A Causal Perspective On Label Bias

Vishwali Mhasawade
vishwalim@nyu.edu
New York University
USA

Alexander D’Amour
alexdamour@google.com
Google DeepMind
USA

Stephen R Pfohl
spfohl@google.com
Google Research
USA

ABSTRACT

Predictive models developed with machine learning techniques are commonly used to inform decision making and resource allocation in high-stakes contexts, such as healthcare and public health. One means through which this practice may propagate equity-related harms is when the data used for model development or evaluation exhibits label bias. In such cases, the target of prediction is a proxy label of a construct of interest that may be difficult or impossible to measure, while the relationship between the proxy and the construct of interest differs systematically across subgroups. Label bias can be especially challenging to identify and mitigate in practice because consequential fairness violations are masked when the model is evaluated with respect to the proxy label. In this work, we aim to develop further formal understanding of label bias to inform the development of approaches for the identification and mitigation of it. To do so, we present desiderata for unbiased and biased proxy labels, introduce candidate causal graphical criteria for label bias, and consider the extent to which proxy labels can be used to reason about fairness with respect to a true construct of interest. We validate our findings with a simulation study and experiments with synthetic health insurance data used in the context of a care management system.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

Label bias, Proxy, Model Evaluation, Fairness

ACM Reference Format:
Vishwali Mhasawade, Alexander D’Amour, and Stephen R Pfohl. 2024. A Causal Perspective On Label Bias. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3630106.3658972

1 INTRODUCTION

There is a growing need to evaluate prediction models for algorithmic fairness and bias in high-stakes decision making contexts [17, 21, 54, 56, 62]. In these contexts, it is typical to fit a model to an imperfect proxy of a construct of interest and then use predictions of the proxy as the basis for a decision. A major concern is that these decisions may harm specific subgroups or otherwise introduce or exacerbate disparities.

As a motivating example, we consider the policy and predictive model studied in Obermeyer et al. [56], where it was shown that the choice of label used for prediction can directly introduce fairness concerns with potential to exacerbate racial health disparities [55]. This an example of the broader *label bias* problem [36, 37, 44], where systematic differences in the relationship between a proxy label and an unobserved construct of interest across subgroups can render standard evaluation approaches misleading and mask consequential fairness and equity-related harms. In this example, patients are referred to a care management program on the basis of predictions of healthcare expenditure from a model developed using administrative insurance claims data, where healthcare expenditures are assumed to be a proxy for the need for healthcare (health status). It was found that despite evidence that the model estimated healthcare expenditures well for all groups, bias was present given that the number of active chronic conditions was greater for the Black population relative to the White population conditioned on the value of the risk score, suggesting a difference in the implicit threshold of referral [31, 67] on the basis of health status for the two groups. Here, healthcare expenditure is a *biased proxy* of health status because the relationship between healthcare expenditure and health status differs across racial groups due to differential exposure to structural racism that systematically limits access to healthcare for the Black population in the United States [5, 6, 51, 76].

In this work, we formalize and characterize label bias using statistical and causal criteria, building on prior works that describe related forms of bias with causal graphical models of measurement [37, 43, 57]. We further consider implications for evaluation of fairness, with particular attention paid to the *sufficiency* fairness criterion that assesses differential miscalibration across subgroups [9, 50]. In cases where multiple proxy labels are available, we study how the joint causal structure between multiple proxy labels and the unobserved label of interest enables assessment of bias with respect to a given proxy. We present multiple causal structures and describe conditions that relate fairness evaluation with respect to a proxy label to fairness evaluation with respect to the unobserved label of interest. Furthermore, we also highlight a secondary issue related to the effect of including subgroup information as a predictor on fairness properties measured with respect to a proxy label and the unobserved label of interest. We evaluate the appropriateness of this formalization using a simulation study and re-analysis of the setting of Obermeyer et al. [56] using the synthetic dataset released by the study authors. We provide technical background in section 2, detail our framework for understanding label bias in section 3,

FAccT ’24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658972

describe related work in section 4, describe our experiments and results in section 5[1], and discuss our findings in section 6.

