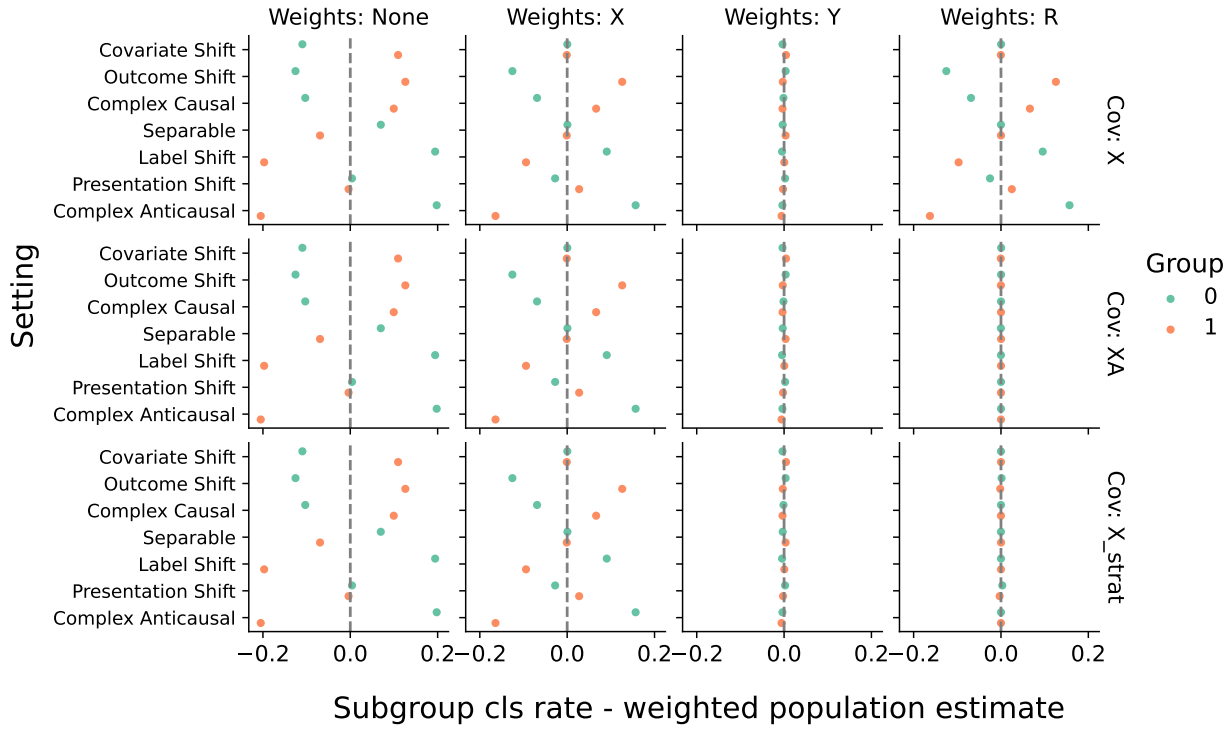

**Supplementary Figure B7: Simulation study: controlled evaluation of specificity**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between the unweighted disaggregated performance with the population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.



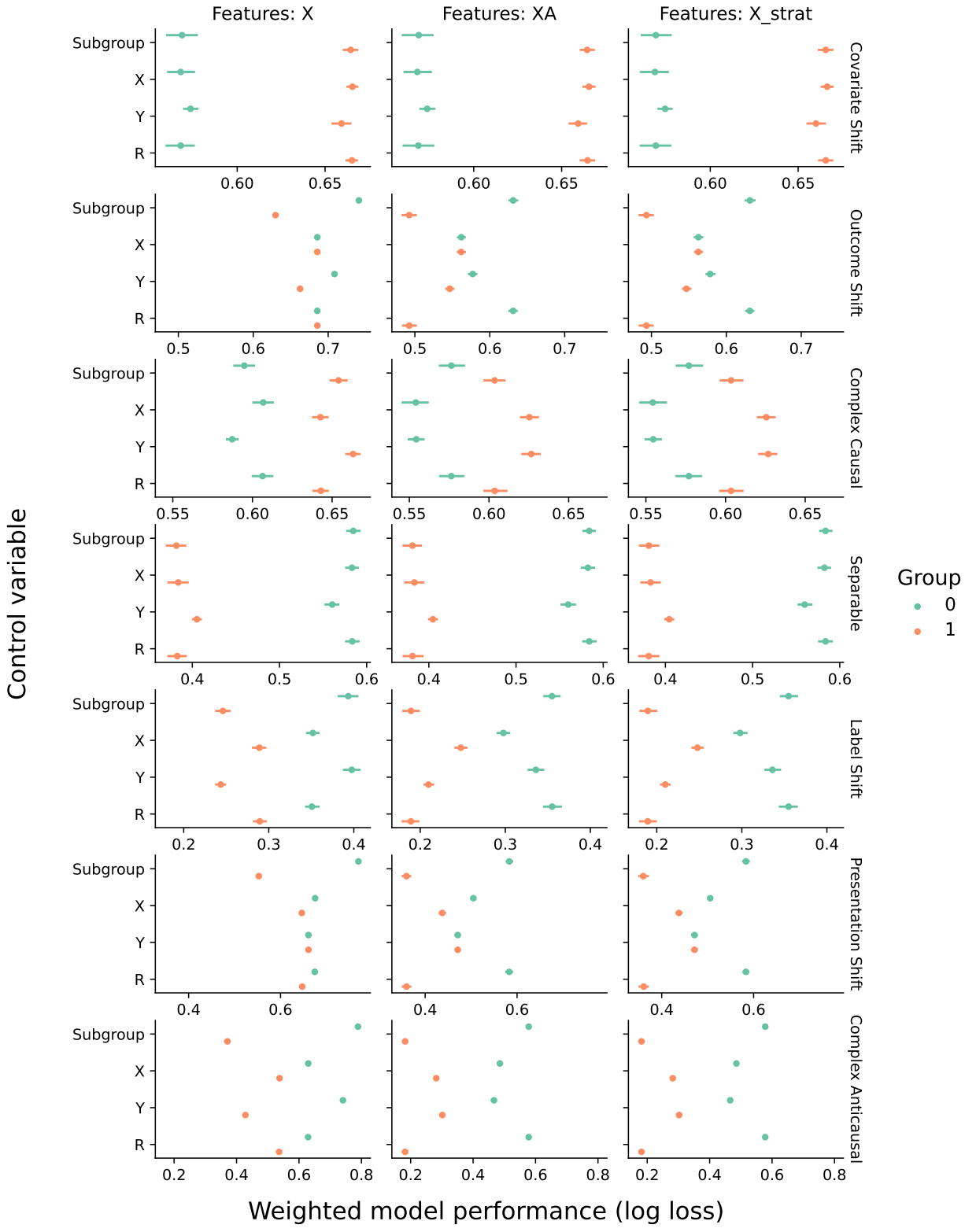

**Supplementary Figure B8: Simulation study: controlled evaluation of net benefit**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between the unweighted disaggregated performance with the population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.



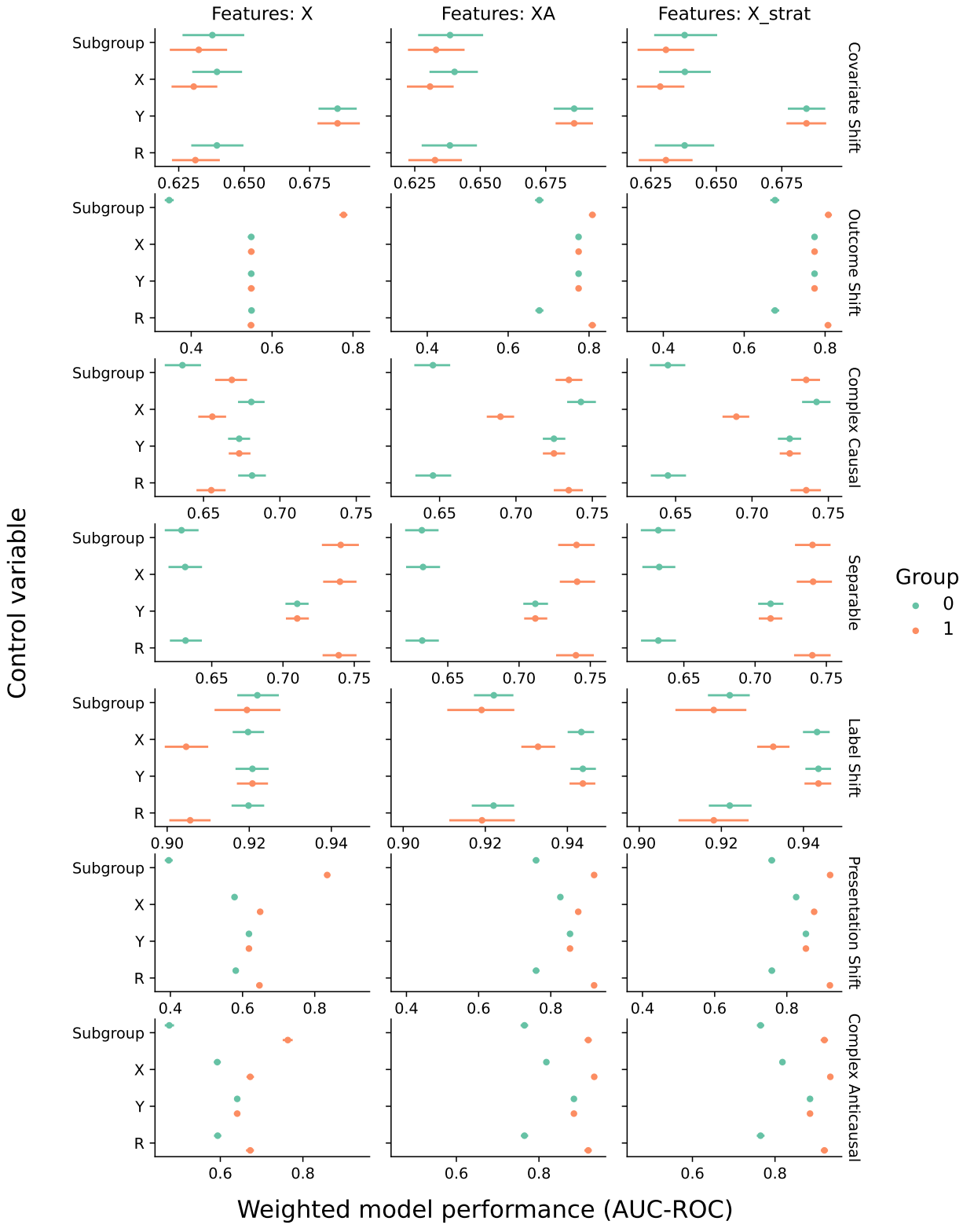

**Supplementary Figure B9: Simulation study: controlled evaluation of precision**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between the unweighted disaggregated performance with the population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.



