Understanding subgroup performance differences of fair predictors using causal models

Stephen R. Pfohl
Google Research
spfohl@google.com

Natalie Harris
Google Research

Chirag Nagpal
Google Research

David Madras
Google Research

Vishwali Mhasawade
New York University

Olawale Salaudeen
University of Illinois Urbana-Champaign

Katherine Heller
Google Research

Sanmi Koyejo
Google Deepmind

Alex D'Amour
Google Deepmind

Abstract

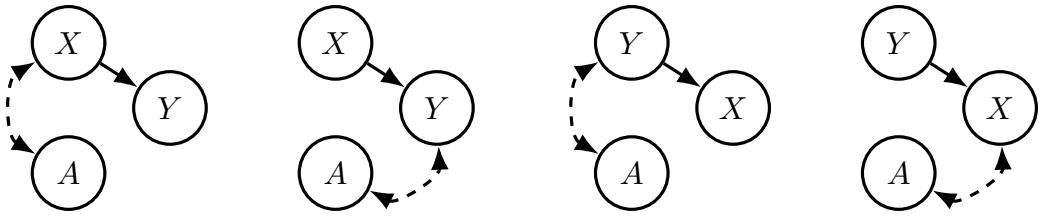
A common evaluation paradigm compares the performance of a machine learning model across subgroups to assess properties related to fairness. In this work, we argue that distributional differences across subgroups can render this approach to evaluation of fairness misleading. We consider distributional differences across subgroups as a source of confounding that can lead to differences in performance metrics across subgroups even if the relationship between covariates and a label of interest is modeled as well as possible for each subgroup. We show that these differences in model performance can be anticipated and characterized based on the causal structure of the data generating process and the choices made during the model fitting procedure (e.g. whether subgroup membership is used as a predictor). We demonstrate how to construct alternative evaluation procedures that control for this source of confounding during evaluation by implicitly matching the distribution of confounding variables across subgroups. We emphasize that the selection of appropriate control variables requires domain knowledge and selection of contextually inappropriate control variables can produce misleading results.

1 Introduction

A significant body of work uses disaggregated evaluation of machine learning models across subgroups (e.g. by race, ethnicity, or gender) in order to assess algorithmic fairness properties [1]. In this paradigm, differences in a performance metric or other statistical property (e.g. calibration or the distribution of predictions or covariates) across subgroups are taken as evidence of fairness violations that could potentially introduce or exacerbate inequity. As examples, Seyyed-Kalantari et al. [2] evaluate differences in classification performance in the context of classification of radiological findings from chest X-rays on the basis of race, sex, and socioeconomic status, and Gichoya et al. [3] investigate the predictability of racial categories from radiological images.

For the purposes of this work, we consider subgroup Bayes-optimality, *i.e.,* estimation of the conditional expectation of the label given the covariates for each subgroup, to be the primary fairness property of interest. As shown in Liu et al. [4], subgroup Bayes-optimality implies the fairness criterion of *sufficiency*, which can be interpreted as a condition of equal calibration curves across subgroups and follows directly from unconstrained empirical risk minimization that leverages covariates and information of subgroup membership (either explicitly or through subgroup-specific

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).



(a) Covariate shift; causal direction (b) Outcome shift; causal direction (c) Label shift; anti-causal direction (d) Presentation shift; anticausal direction

Figure 1: Causal graphs encoding assumptions regarding subgroup heterogeneity.

models) with large sample sizes and a well-specified model class. Subgroup Bayes-optimality can be motivated on the basis that it is consistent with modeling the data well and with optimal decision making without explicit trade-offs between fairness and predictive performance [5, 6].

In this work, we aim to develop understanding of a particular failure mode of the disaggregated evaluation paradigm, where models can achieve different average performance across subgroups despite satisfying subgroup Bayes-optimality and sufficiency. In prior work, this phenomenon is often discussed in terms of incompatibility between fairness criteria that renders it impossible to satisfy different notions of fairness simultaneously [4, 7–9]. Here, we argue that, when the data distribution differs across subgroups, a lack of parity in performance across subgroups is not necessarily unexpected, and may not motivate active algorithmic mitigation.