2 BACKGROUND

2.1 Causal direct acyclic graphs

We study the label bias problem using causal directed acyclic graphs (DAGs). Following Pearl et al. [57] and Mhasawade and Chunara [53], we define a causal model as a triple of sets $(\mathbf{U}, \mathbf{V}, F)$ such that:

- $\mathbf{U}$ are a set of latent variables, which are not caused by any of the observed variables in $\mathbf{V}$.
- F is set of functions such that for each $V_i \in \mathbf{V}, V_i = f_i\left(pa_i, U_{pa_i}\right)$, $pa_i \subseteq \mathbf{V} \setminus V_i$ and $U_{pa_i} \subseteq \mathbf{U}$, where pa_i refers to the *causal parents* of V_i.
- The joint distribution over all variables is given by the product of the conditional distribution of each variable given its causal parents: $\Pr(\mathbf{V}) = \prod_i \Pr\left(V_i | pa_i\right)$.

2.2 Modeling in the well-specified setting

We consider development and evaluation of a model h using data samples from $P(X, Y, A)$, comprising covariates $X \subseteq \mathcal{X} = \mathbb{R}^m$, a categorical group attribute A, and a label Y. For simplicity, we consider A to be binary to indicate two subgroups. Unless otherwise stated, Y is assumed to be binary in $\{0, 1\}$. The model h is assumed to take as input $Z \subset \{X, \{X, A\}\}$ and to output a continuously-valued score that can be compared to a threshold τ to produce a hard prediction $\hat{Y}$. For the purposes of this work, we assume that the data used for model development and evaluation are independent and identically distributed samples from the same underlying population and that the distribution of that population matches that of the target population that the model is intended to be used for [68].

2.3 Sufficiency, calibration, and subgroup Bayes-optimality

For evaluation of fairness, we primarily focus on the *subgroup Bayes-optimality*, *calibration*, and *sufficiency* criteria [9, 50]. The sufficiency criteria is given by $\mathbb{E}[Y \mid h(Z), A = a_i] = \mathbb{E}[Y \mid h(Z)]\, \forall a_i$, *i.e.*, that the calibration curves are the same for all groups. For a binary outcome Y, this is equivalent to $Y \perp\!\!\!\perp A \mid h(Z)$. This is related to calibration for all groups, *i.e.*, $\mathbb{E}[Y \mid h(Z), A = a_i] = h(Z)\, \forall a_i$, in that calibration for all groups implies sufficiency. Subgroup Bayes-optimality is the condition that the conditional expectation of Y given $\{X, A\}$ is modeled as well as possible for each subgroup, *i.e.*, $h = \mathbb{E}[Y \mid X, A]$. The relationship between the criteria is such that subgroup Bayes-optimality implies calibration for all groups, which implies sufficiency [9]. The converse statements do not generally hold, in that sufficiency does not imply calibration for all groups, and calibration for all groups does not imply subgroup Bayes-optimality [21]. However, violation of sufficiency *does* imply violation of subgroup Bayes-optimality, and thus that the model is a sub-optimal predictor for at least one subgroup [50].