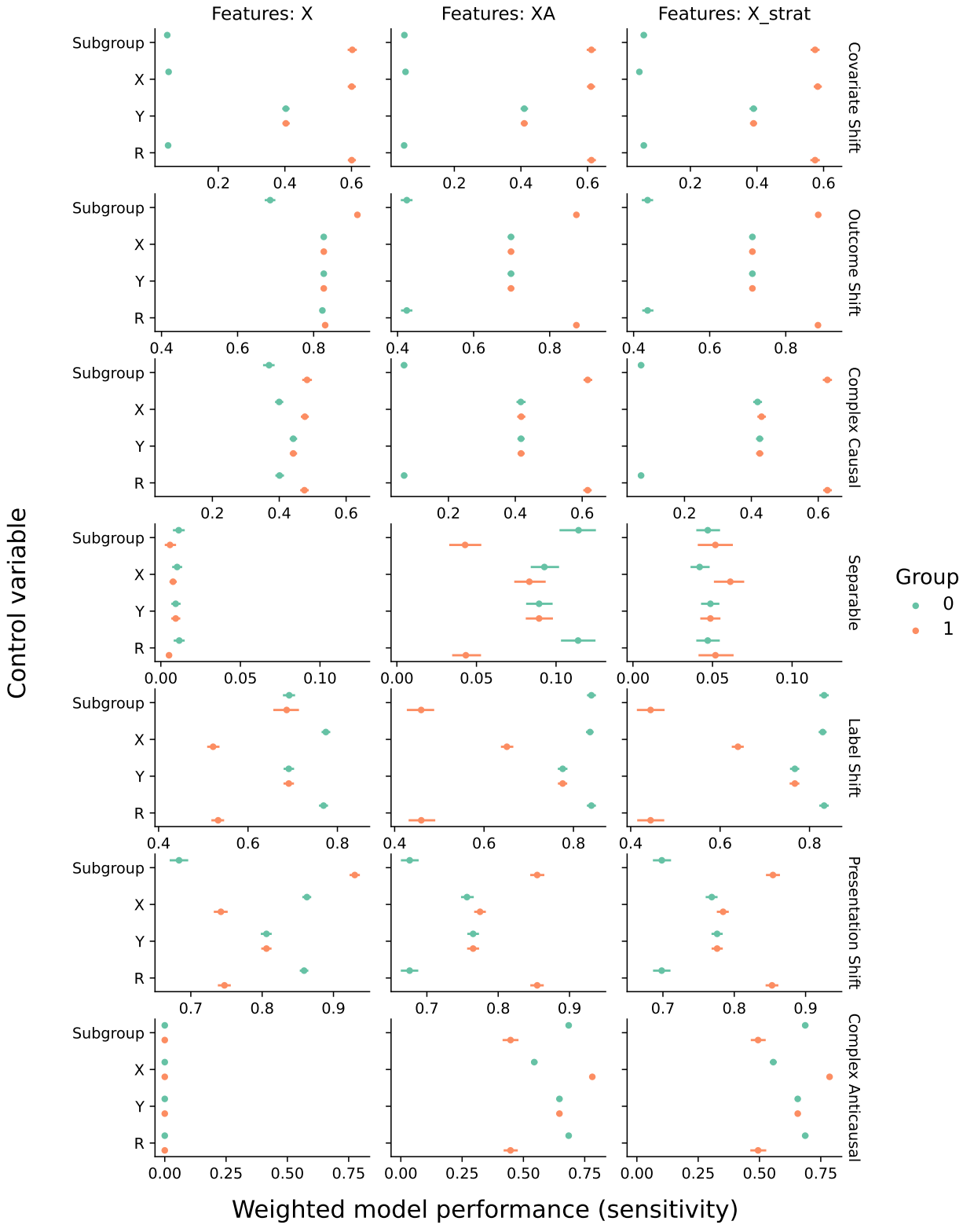

**Supplementary Figure B10: Simulation study: controlled evaluation of classification rate**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between the unweighted disaggregated performance with the population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.



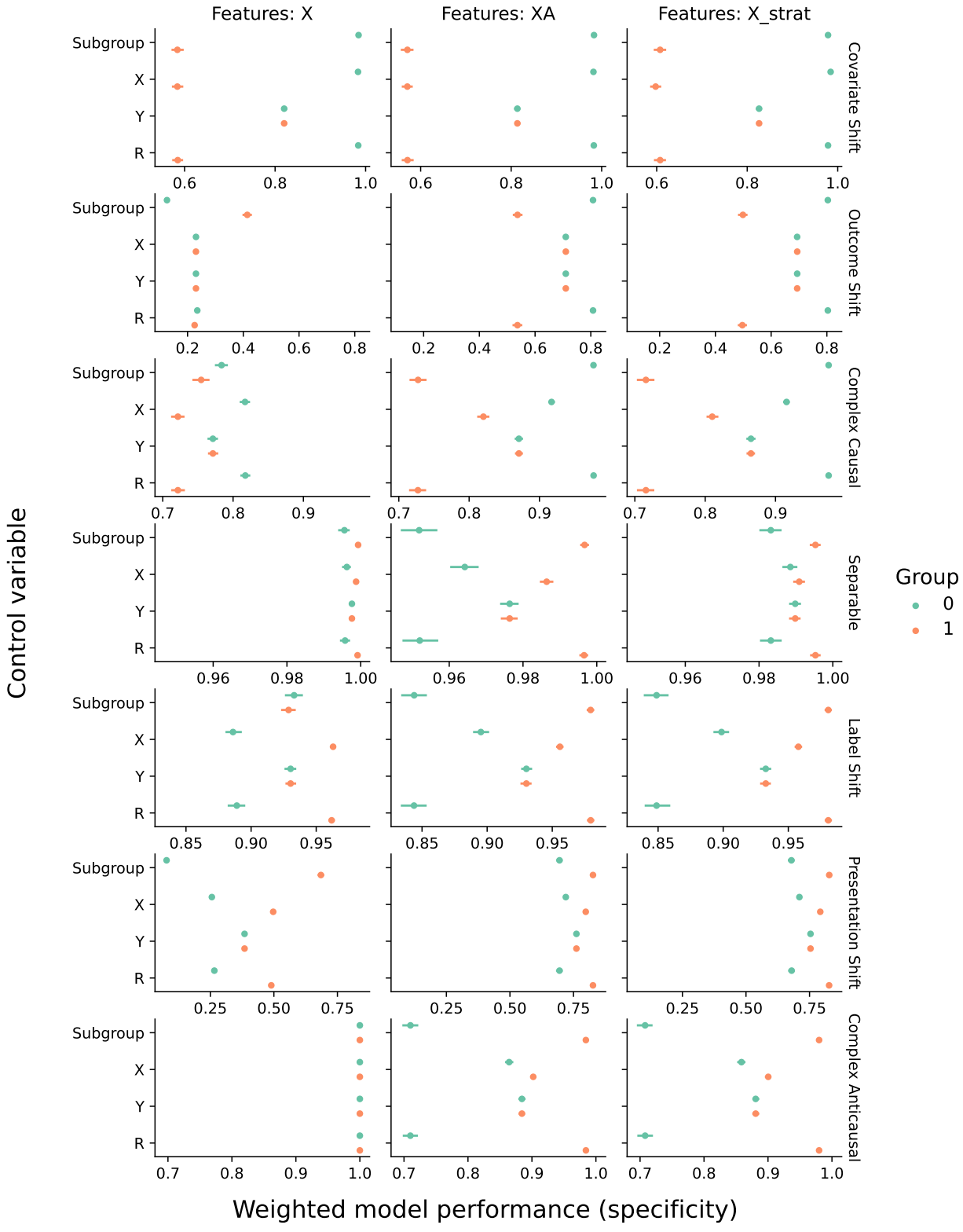

**Supplementary Figure B11: Simulation study: weighted estimation of log loss**. Plotted are the weighted estimates of performance $M_a$ with 95% confidence intervals, corresponding to weighted estimates of population performance weighted to match the distribution of $X$, $Y$, or $R$ for each subgroups. The entry labeled "subgroup" corresponds to the unweighted estimate of subgroup performance. The first column corresponds to subgroup-agnostic prediction, the second column to prediction with $A$ as an additional covariate, and the third column to stratified prediction by $A$.