Our approach is to characterize the conditional independence properties of models (both arbitrary and Bayes-optimal) learned and evaluated under different data generating processes that encode different assumptions on the causal relationships between the covariates, labels, and subgroup membership. This approach mirrors related work that uses causal directed acyclic graphs to study the relationship between fairness and robustness to distribution shift [10–15]. However, the setting we study here differs in that we specifically focus on a setting without explicit distribution shift, *i.e.,* where the model is learned and evaluated on a fixed data distribution, with distribution shift implicitly present *across* subgroups. We show that in simple, prototypical cases, performance differs across subgroups, but can be shown to be equal conditioned on a confounder that varies in distribution marginally across subgroups. We further investigate the use of controlled comparisons that fix the distribution of confounding variables across subgroups as a complementary approach to standard uncontrolled comparisons.

2 Problem formulation and methodology

We consider data with covariates $X \in \mathbb{R}^n$, a binary label $Y \in \{0, 1\}$, and a categorical subgroup indicator A. We reason about properties of a model f that takes as inputs $Z \subseteq \{X, A\}$ to produce scores $R = f(Z)$ that can be compared to a threshold τ to yield binary predictions $\hat{Y} = \mathbb{1}[R \geq \tau]$. We note that because our scope is limited to modeling binary outcomes, it follows that $\mathbb{E}[Y \mid X] = P(Y = 1 \mid X)$, and thus $\mathbb{E}[Y \mid X]$ fully characterizes $P(Y \mid X)$.

Our approach relies on the use of causal directed acyclic graphs to describe the causal structure of the data generating processes of interest [16]. While we include A in these graphs to describe the role of subgroup heterogeneity 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 [17].

We consider four settings that are analogous to those commonly studied in distribution shift settings (Figure 1). In the first setting, we consider *covariate shift* across subgroups, where X and A are not independent (i.e., the distribution of X differs across subgroups), but $P(Y \mid X)$ is stable across subgroups (Figure 1a). In the second setting, we consider an *outcome shift*, where an unobserved confounder not independent of A has influence on Y, unmediated by X, such that $P(Y \mid X)$ differs across subgroups (Figure 1b). The third and fourth settings are examples of anticausal graphs, where the observed covariates X are downstream of the label Y. In this context, we consider a *label shift* setting, where the $Y \rightarrow X$ relationship is stable, but the prevalence of Y differs across subgroups

Table 1: Conditional independence properties of Bayes-optimal models.

	Sufficiency		Separation	
Setting	$Y \perp A \mid R^*$	$Y \perp A \mid R^*_A$	$R^* \perp A \mid Y$	$R^*_A \perp A \mid Y$
Covariate shift (Causal)	✓	✓	✗	✗
Outcome shift (Causal)	✗	✓	✗	✗
Label shift (Anticausal)	✗	✓	✓	✗
Presentation shift (Anticausal)	✗	✓	✗	✗

(Figure 1c), and a *presentation shift* setting, where the prevalence of Y is the same across subgroups, but $P(X \mid Y)$ differs (Figure 1d).

We focus on two categories of results. First, we use the structure of the causal directed acylic graphs to analytically describe the expected conditional independence properties of Bayes-optimal models learned and evaluated in these settings. Second, we consider the extent to which differences in performance metrics across subgroups may be anticipated and controlled for in these settings.

We note that our claims are weaker than fairness impossibility results [7–9] in that we focus on the statements of conditional independence and performance parity that can be shown to be necessarily satisfied, rather than necessarily violated. As such, when we claim that conditional independence or performance parity is not satisfied for a particular graph, it is a statement that the condition does not hold in general, even if it is possible to construct a distribution consistent with the graph or select a specific performance metric for which the condition of interest does hold.