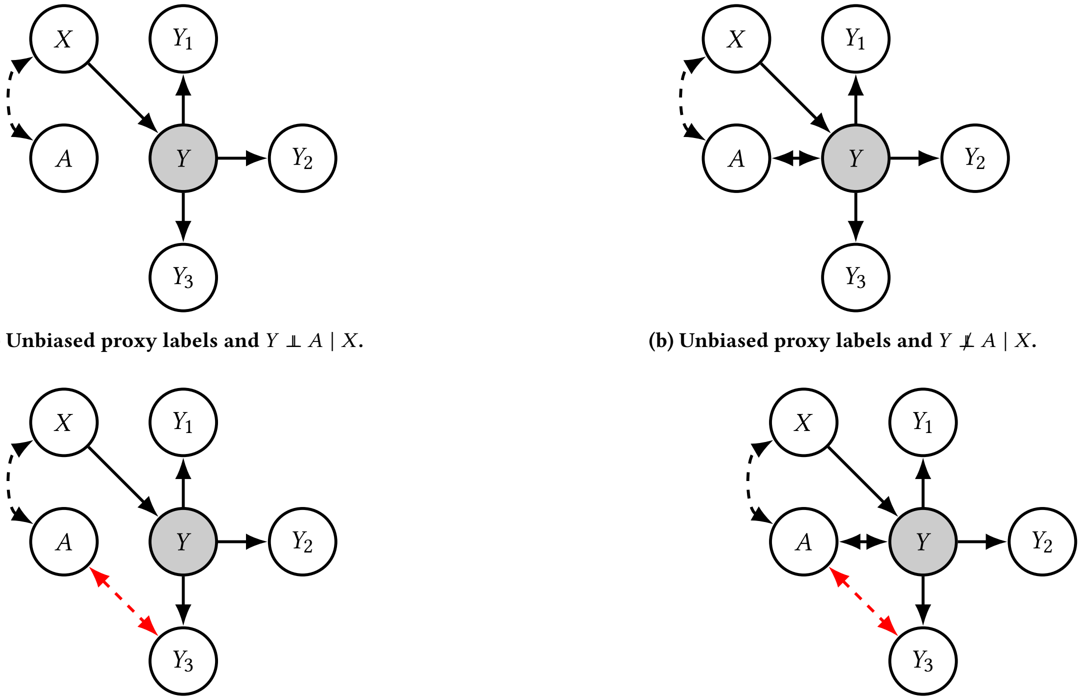
Subgroup Bayes-optimality and sufficiency are important criteria for reasoning about fairness and decision-making. A decision rule that thresholds the subgroup Bayes-optimal model yields an optimal decision rule overall and for each subgroup if certain assumptions of the decision making context are met (*e.g.*, if the utility of the decision monotonically increases as a function of the calibrated risk score and is independent of group membership, such as when fixed costs or utilities are associated with each of true positive, true negative, false positive, and false negative classification errors) [7, 60]. Furthermore, if a model satisfies sufficiency, a single-threshold decision rule informally corresponds to an equal treatment condition across subgroups with respect to the conditional probability of the outcome given the predicted score, such that violation of sufficiency implies different implicit thresholds across subgroups [31]. Other common fairness criteria, such as demographic parity [15, 27], equalized odds [40], and predictive parity [19], can be misleading in this setting because they can be violated for subgroup Bayes-optimal models when the data distribution differs across subgroups [20, 50].

Analysis of causal structure provides some means to anticipate the fairness properties of models learned from data faithful to the causal structure [61]. For example, if the graph is such that $Y \perp\!\!\!\perp A \mid X$ (Figure 1a), then a *population Bayes-optimal* model that accurately estimates $\mathbb{E}[Y \mid X]$ is also optimal for each subgroup given that $Y \perp\!\!\!\perp A \mid X$ implies that $\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid X, A]$. However, if $Y \not\perp\!\!\!\perp A \mid X$, *i.e.* if X does not d-separate X and A, then $\mathbb{E}[Y \mid X] \neq \mathbb{E}[Y \mid X, A]$ in general, the population Bayes-optimal model that depends only on X need not be subgroup Bayes-optimal and may violate sufficiency. We represent this setting causally with a bidirected arrow between A and Y to indicate the presence of an unobserved confounder that influences Y, unmediated by X, with a differing distribution across subgroups A (Figure 1b). In such cases, the gap between the population and subgroup Bayes-optimal predictors can be addressed by incorporating subgroup membership information as an input to the model, which can take the form of fitting a separate model for each subgroup or by considering an indicator of subgroup membership as an additional covariate along with X. Here, we notate models that do not incorporate subgroup information as $h_{\not A}$ and those that do as h_A. We accordingly refer to the population Bayes-optimal model $\mathbb{E}[Y \mid X]$ as h^* or $h^*_{\not A}$ and the subgroup Bayes-optimal model $\mathbb{E}[Y \mid X, A]$ as h^*_A.