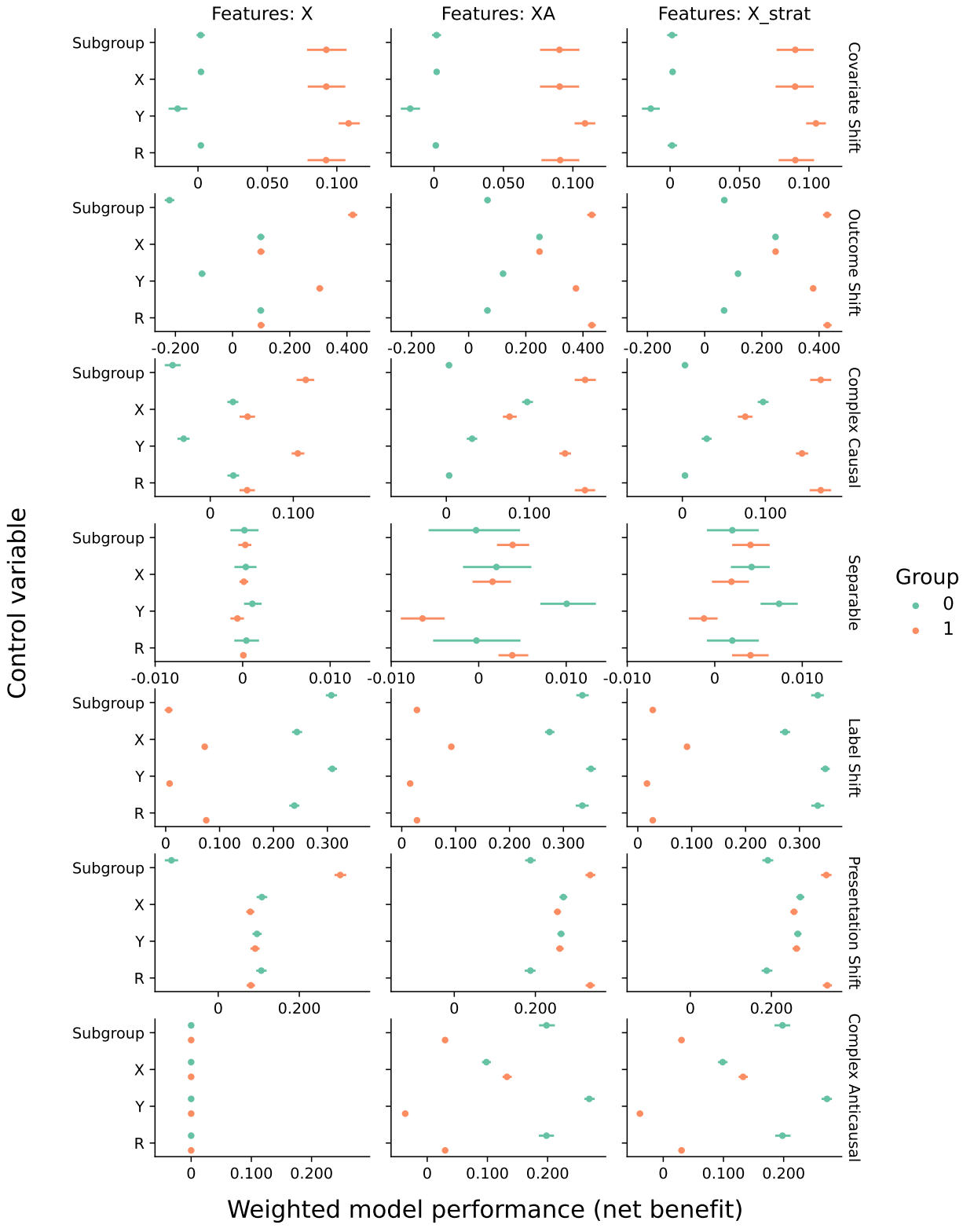

**Supplementary Figure B12: Simulation study: weighted estimation of AUC-ROC**. Plotted are the weighted estimates of performance $M_a$ with 95% confidence intervals, corresponding to weighted estimates of population performance weighted to match the distribution of $X$, $Y$, or $R$ for each subgroups. The entry labeled "subgroup" corresponds to the unweighted estimate of subgroup performance. The first column corresponds to subgroup-agnostic prediction, the second column to prediction with $A$ as an additional covariate, and the third column to stratified prediction by $A$.



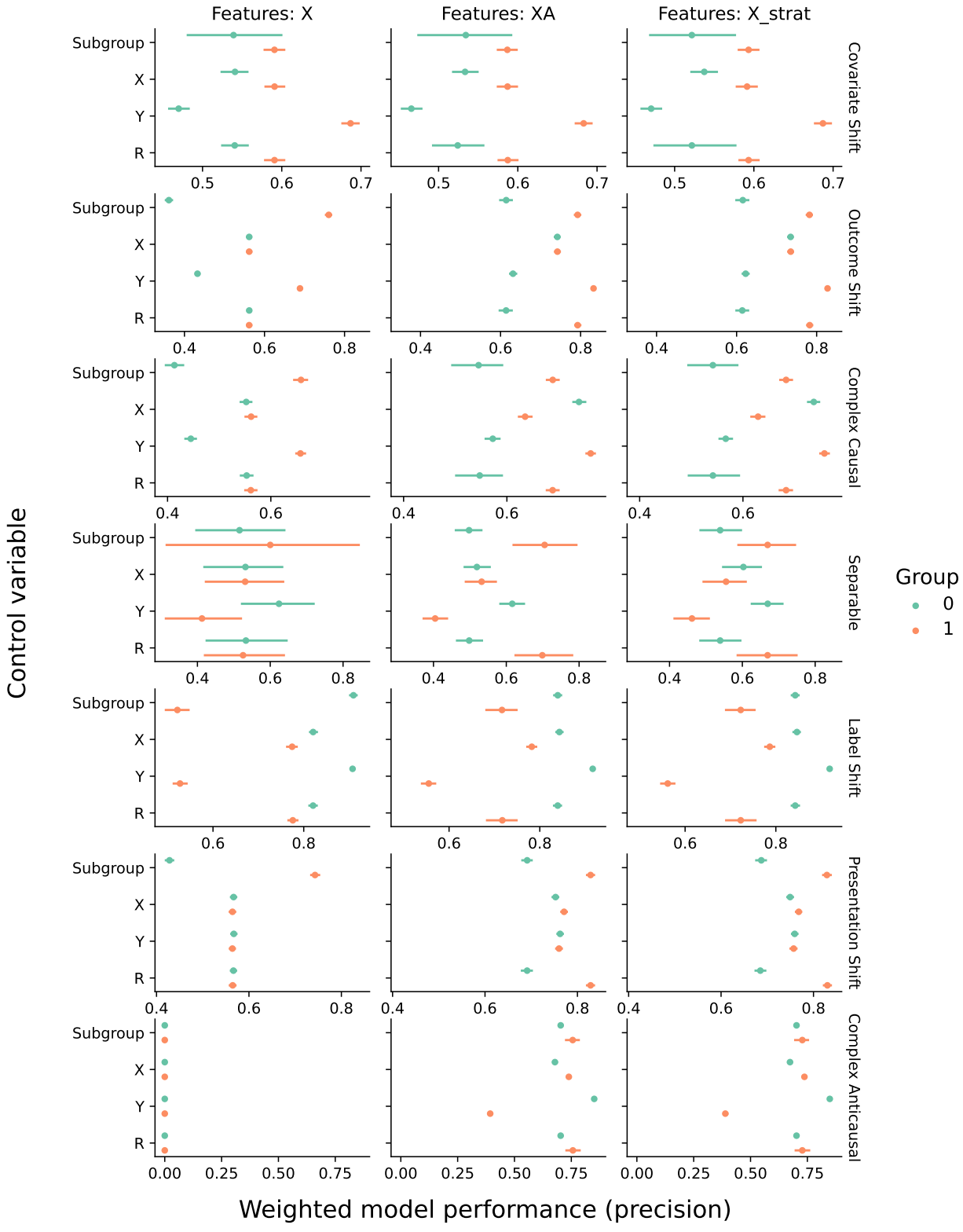

**Supplementary Figure B13: Simulation study: weighted estimation of sensitivity**. Plotted are the weighted estimates of performance $M_a$ with 95% confidence intervals, corresponding to weighted estimates of population performance weighted to match the distribution of $X$, $Y$, or $R$ for each subgroups. The entry labeled "subgroup" corresponds to the unweighted estimate of subgroup performance. The first column corresponds to subgroup-agnostic prediction, the second column to prediction with $A$ as an additional covariate, and the third column to stratified prediction by $A$.



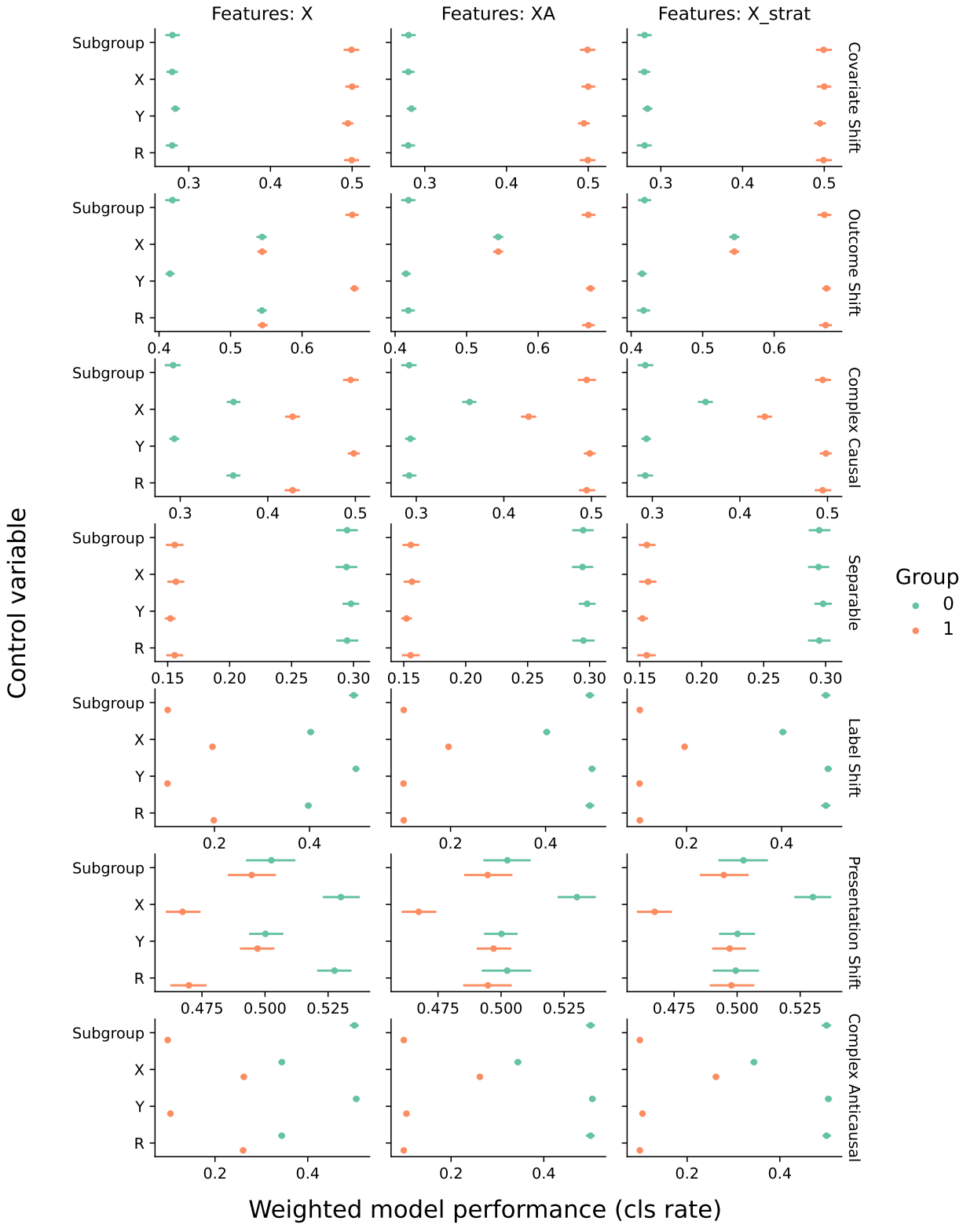

**Supplementary Figure B14: Simulation study: weighted estimation of specificity**. Plotted are the weighted estimates of performance $M_a$ with 95% confidence intervals, corresponding to weighted estimates of population performance weighted to match the distribution of $X$, $Y$, or $R$ for each subgroups. The entry labeled "subgroup" corresponds to the unweighted estimate of subgroup performance. The first column corresponds to subgroup-agnostic prediction, the second column to prediction with $A$ as an additional covariate, and the third column to stratified prediction by $A$.