2.1 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 that estimate $\mathbb{E}[Y \mid Z]$ for $Z \subseteq \{X, A\}$. We define f^* as the *population* Bayes-optimal model that estimates the conditional expectation of Y given covariates X, such that $R^* = f^*(X) = \mathbb{E}[Y \mid X]$. We define the *subgroup* Bayes-optimal model as the model that estimates the conditional expectation of Y given covariates X and information of subgroup membership, i.e., $\mathbb{E}[Y \mid X, A]$. This can be represented as a single model $R^*_A = f^*_A(X, A) = \mathbb{E}[Y \mid X, A]$ or a set of subgroup-specific Bayes-optimal models ($\{f^*_a\}_{a \in \mathcal{A}}$ for $R^*_a = f^*_A(X, A = a) = f^*_a(X) = \mathbb{E}[Y \mid X, A = a]$).

To analyze the properties of Bayes-optimal models, we reason about the (conditional) independence properties of X, Y, and A that follow from the causal graph for each setting, subject to the constraint that $Y \perp Z \mid f(Z)$ under Bayes-optimality. We focus on the sufficiency ($Y \perp A \mid R$) and separation ($R \perp A \mid Y$) fairness criteria. The key results are summarized in Table 1.

To assess sufficiency, we reason about the stronger condition of subgroup Bayes-optimality as a sufficient condition for sufficiency with binary Y [4]. For each of the settings considered, the Bayes-optimal model that depends on both X and A is subgroup Bayes-optimal and satisfies sufficiency. However, of the settings that we study, the covariate shift setting is the only one where subgroup Bayes-optimality is obtained using only X as input. This follows because $Y \perp A \mid X$ implies that $R^* = \mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid X, A]$ and thus $\mathbb{E}[Y \mid R^*] = \mathbb{E}[Y \mid R^*, A]$. In each of the other graphs, we have that $Y \not\perp A \mid X$, which implies that the population Bayes-optimal model is not, in general, subgroup Bayes-optimal. In the outcome shift setting, the bidirected edge between A and Y implies that the relationship between X and Y varies across subgroups. In the presentation shift setting, X takes the role of a collider variable that introduces an association between Y and A conditioned on X, despite the marginal independence of Y and A. In the label shift setting, while the subgroup Bayes-optimal model differs, in general, from the population Bayes-optimal model, the subgroup Bayes-optimal model can be derived from the population Bayes-optimal with a straightforward post-hoc adjustment [18, 19].

Of the settings of interest, we find that separation only necessarily holds under the label shift graph, and there only for models that only depend on X, but not those that depend on X and A. Furthermore,

Table 2: Model performance properties over subgroups in different settings.

	Setting		$\{R, Y\} \perp A \mid V$			
Graph	$Z \subseteq \{X, A\}$	Model	$V = \{\}$	X	Y	R
Covariate shift (Causal)	X	f^*	✗	✓	✗	✓
		f	✗	✓	✗	✗
	$\{X, A\}$	f_A^*	✗	✓	✗	✓
		f	✗	✗	✗	✗
Outcome shift (Causal)	X	f^*	✗	✗	✗	✗
		f	✗	✗	✗	✗
	$\{X, A\}$	f_A^*	✗	✗	✗	✓
		f	✗	✗	✗	✗
Label shift (Anticausal)	X	f^*	✗	✗	✓	✗
		f	✗	✗	✓	✗
	$\{X, A\}$	f_A^*	✗	✗	✗	✓
		f	✗	✗	✗	✗
Presentation shift (Anticausal)	X	f^*	✗	✗	✗	✗
		f	✗	✗	✗	✗
	$\{X, A\}$	f_A^*	✗	✗	✗	✓
		f	✗	✗	✗	✗

in the label shift case, the argument does not require Bayes-optimality, because controlling the distribution of Y controls the distribution of X and thus of an arbitrary R. In all other cases that we study, the separation criteria is not implied by the graph. This result is consistent with works that show that fitting performant, well-calibrated predictive models for each subgroup is in conflict with separation and the related equalized odds criterion [4, 6, 5].