3 PROXY LABELS AND LABEL BIAS

Here, we formalize candidate statistical definitions for unbiased and biased proxy labels in terms of the concepts introduced in section 2 and discuss their relationship to related causal graphical criteria. In a setting with proxy labels, we do not observe the true outcome of interest Y but instead observe a set of proxy labels $\mathbf{Y_P} = \{Y_1, \cdots, Y_n\}$. We consider a mode of evaluation considered in prior work [56, 65, 78] where violation or satisfaction of sufficiency with respect to a proxy variable is used to reason about sufficiency with respect to the true label. Within this framework, we may consider a proxy label Y_i to be unbiased if it enables reasoning about whether a predictive model satisfies sufficiency with respect to the true label Y on the basis of a test for sufficiency with respect to Y_i. Moreover, we may consider a proxy label Y_i to be biased if a subgroup Bayes-optimal estimate for Y_i violates sufficiency with respect to Y. This translates to a statistical criterion defined in terms

[1] Code to replicate experiments is available at https://github.com/google-research/google-research/tree/master/causal_label_bias.



(a) Unbiased proxy labels and $Y \perp\!\!\!\perp A \mid X$.

(b) Unbiased proxy labels and $Y \not\perp\!\!\!\perp A \mid X$.

(c) Y_3 is a biased proxy label and $Y \perp\!\!\!\perp A \mid X$.

(d) Y_3 is a biased proxy label and $Y \not\perp\!\!\!\perp A \mid X$.

Figure 1: a) Example causal DAG with unobserved outcome of interest Y represented by the gray node and three observed proxy labels Y_1, Y_2, Y_3. All the proxy labels are downstream of the unobserved outcome of interest with $Y \rightarrow Y_i$. Bidirected edges represent confounding between variables. A red edge is used to indicate a biased proxy relationship. All three proxies are unbiased in (a) and (b), while Y_3 is a biased proxy in (c) and (d). We show cases where $Y \perp\!\!\!\perp A \mid X$ in (a) and (c) and cases where $Y \not\perp\!\!\!\perp A \mid X$ in (b) and (d).