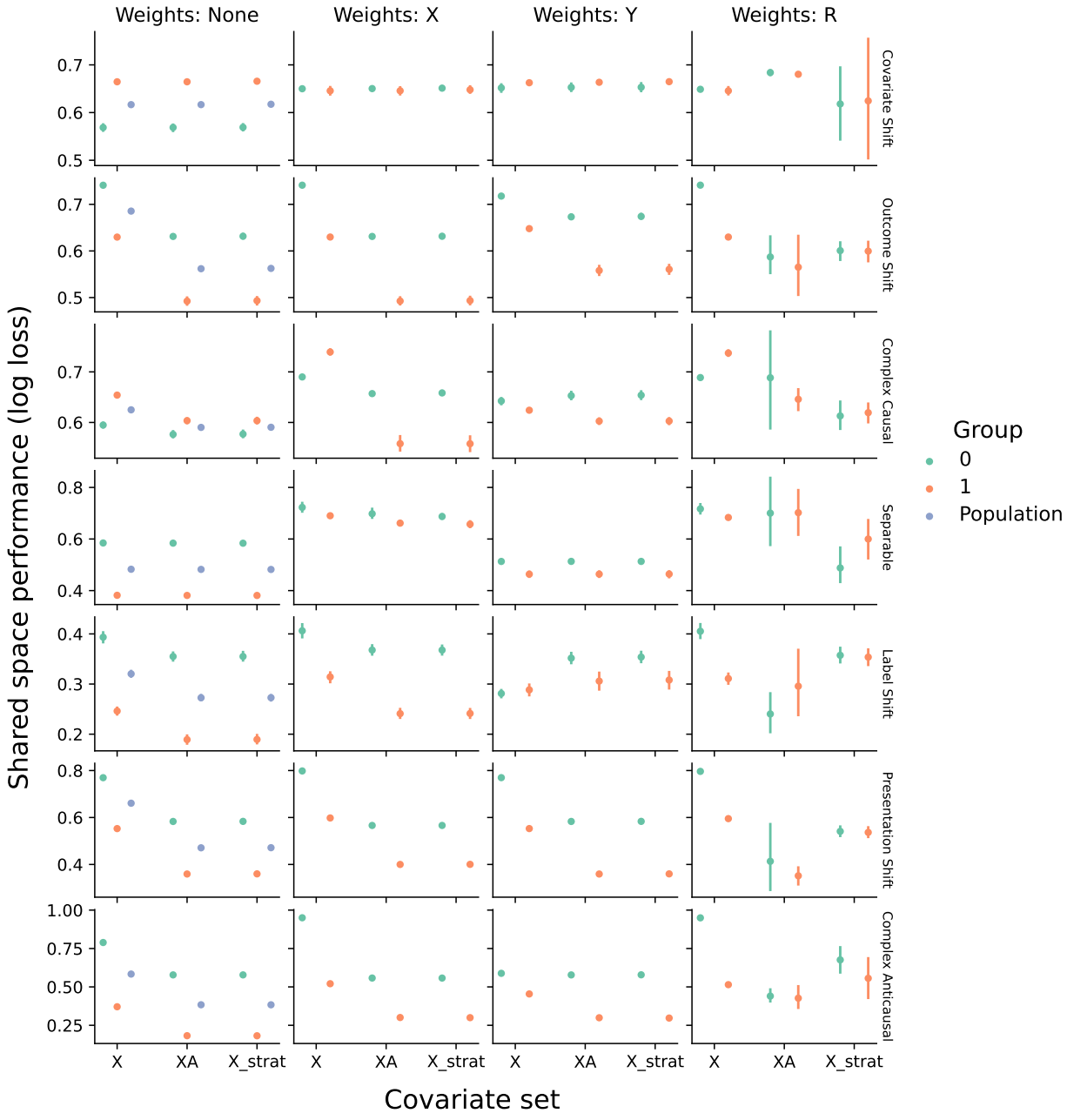

**Supplementary Figure B15: Simulation study: weighted estimation of net benefit**. Plotted are the weighted estimates of performance $M_a$ with 95% confidence intervals, corresponding to weighted estimates of population performance weighted to match the distribution of $X$, $Y$, or $R$ for each subgroups. The entry labeled "subgroup" corresponds to the unweighted estimate of subgroup performance. The first column corresponds to subgroup-agnostic prediction, the second column to prediction with $A$ as an additional covariate, and the third column to stratified prediction by $A$.



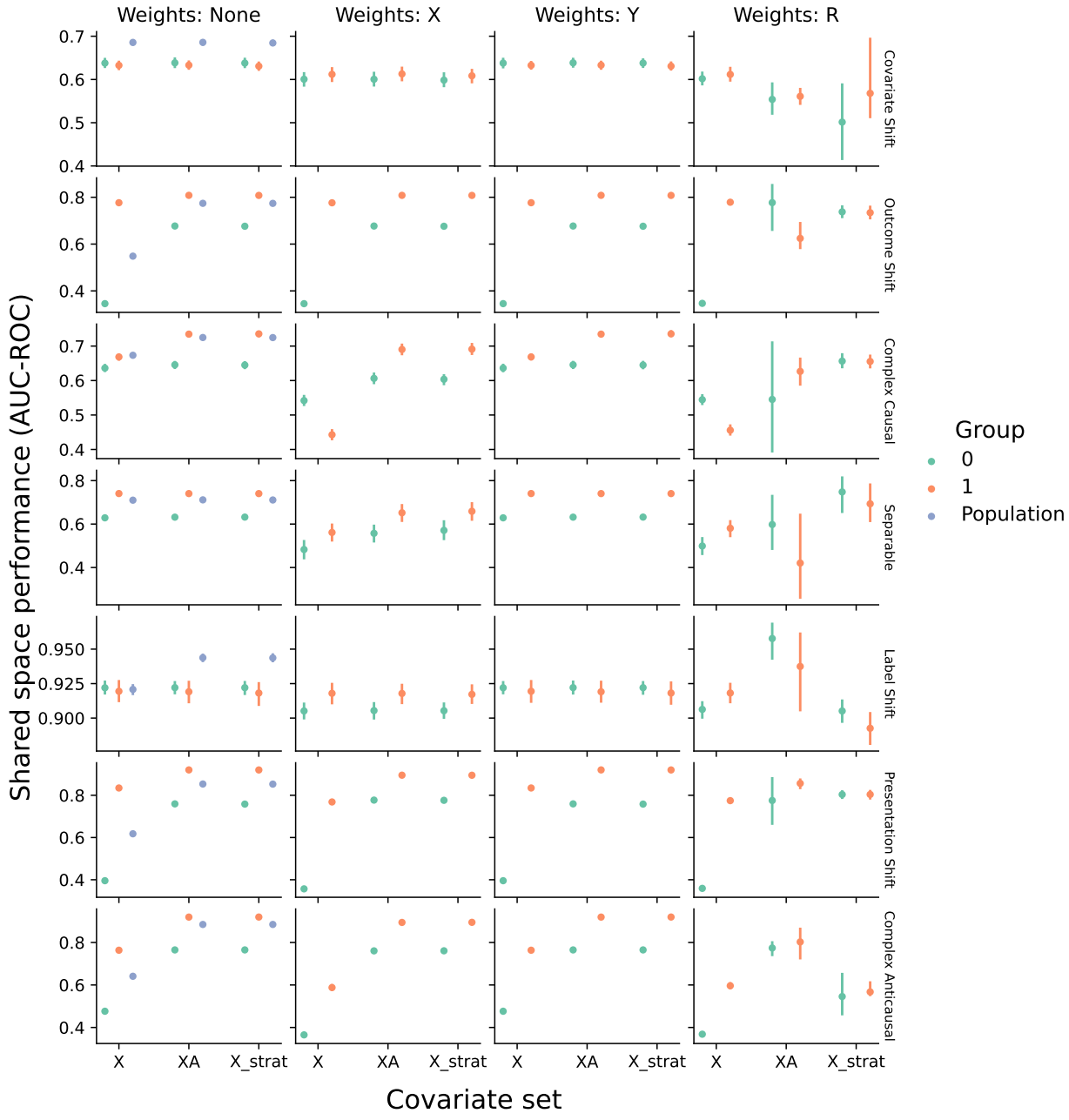

**Supplementary Figure B16: Simulation study: weighted estimation of precision**. Plotted are the weighted estimates of performance $M_a$ with 95% confidence intervals, corresponding to weighted estimates of population performance weighted to match the distribution of $X$, $Y$, or $R$ for each subgroups. The entry labeled "subgroup" corresponds to the unweighted estimate of subgroup performance. The first column corresponds to subgroup-agnostic prediction, the second column to prediction with $A$ as an additional covariate, and the third column to stratified prediction by $A$.