2.2 Controlled evaluation to characterize subgroup performance differences

In this section, we describe scenarios in which average performance is expected to be equal across subgroups for models trained and evaluated in-distribution. We consider performance metrics $m : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ that can be computed at an instance-level and aggregated as a mean over a distribution. We note that $\{R, Y\} \perp A$ is a sufficient condition for equal average performance across subgroups, as fixing the distribution $P(R, Y)$ fixes the distribution of m. It follows then that unequal performance across subgroups implies $\{R, Y\} \not\perp A$. To aid in reasoning about conditional independence properties that involve R and m, we include expanded causal graphs that explicitly represent them as deterministic transformations of their parents (Supplementary Figure B1).

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 an arbitrary reference distribution $P_0(V)$ ensures equal performance across subgroups. In practice, controlled evaluations can be constructed through weighting procedures that implicitly match the distribution of V across subgroups.

We briefly describe the approach here, and include a more thorough description in supplementary section A.1. We consider a source distribution $\mathcal{P}$ and target distribution $\mathcal{Q}$ over V such that we compute the performance using data from $\mathcal{P}$ after fixing the marginal of V to $P_{\mathcal{Q}}(V)$. The expectation is given by $\int w * m(R, Y)P_{\mathcal{P}}(R, Y \mid v)P_{\mathcal{P}}(v)dv$ for weights $w \propto \frac{P_{\mathcal{Q}}(V)}{P_{\mathcal{P}}(V)}$. There are several ways to define $\mathcal{P}$ and $\mathcal{Q}$ for controlled comparison of subgroups. In our experiments, we consider cases with two subgroups and adopt the strategy proposed in Namkoong et al. [20], where we consider a

fixed target distribution $\mathcal{Q} \propto \frac{P(V|A=a)P(V|A=a')}{P(V|A=a)+P(V|A=a')}$ that has zero density in cases where the control variable has zero support for either subgroup.

In Table 2, we summarize the key results that can be inferred graphically about performance parity across subgroups in different settings, including for different control variable sets (X, Y, R, or $\{\}$, where $\{\}$ corresponds to an uncontrolled marginal comparison). In supplementary section A.2, we describe a simulated experiment that verifies the results in Table 2.

For the settings that we study, we find that performance is, in general, unequal for Bayes-optimal models (both population and subgroup Bayes-optimal) and for arbitrary non-optimal models. For models that depend only on X, 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 depends on both X and A 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 the outcome shift and presentation shift graphs, 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 distribution of the *optimal risk score* across subgroups. Furthermore, if sufficiency holds without subgroup Bayes-optimality, it still follows that controlling for R implies equal performance across subgroups. This reveals a connection between testing for sufficiency and controlled comparison for the distribution of the risk score, in that unequal performance after controlling for R implies sufficiency is violated. We note that these observations hold for all of the graphs considered and for graphs involving compound shifts where the relationship of A with both X and Y is confounded.

3 Discussion and conclusion

Our work highlights the challenges associated with the interpretation of disaggregated evaluations of machine learning models over subgroups and has potential for far-reaching implications due to the ubiquity of disaggregated evaluation for assessment of fairness and robustness (e.g. Koh et al. [21]). In particular, we emphasize that modeling the relationship between covariates and labels well for each subgroup, and in a manner consistent with fairness, does not imply equal performance across subgroups. However, these differences can be anticipated and controlled for if aspects of the causal structure of the data are known. An important aspect of our analysis is that we largely focus on the properties of Bayes-optimal models in-distribution, corresponding to a setting where data collection is large-scale, diverse, representative, and free of observational biases (e.g. measurement error, selection bias, or missing outcomes). This highlights that the properties that we study remain even with arbitrarily large datasets, models that fit the data well, and with measures taken to ensure that the data is representative and free of bias. Important areas of future methods development includes an extension of our approach to incorporate observational biases, explicit distribution shift, and finite-sample effects, including estimation error.