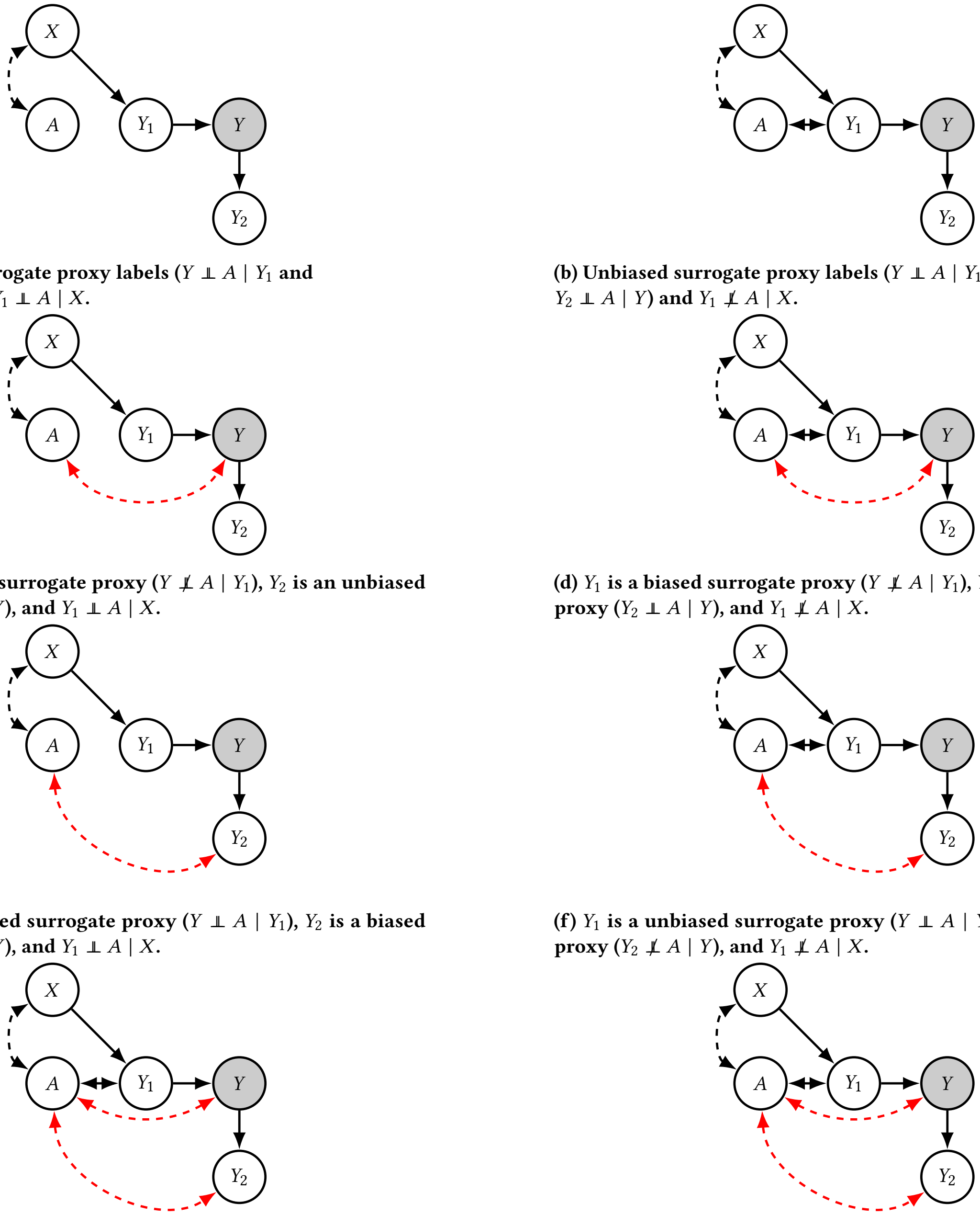
of $\{Y, Y_i, X, A\}$, where

$$\mathbb{E}[Y \mid \mathbb{E}[Y_i \mid X, A], A = a_j] \neq \mathbb{E}[Y \mid \mathbb{E}[Y_i \mid X, A], A = a_k] \quad (1)$$

for two groups a_j and a_k if Y_i is a biased proxy label. This notion of bias can be motivated decision-theoretically in that decision making at a single threshold on $h^*_A(Z)$ implies decision making at a different implicit thresholds for Y across subgroups if $h^*_A(Z)$ is a subgroup Bayes-optimal model for Y_i and $\mathbb{E}[Y \mid h^*_A(Z)]$ monotonically increases as a function of $h^*_A(Z)$ [7, 31, 60]. Under these criteria for (un)biasedness, fitting a subgroup Bayes-optimal model for a biased proxy induces sufficiency violation with respect to the true label, which can be detected through sufficiency violation with respect to an unbiased proxy, thus implying that efforts taken to improve sufficiency fairness with respect to a biased proxy label, such as by fitting separate models for each subgroup, can directly introduce unfairness with respect to true label.

We study these phenomena in the context of a restricted set of causal graphs intended to reflect a set of settings illustrative, but not exhaustive, of data generating processes and measurement mechanisms relevant to label bias. We consider two key classes of DAGs defined in terms of whether the proxy of interest is causally downstream ($Y \longrightarrow Y_i$; Figure 1) or upstream ($Y_i \longrightarrow Y$; Figure 2) of the true outcome Y. For all settings considered, we assume that X is a causal parent of Y_j with no effect on Y_k unmediated by Y_j, where Y_j and Y_k are the first and second of $\{Y, Y_i\}$ in the causal ordering. Within the context of the DAGs considered, we consider a simple causal graphical criterion for label bias based on the presence or absence of a direct effect (or backdoor confounding path) between A on Y_k unmediated by Y_j. Expressed in terms of conditional independence criteria implied by d-separation, we say that if Y is a causal parent of Y_i, then Y_i is a a *graphically* unbiased proxy of Y if $Y_i \perp\!\!\!\perp A \mid Y$ and a graphically biased proxy if $Y_i \not\perp\!\!\!\perp A \mid Y$; if Y_i is a causal parent of Y, then Y_i is a graphically unbiased *surrogate* proxy of Y if $Y \perp\!\!\!\perp A \mid Y_i$ and a graphically biased surrogate proxy if $Y \not\perp\!\!\!\perp A \mid Y_i$. We note that in this context, we refer to proxies as being *graphically (un)biased* to differentiate these candidate causal graphical criteria for label bias from the statistical criteria for label bias related to sufficiency. We refer to upstream proxies as surrogate proxies due to the connection to the surrogate outcomes setting [4, 63].