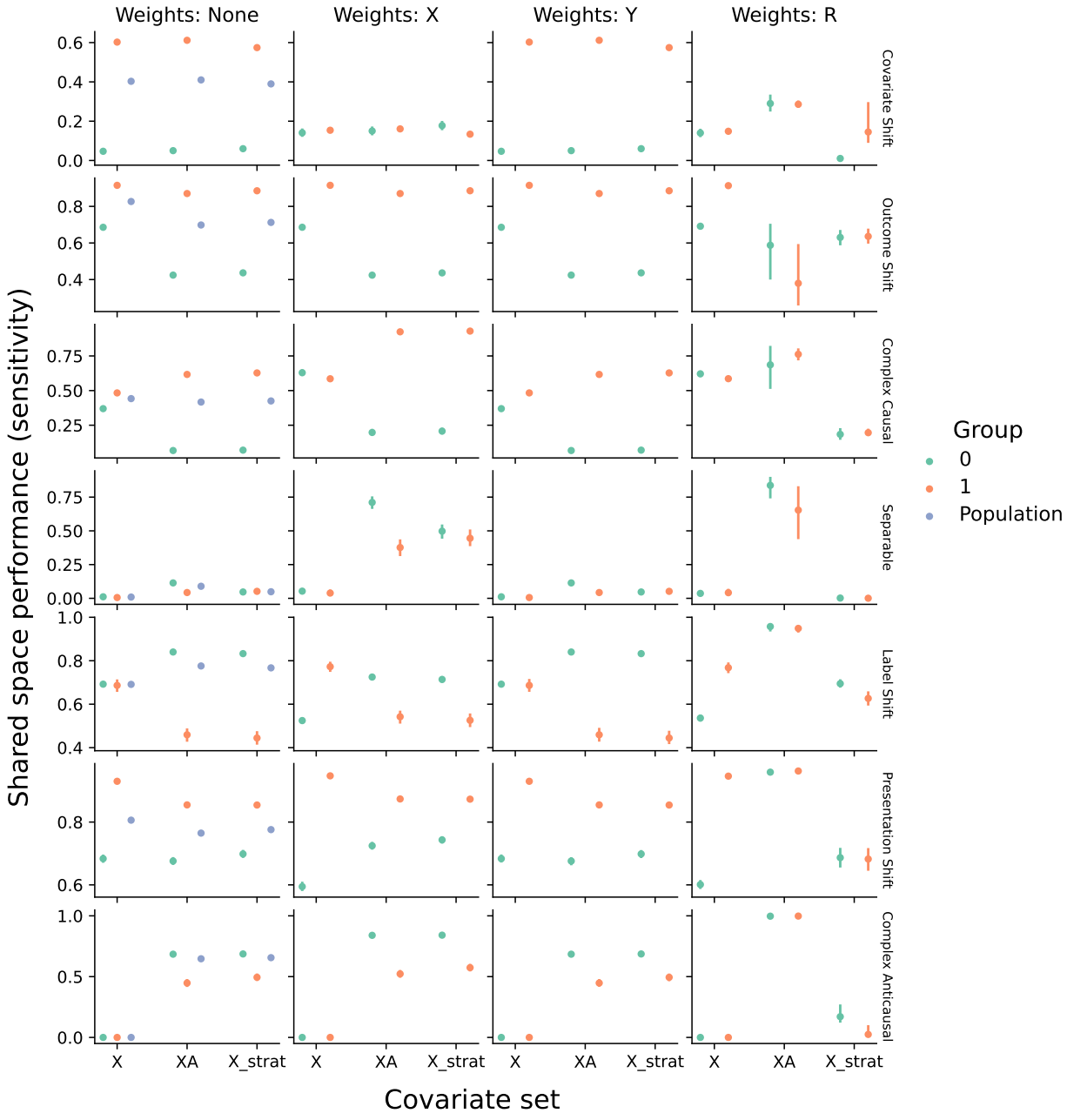

**Supplementary Figure B17: Simulation study: weighted estimation of classification rate**. Plotted are the weighted estimates of performance $M_a$ with 95% confidence intervals, corresponding to weighted estimates of population performance weighted to match the distribution of $X$, $Y$, or $R$ for each subgroups. The entry labeled "subgroup" corresponds to the unweighted estimate of subgroup performance. The first column corresponds to subgroup-agnostic prediction, the second column to prediction with $A$ as an additional covariate, and the third column to stratified prediction by $A$.



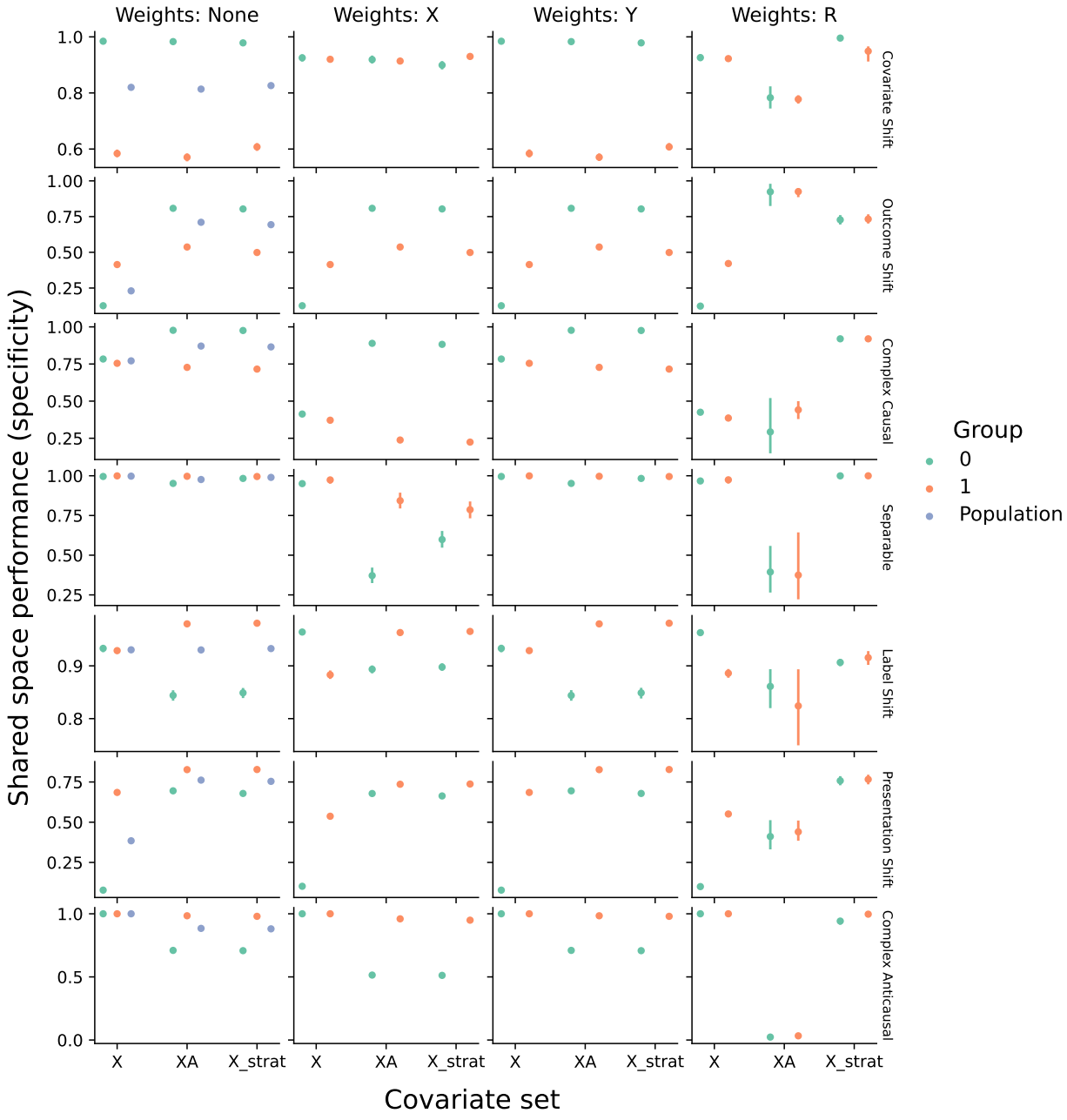

**Supplementary Figure B18: Simulation study: shared space subgroup log loss**. Plotted are the average performance with 95% confidence intervals for subgroup performance following weighting to a shared space, using the approach of Cai et al. [16]. Columns correspond to different conditioning variables used to construct the weights and rows correspond data generating processes.



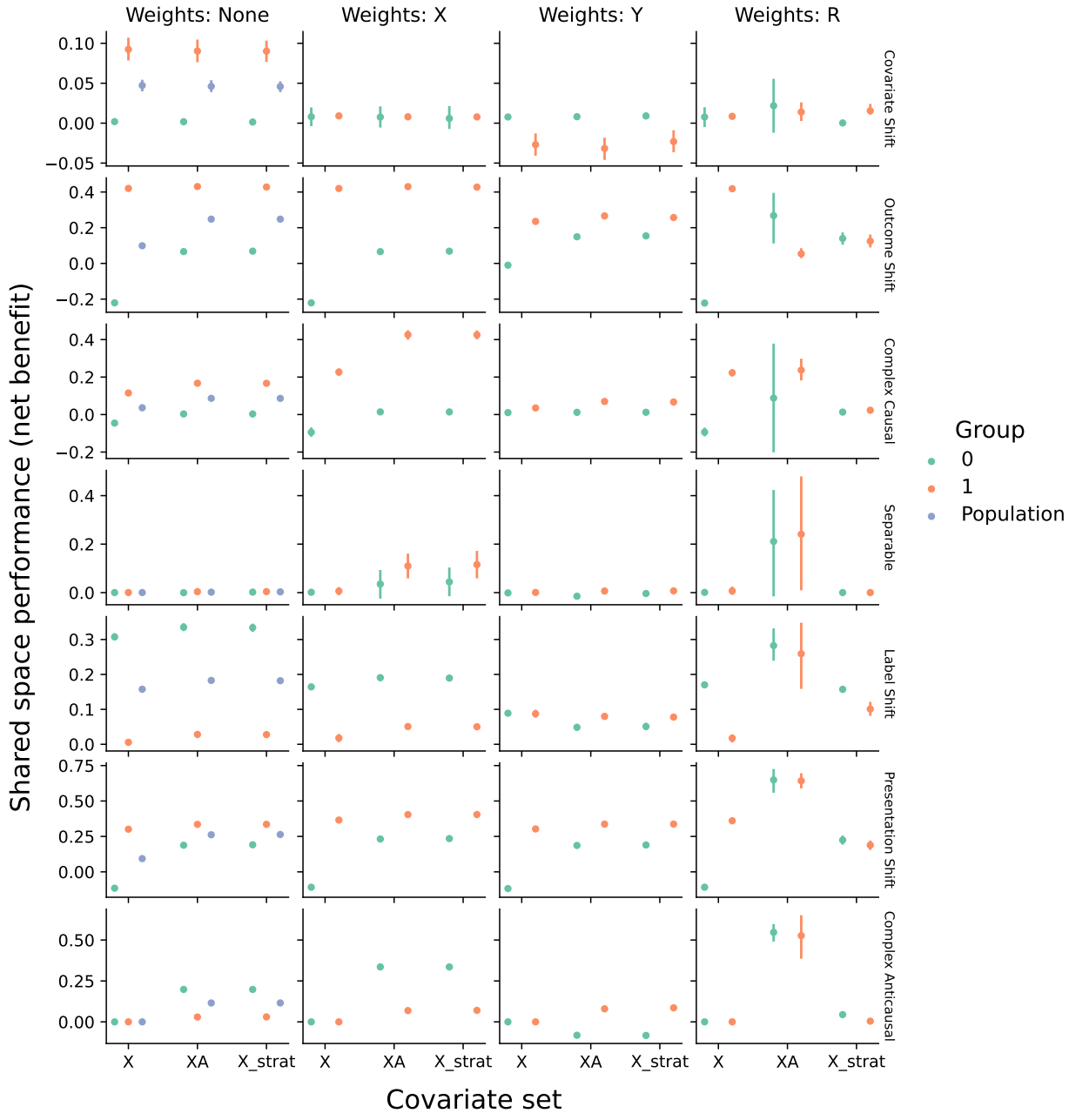

**Supplementary Figure B19: Simulation study: shared space subgroup AUC-ROC**. Plotted are the average performance with 95% confidence intervals for subgroup performance following weighting to a shared space, using the approach of Cai et al. [16]. Columns correspond to different conditioning variables used to construct the weights and rows correspond data generating processes.