It remains to be seen how to best use controlled evaluation procedures to complement standard uncontrolled fairness analyses. The distributional differences across subgroups that our approach controls for are more-often-than-not the direct result social and structural inequities that, for example, lead to differences in the measures of disease (X) and health outcomes (Y) across populations [22]. The lens that we take here implicitly assumes 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. 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.

Acknowledgments and Disclosure of Funding

This study was funded by Google LLC and/or a subsidiary thereof ('Google').

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] 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.

[5] 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.

[6] 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.

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

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

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

[10] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*, 2021.

[11] 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.

[12] 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`.

[13] 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.

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

[15] 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.

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

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

[18] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[19] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

[20] Hongseok Namkoong, Steve Yadlowsky, et al. Diagnosing model performance under distribution shift. *arXiv preprint arXiv:2303.02011*, 2023.

[21] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[22] 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.

[23] 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.

Supplementary Material

A Supplementary methods

A.1 Weighting approaches to controlled evaluation

To begin, we consider estimating model performance under a distribution shift, such that if we have some source distribution $\mathcal{P}$ and a target distribution $\mathcal{Q}$ over $\{R, Y\}$, then we can estimate performance on $\mathcal{Q}$ 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 ($P_{\mathcal{Q}}(R, Y) > 0 \rightarrow P_{\mathcal{P}}(R, Y) > 0$). In the context of subgroup comparisons, there are several reasonable ways to set $\mathcal{P}$ and $\mathcal{Q}$. For example, one can compute subgroup performance from the aggregate population using $\mathcal{P} = P(R, Y)$ and $\mathcal{Q} = P(R, Y \mid A = a)$ with weights $w \propto P(A = a \mid V)$ or one can compare subgroups $A = a$ to $A = a'$ pairwise with $\mathcal{P} = P(R, Y \mid A = a)$, $\mathcal{Q} = P(R, Y \mid A = a')$ with weights $w \propto \frac{P(A=a'|R,Y)}{P(A=a|R,Y)}$.

As described in section 2.2, 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 expectation is given by $\int w * m(R, Y) P_{\mathcal{P}}(R, Y \mid v) P_{\mathcal{P}}(v) dv$ for weights $w \propto \frac{P_{\mathcal{Q}}(V)}{P_{\mathcal{P}}(V)}$, or $w \propto P(A = a \mid V)$ and $w \propto \frac{P(A=a'|V)}{P(A=a|V)}$ for the aggregate vs. subgroup and pairwise comparison settings, respectively. From this expression, 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 in $\mathcal{P}$ is equal to the marginal performance in $\mathcal{Q}$. If weighted performance in $\mathcal{P}$ is not equal to the marginal performance in $\mathcal{Q}$, then $\{R, Y\} \not\perp A \mid V$.

In our experiments, we use an alternative weighting strategy for pairwise comparisons proposed in Namkoong et al. [20], where we consider 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 $A = a'$ [20].

A.2 Simulation study

We conduct a small simulation study to investigate the effect of controlling for variability in X, Y, or R across subgroups. We construct one data generating process for each of the four settings that we study (Figure 1). The data generating processes are provided in section A.3. For each data generating process, we sample 20,000 samples I.I.D and use 10,000 for training and reserve 10,000 as a test dataset for evaluation. All model fitting and evaluation procedures are separately for the case where X is used for prediction and for the case where both X and A are used. When both X and A are used, we fit separate models for each subgroup. To fit the models for Y, we use the scikit-learn [23] implementation of ridge regression with five-fold cross-validation for the inverse regularization parameter C over the grid $[0.01, 0.1, 1., 10., 100]$, refitting the model over the training data with the best value of C on the basis of the average held-out log-loss over the cross-validation folds.