In Figure 1, we depict illustrative causal DAGs corresponding to settings where a set of proxy labels $\{Y_1, Y_2, Y_3\}$ are causal children of the outcome of interest Y and verify properties of models trained on data sampled from these causal graphs in a simulation study in Section 5. We consider all three downstream proxies to be graphically unbiased in Figure 1a,b. In Figure 1c,d we consider Y_3 as a graphically biased proxy, and represent this through a bidirected red arrow between A and Y_3, corresponding to the presence of an



(a) Unbiased surrogate proxy labels ($Y \perp\!\!\!\perp A \mid Y_1$ and $Y_2 \perp\!\!\!\perp A \mid Y$) and $Y_1 \perp\!\!\!\perp A \mid X$.

(b) Unbiased surrogate proxy labels ($Y \perp\!\!\!\perp A \mid Y_1$ and $Y_2 \perp\!\!\!\perp A \mid Y$) and $Y_1 \not\perp\!\!\!\perp A \mid X$.

(c) Y_1 is a biased surrogate proxy ($Y \not\perp\!\!\!\perp A \mid Y_1$), Y_2 is an unbiased proxy ($Y_2 \perp\!\!\!\perp A \mid Y$), and $Y_1 \perp\!\!\!\perp A \mid X$.

(d) Y_1 is a biased surrogate proxy ($Y \not\perp\!\!\!\perp A \mid Y_1$), Y_2 is an unbiased proxy ($Y_2 \perp\!\!\!\perp A \mid Y$), and $Y_1 \not\perp\!\!\!\perp A \mid X$.

(e) Y_1 is a unbiased surrogate proxy ($Y \perp\!\!\!\perp A \mid Y_1$), Y_2 is a biased proxy ($Y_2 \not\perp\!\!\!\perp A \mid Y$), and $Y_1 \perp\!\!\!\perp A \mid X$.

(f) Y_1 is a unbiased surrogate proxy ($Y \perp\!\!\!\perp A \mid Y_1$), Y_2 is a biased proxy ($Y_2 \not\perp\!\!\!\perp A \mid Y$), and $Y_1 \not\perp\!\!\!\perp A \mid X$.

(g) Y_1 is a biased surrogate proxy ($Y \not\perp\!\!\!\perp A \mid Y_1$), Y_2 is a biased proxy ($Y_2 \not\perp\!\!\!\perp A \mid Y$), and $Y_1 \perp\!\!\!\perp A \mid X$.

(h) Y_1 is a biased surrogate proxy ($Y \not\perp\!\!\!\perp A \mid Y_1$), Y_2 is a biased proxy ($Y_2 \not\perp\!\!\!\perp A \mid Y$), and $Y_1 \not\perp\!\!\!\perp A \mid X$.

Figure 2: Example causal DAG with the unobserved outcome of interest Y represented by the gray node and two observed proxy labels, Y_1 and Y_2. Y_1 is a *surrogate* proxy label upstream of the unobserved outcome of interest Y ($Y_1 \rightarrow Y$). Y_2 is a proxy label downstream of the unobserved outcome of interest Y ($Y \rightarrow Y_2$). Bidirected edges represent confounding between variables. A red edge is used to indicate a biased proxy relationship. Y_1 is an unbiased surrogate proxy of Y in (a), (b), (e) and (f) while it is a biased proxy of Y in (c), (d), (g) and (h). Y_2 is an unbiased proxy in (a), (b), (c), (d) and a biased proxy in (e), (f), (g), (h).

unobserved confounder with influence on the proxy Y_3, unmediated by Y, that differs in distribution across subgroups. In Figure 2, we show a setting that incorporates a surrogate outcome Y_1 that is a causal parent of Y and a downstream proxy Y_2 that is a causal child of Y. Here, the upstream surrogate proxy Y_1 is biased when there is an edge between A and Y (Figure 2c,d,g,h) and the downstream proxy Y_2 is biased when there is an edge between A and Y_2 (Figure 2e,f,g,h).