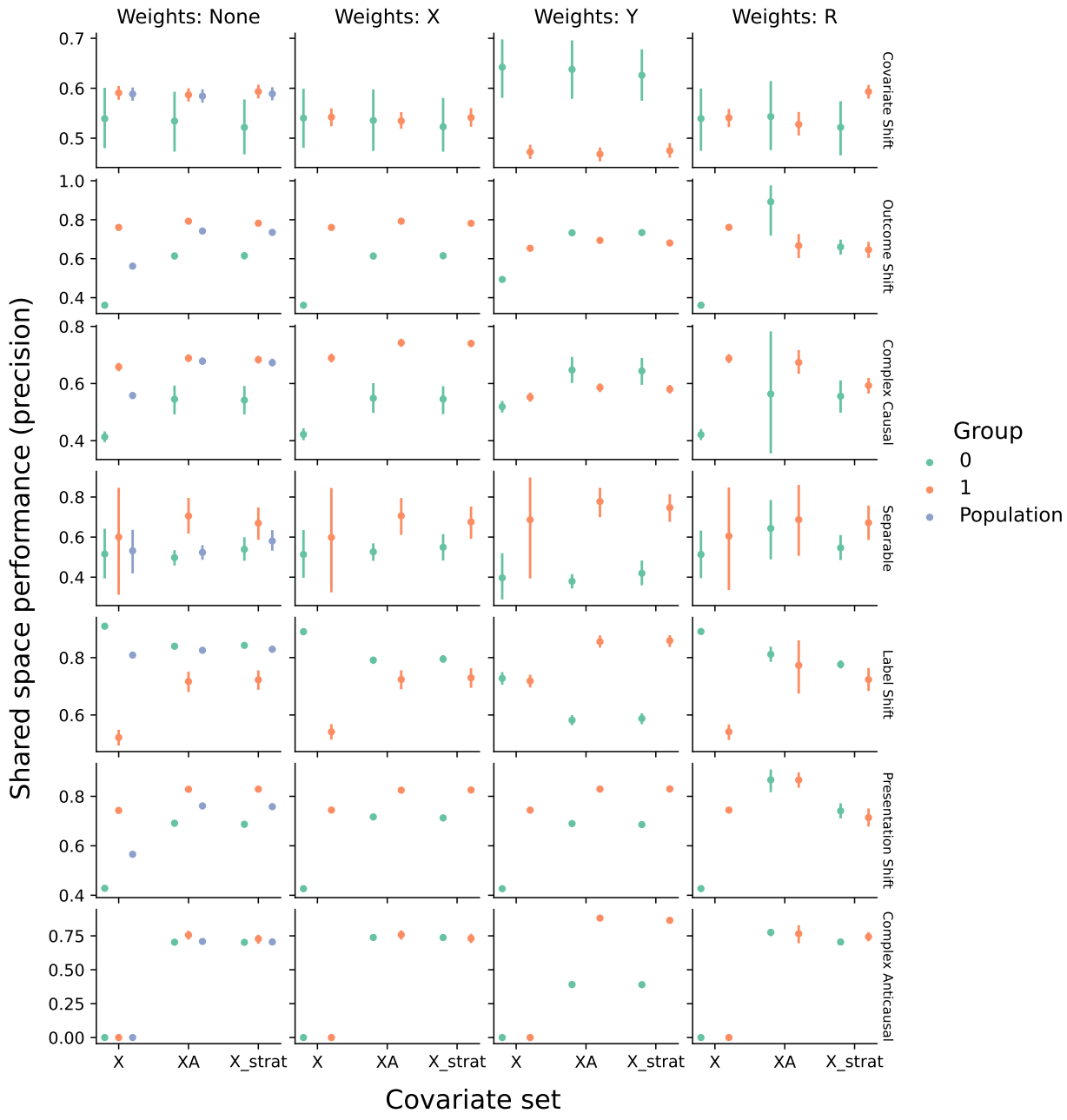

**Supplementary Figure B20: Simulation study: shared space subgroup sensitivity**. Plotted are the average performance with 95% confidence intervals for subgroup performance following weighting to a shared space, using the approach of Cai et al. [16]. Columns correspond to different conditioning variables used to construct the weights and rows correspond data generating processes.



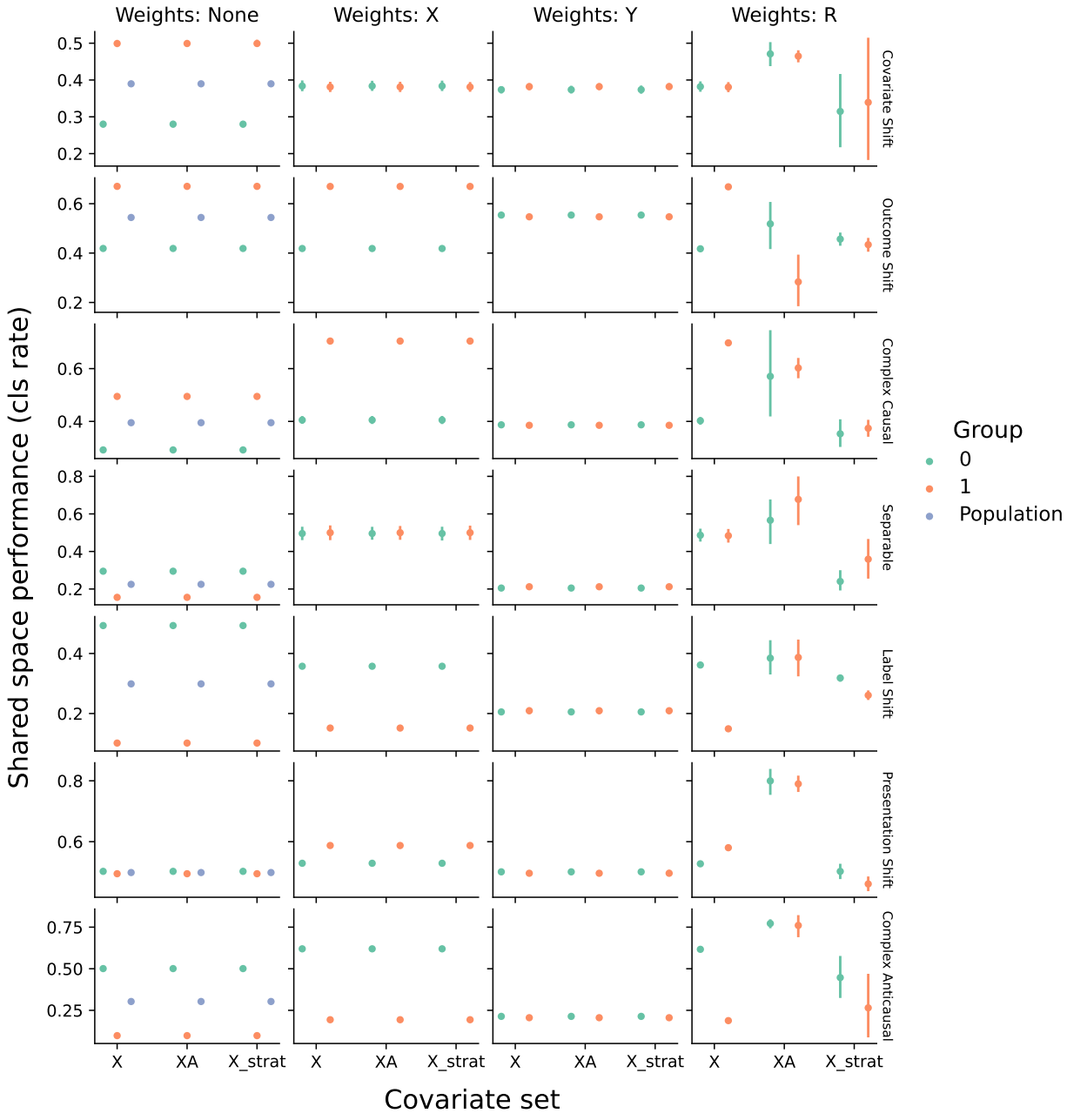

**Supplementary Figure B21: Simulation study: shared space subgroup specificity**. Plotted are the average performance with 95% confidence intervals for subgroup performance following weighting to a shared space, using the approach of Cai et al. [16]. Columns correspond to different conditioning variables used to construct the weights and rows correspond data generating processes.



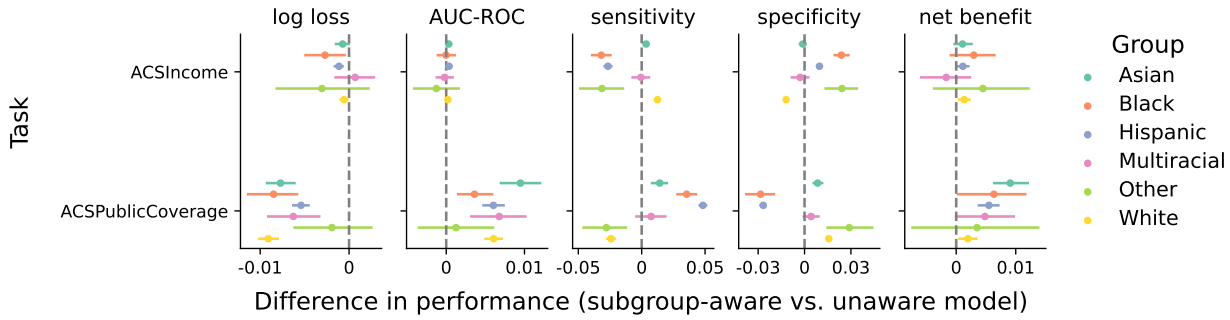

**Supplementary Figure B22: Simulation study: shared space subgroup net benefit**. Plotted are the average performance with 95% confidence intervals for subgroup performance following weighting to a shared space, using the approach of Cai et al. [16]. Columns correspond to different conditioning variables used to construct the weights and rows correspond data generating processes.