We use the weighting formula of Namkoong et al. [20] so that weighted performance for pairs of subgroups may be directly compared. We fit models for $P(A = a' \mid V)$ for $V = X, Y$, or R using the test data, without cross-fitting. We use the scikit-learn implementation of histogram-based gradient boosting classification trees with five-fold cross-validation with a grid over the maximum number of leaf nodes in $[10, 25, 50]$. The best hyperparameters are used to refit the group membership model and predict group membership for the full test dataset. We compute weighted performance estimates for the log-loss (Supplementary Table B1), area under the receiver operating characteristic curve (Supplementary Table B2), recall at a threshold of 0.5 (Supplementary Table B3), and specificity at a threshold of 0.5 (Supplementary Table B4).

A.3 Data generating processes

A.3.1 Causal data generating processes

This description encompasses the covariate shift and outcome 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 $\gamma_A = 1$, $\beta_{a_0} = \beta_{a_1} = 0.5$, and $\beta_{a_0} = \beta_{a_1} = 0$. For the outcome shift setting, we set $\gamma_A = 0$, $\beta_{a_0} = 0.5$, $\beta_{a_1} = -1$, and $\alpha_{a_0} = \alpha_{a_1} = 0$.

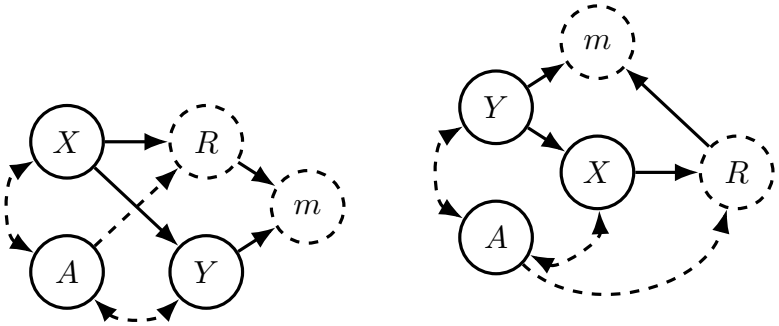
$$
\begin{aligned}
U &\sim \text{Bernoulli}\big(0.5\big) \\
X \mid U = 0 &\sim \mathcal{N}(-2, 1) \\
X \mid U = 1 &\sim \mathcal{N}(0, 1) \\
A \mid U &\sim \gamma_A U + (1 - \gamma_A) * \text{Bernoulli}\big(0.5\big) \\
Y \mid A = 0 &\sim \text{Bernoulli}\Big(\text{logit}^{-1}\big(\beta_{a_0} x + \alpha_{a_0}\big)\Big) \\
Y \mid A = 1 &\sim \text{Bernoulli}\Big(\text{logit}^{-1}\big(\beta_{a_1} x + \alpha_{a_1}\big)\Big)
\end{aligned}
$$

A.3.2 Anticausal data generating processes

This description encompasses the covariate shift and outcome 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 $\pi_{Y_0} = 0.5$, $\pi_{Y_1} = 0.5$, $\mu_{A_0Y_0} = 1$, $\mu_{A_0Y_1} = 0$, $\mu_{A_1Y_0} = -1$, $\mu_{A_1Y_1} = 1$.

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

B Supplementary figures and tables



(a) Causal direction (b) Anticausal direction

Supplementary Figure B1: Extended graphs that incorporate the model output R and instance-level value of the metric m. Nodes with dashed outline are deterministic transformations of their parents.

Supplementary Table B1: Results of simulation study showing estimated **log-loss** after controlling for confounding variables.