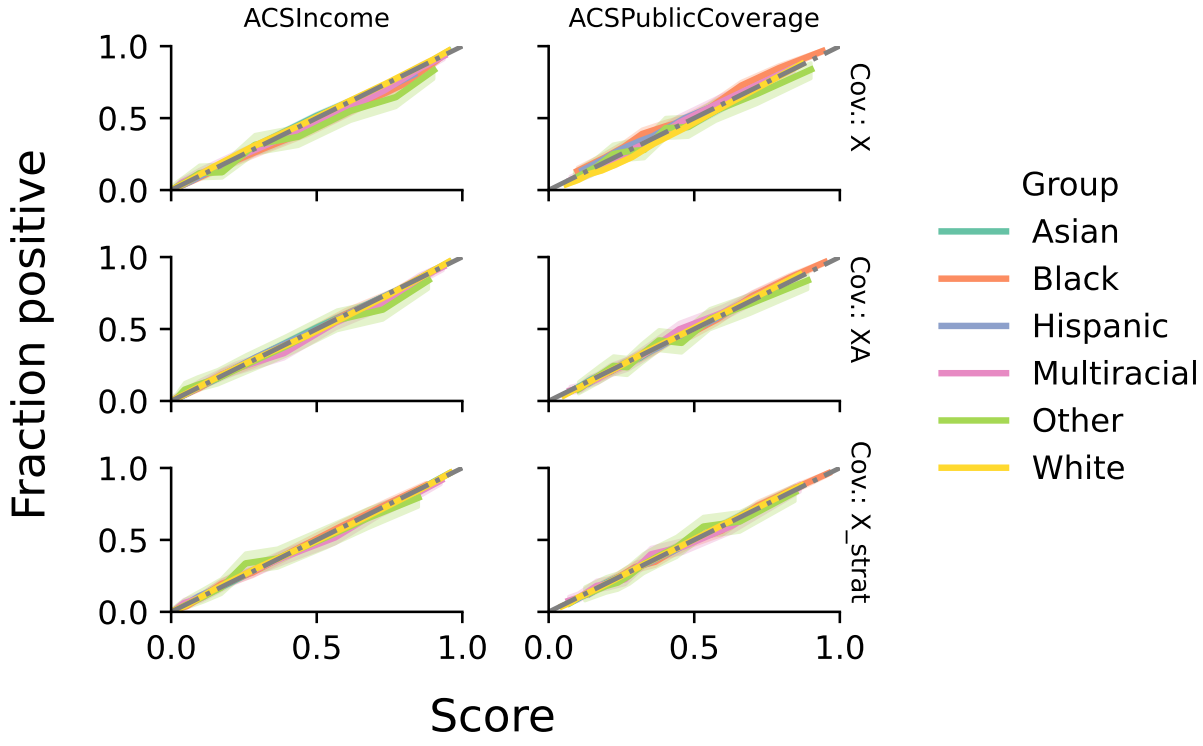

**Supplementary Figure B23: Simulation study: shared space subgroup precision**. Plotted are the average performance with 95% confidence intervals for subgroup performance following weighting to a shared space, using the approach of Cai et al. [16]. Columns correspond to different conditioning variables used to construct the weights and rows correspond data generating processes.



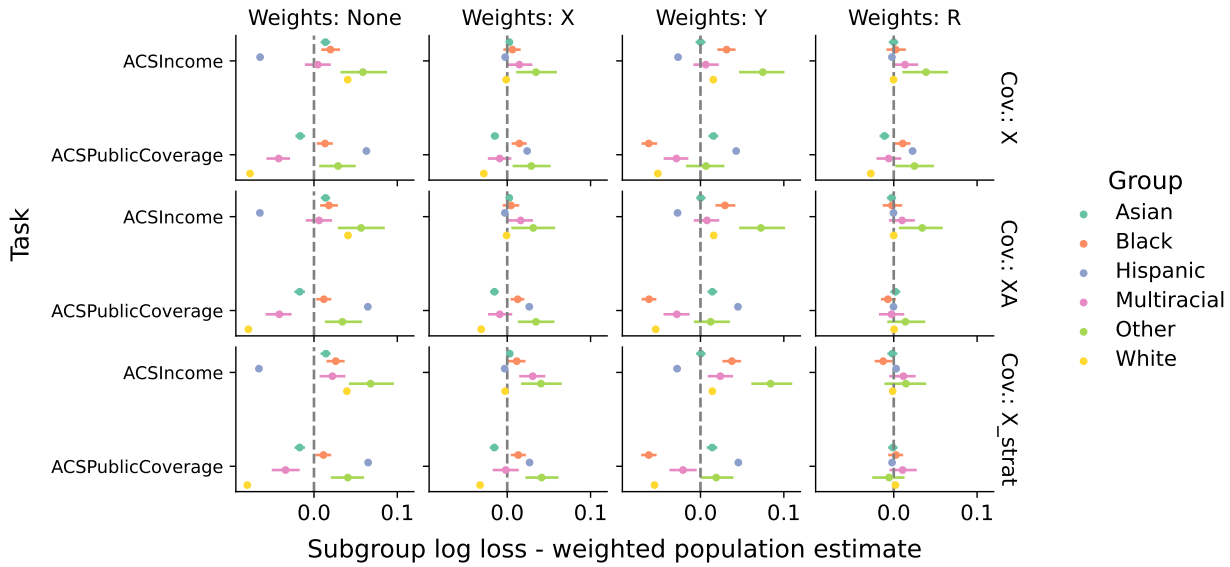

**Supplementary Figure B24: Simulation study: shared space subgroup classification rate**. Plotted are the average performance with 95% confidence intervals for subgroup performance following weighting to a shared space, using the approach of Cai et al. [16]. Columns correspond to different conditioning variables used to construct the weights and rows correspond data generating processes.



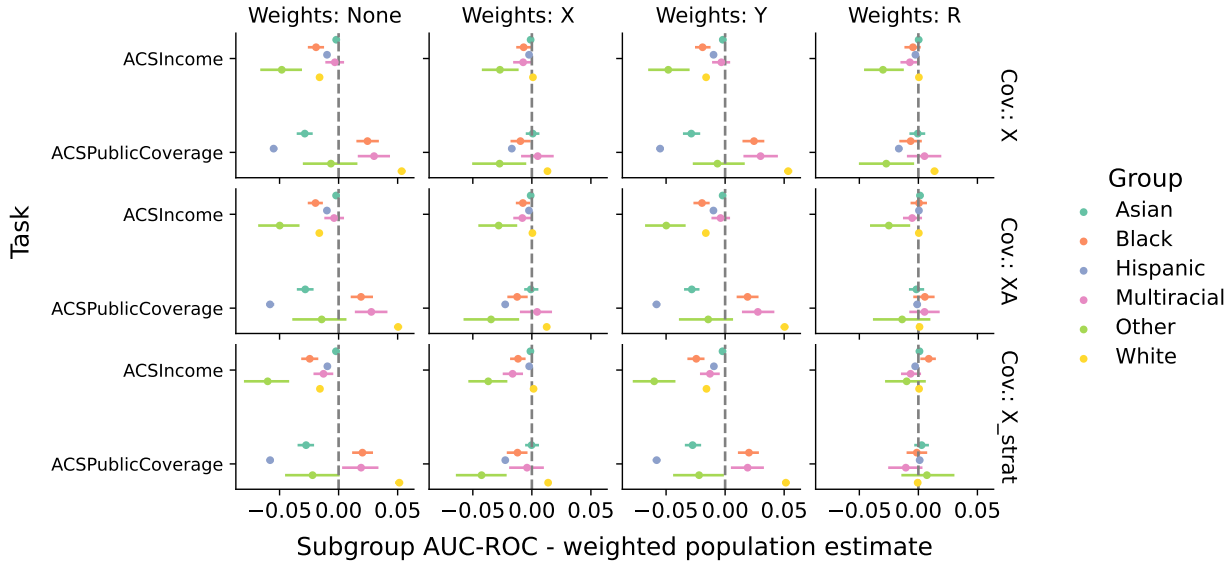

**Supplementary Figure B25: ACS PUMS: the effect of subgroup-aware prediction on model performance**. We report the difference in performance between models that have access to subgroup membership as an additional covariate as compared to those that do not. Plotted are average differences with 95% confidence intervals for each setting and for several performance metrics.



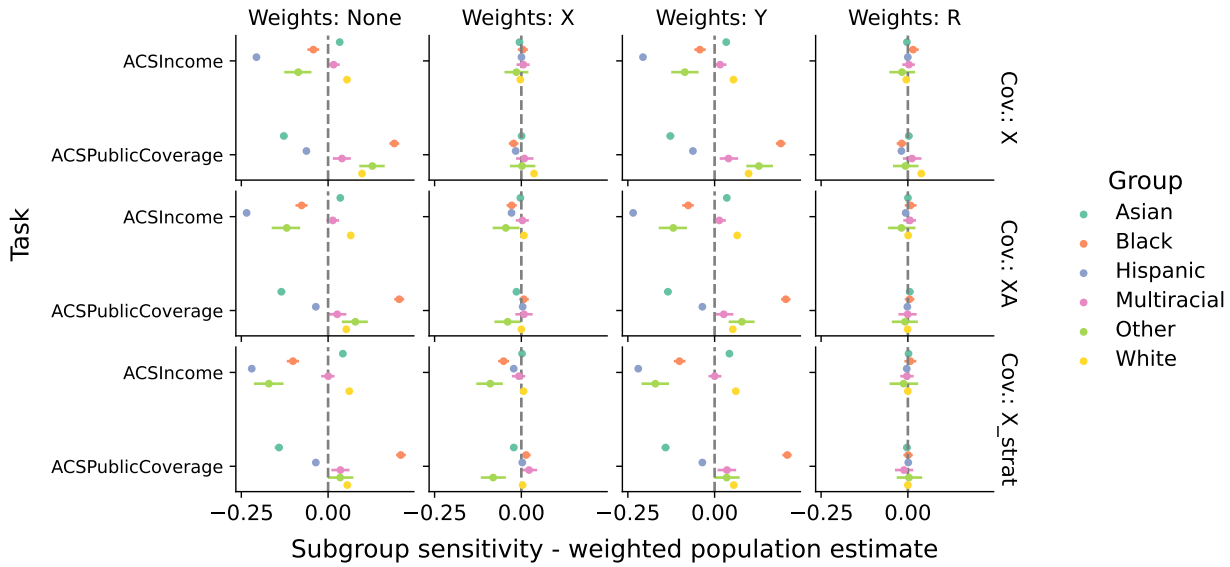

**Supplementary Figure B26: ACS PUMS: calibration curves**. Plotted are calibration curves for each subgroup with 95% confidence intervals. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.



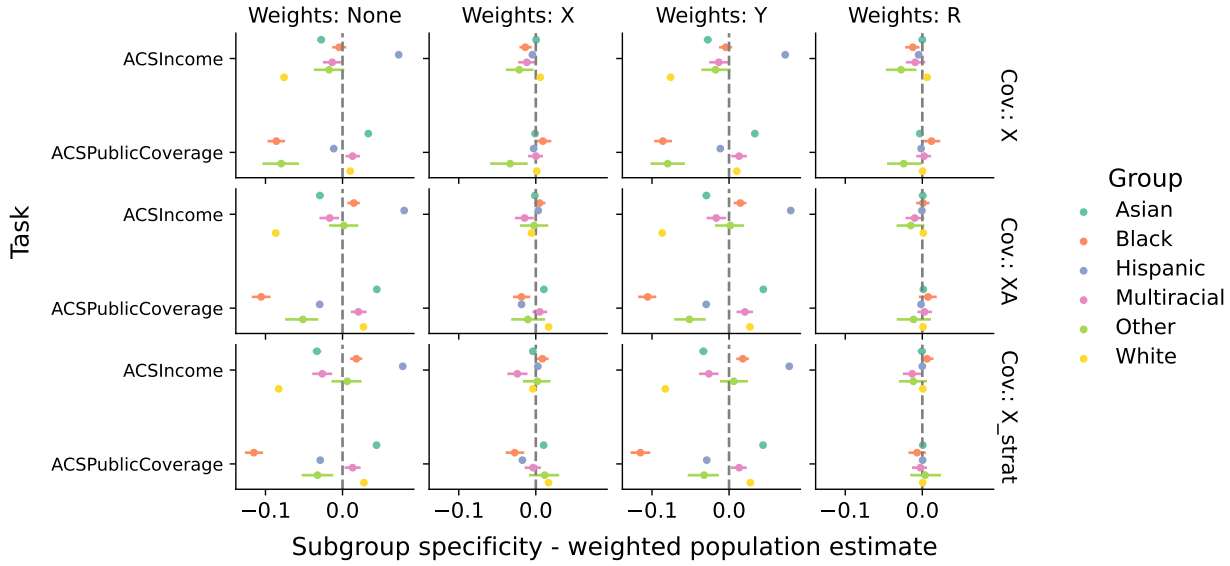

**Supplementary Figure B27: ACS PUMS: controlled evaluation of log loss**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between the unweighted disaggregated performance with the population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.



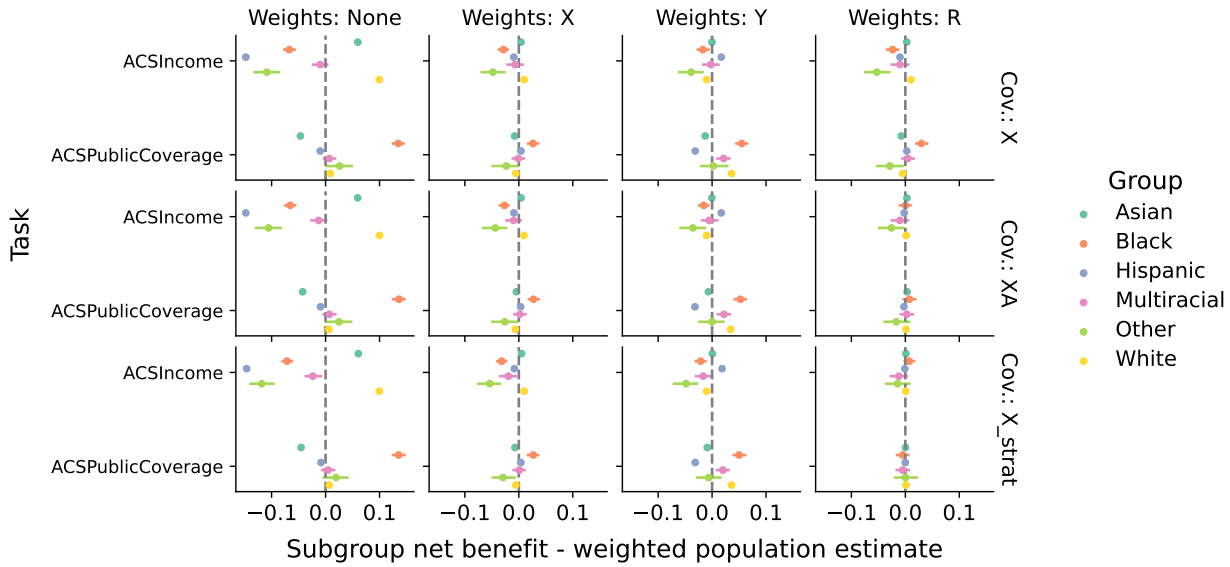

**Supplementary Figure B28: ACS PUMS: controlled evaluation of AUC-ROC**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between the unweighted disaggregated performance with the population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.



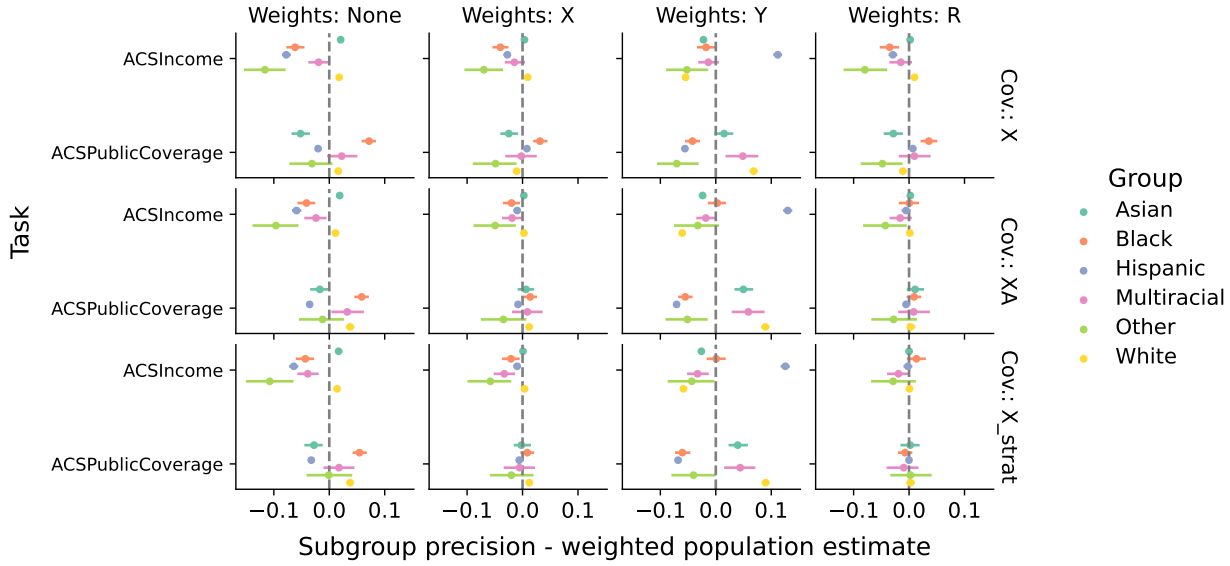

**Supplementary Figure B29: ACS PUMS: controlled evaluation of sensitivity**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between the unweighted disaggregated performance with the population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.



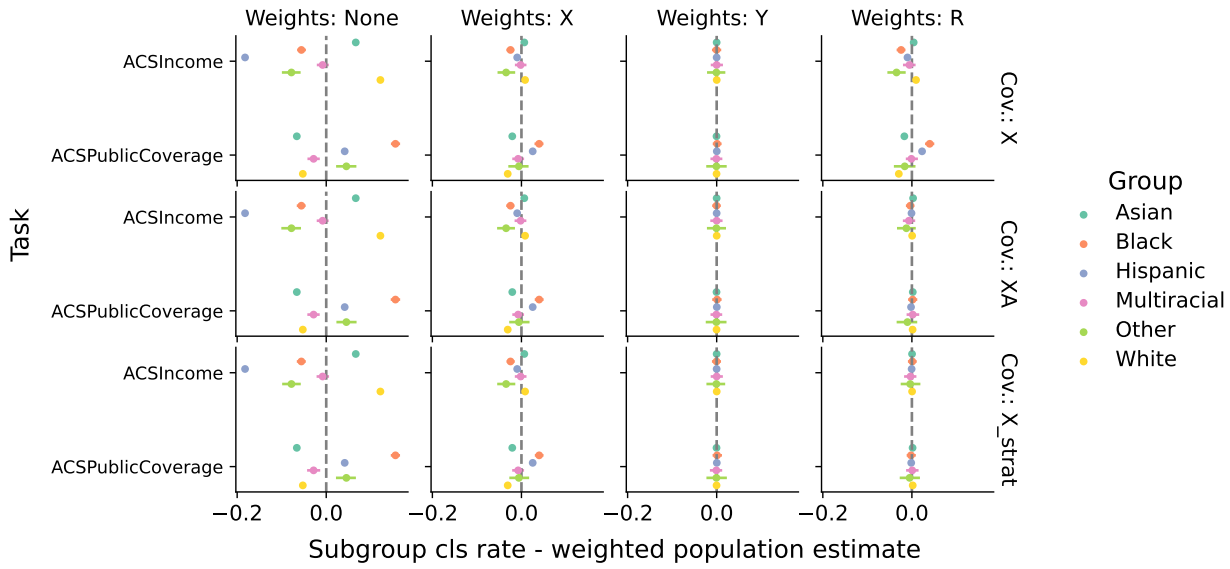

**Supplementary Figure B30: ACS PUMS: controlled evaluation of specificity**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between the unweighted disaggregated performance with the population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.




**Supplementary Figure B31: ACS PUMS: controlled evaluation of net benefit**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between the unweighted disaggregated performance with the population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.




**Supplementary Figure B32: ACS PUMS: controlled evaluation of precision**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between the unweighted disaggregated performance with the population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction, the second row to prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.




**Supplementary Figure B33: ACS PUMS: controlled evaluation of classification rate**. Plotted are the statistics $T_a$ with 95% confidence intervals, corresponding to differences between the unweighted disaggregated performance with the population performance weighted to match the distribution of $X$, $Y$, or $R$ on the subgroups. The first row corresponds to subgroup-agnostic prediction, the second to row prediction with $A$ as an additional covariate, and the third row to stratified prediction by $A$.