		$Z = X$		$Z = \{X, A\}$	
Setting	V	$A = a$	$A = a'$	$A = a$	$A = a'$
Covariate shift	$\{\}$	0.567	0.663	0.567	0.663
	X	0.639	0.647	0.640	0.648
	Y	0.658	0.666	0.657	0.665
	R	0.639	0.647	0.633	0.649
Outcome shift	$\{\}$	0.736	0.641	0.622	0.499
	X	0.735	0.640	0.623	0.497
	Y	0.715	0.657	0.676	0.565
	R	0.735	0.640	0.636	0.627
Label shift	$\{\}$	0.421	0.249	0.372	0.193
	X	0.445	0.315	0.394	0.242
	Y	0.291	0.296	0.366	0.320
	R	0.445	0.315	0.342	0.327
Presentation shift	$\{\}$	0.784	0.565	0.591	0.363
	X	0.792	0.588	0.581	0.400
	Y	0.783	0.565	0.591	0.363
	R	0.792	0.588	0.549	0.544

Supplementary Table B2: Results of simulation study showing estimated **AUC** (area under the receiver operating characteristic curve) after controlling for confounding variables.

		$Z = X$		$Z = \{X, A\}$	
Setting	V	$A = a$	$A = a'$	$A = a$	$A = a'$
Covariate shift	$\{\}$	0.622	0.636	0.622	0.636
	X	0.605	0.606	0.605	0.606
	Y	0.622	0.636	0.622	0.636
	R	0.605	0.606	0.610	0.603
Outcome shift	$\{\}$	0.322	0.800	0.678	0.800
	X	0.324	0.801	0.676	0.801
	Y	0.322	0.800	0.678	0.800
	R	0.324	0.802	0.683	0.705
Label shift	$\{\}$	0.914	0.913	0.914	0.913
	X	0.892	0.914	0.892	0.914
	Y	0.914	0.913	0.914	0.913
	R	0.892	0.914	0.879	0.887
Presentation shift	$\{\}$	0.248	0.919	0.752	0.919
	X	0.236	0.894	0.764	0.894
	Y	0.248	0.919	0.752	0.919
	R	0.236	0.894	0.795	0.800

Supplementary Table B3: Results of simulation study showing estimated **recall** at a threshold of 0.5 after controlling for confounding variables.

		$Z = X$		$Z = \{X, A\}$	
Setting	V	$A = a$	$A = a'$	$A = a$	$A = a'$
Covariate shift	$\{\}$	0.049	0.613	0.049	0.604
	X	0.156	0.144	0.156	0.137
	Y	0.049	0.613	0.049	0.604
	R	0.156	0.144	0.149	0.128
Outcome shift	$\{\}$	0.734	0.934	0.434	0.865
	X	0.741	0.933	0.434	0.866
	Y	0.734	0.934	0.434	0.865
	R	0.741	0.933	0.587	0.596
Label shift	$\{\}$	0.696	0.709	0.828	0.413
	X	0.524	0.792	0.701	0.496
	Y	0.696	0.709	0.828	0.413
	R	0.524	0.792	0.515	0.524
Presentation shift	$\{\}$	0.440	0.775	0.691	0.835
	X	0.369	0.799	0.737	0.860
	Y	0.440	0.775	0.691	0.835
	R	0.369	0.799	0.733	0.721

Supplementary Table B4: Results of simulation study showing estimated **specificity** at a threshold of 0.5 after controlling for confounding variables.

		$Z = X$		$Z = \{X, A\}$	
Setting	V	$A = a$	$A = a'$	$A = a$	$A = a'$
Covariate shift	$\{\}$	0.985	0.570	0.986	0.581
	X	0.931	0.923	0.936	0.928
	Y	0.985	0.570	0.986	0.581
	R	0.931	0.923	0.941	0.931
Outcome shift	$\{\}$	0.104	0.362	0.800	0.526
	X	0.103	0.366	0.798	0.528
	Y	0.104	0.362	0.800	0.526
	R	0.103	0.366	0.682	0.701
Label shift	$\{\}$	0.922	0.923	0.838	0.982
	X	0.957	0.877	0.891	0.967
	Y	0.922	0.923	0.838	0.982
	R	0.957	0.877	0.945	0.946
Presentation shift	$\{\}$	0.216	0.885	0.683	0.842
	X	0.248	0.817	0.659	0.750
	Y	0.216	0.885	0.683	0.842
	R	0.248	0.817	0.718	0.729