In Table 1, we detail the relationships between the sufficiency fairness criterion with respect to proxy labels and the sufficiency fairness criterion with respect to an unobserved true label of interest for an arbitrary predictive model h, for the causal DAGs of interest. There are two main categories of properties: those that hold in general, and those require additional assumptions dependent on the claim and DAG of interest. A consequence is that the graphical notions of bias are not strong enough, on their own, to imply the ideal desiderata of proxy labels discussed above.

For the case where the proxy Y_i is downstream of Y ($Y \rightarrow Y_i$) and graphically unbiased (*e.g.*, Y_1 or Y_2 in any panel of Figure 1), sufficiency satisfaction with respect to Y necessarily implies sufficiency satisfaction with respect to Y_i. This follows from Lemma 4.2 of Dawid [23], which states that if $Y \perp\!\!\!\perp A \mid h(Z)$, then $f(Y) \perp\!\!\!\perp A \mid h(Z)$, where f is a function that depends only on Y. By a contrapositive argument, and without further assumptions, sufficiency violation with respect to Y_i implies sufficiency violation with respect to Y. However, reasoning about whether sufficiency is satisfied with respect to Y requires additional assumptions regarding the informativeness of the proxy Y_i about Y. This is required because it is possible for sufficiency violation with respect to Y to not be observable in Y_i if the mapping from Y to Y_i is noisy or such that the dependence between A and Y is masked within levels of $h(Z)$. We leave a full technical formalization of the necessary assumptions to future work, but note that the required assumptions extend causal faithfulness [69], *i.e.*, that no (conditional) independencies are present other than those implied by the DAG, with additional requirements regarding the functional forms of the $Y \rightarrow Y_i$ relationship and h. However, if the necessary assumptions hold such that $Y \not\!\perp\!\!\!\perp A \mid h(Z) \implies Y_i \not\!\perp\!\!\!\perp A \mid h(Z)$, it follows by the contrapositive that $Y_i \perp\!\!\!\perp A \mid h(Z) \implies Y \perp\!\!\!\perp A \mid h(Z)$, *i.e.*, that sufficiency with respect to Y_i implies sufficiency with respect to Y.

To reason about the properties of downstream, graphically biased proxies (Y_3 in Figure 1c,d), we note that additional assumptions beyond the presence of a direct or backdoor path between A and graphically biased Y_i is needed in order to conclude that sufficiency satisfaction with respect to Y_i implies sufficiency violation with respect to Y ($Y_i \perp\!\!\!\perp A \mid h(Z) \implies Y \not\!\perp\!\!\!\perp A \mid h(Z)$). As before, we note that these assumptions are related to causal faithfulness, but leave further formalization for future work. For intuition, consider the contrapositive statement, where sufficiency satisfaction with respect to Y implies sufficiency violation with respect to Y_i ($Y \perp\!\!\!\perp A \mid h(Z) \implies Y_i \not\!\perp\!\!\!\perp A \mid h(Z)$), which may be false if the effect between A and Y_i unmediated by Y does not introduce dependence within levels of $h(Z)$. However, if the necessary assumptions related to faithfulness hold, we can conclude that fitting a model that satisfies sufficiency with respect to Y_i will violate sufficiency with respect to Y. Furthermore, we note violation of sufficiency with respect to a biased downstream proxy does not imply satisfaction of sufficiency with respect to the true outcome of interest.

For the case of upstream surrogate proxy labels ($Y_i \rightarrow Y$), the results are analogous to the case of downstream proxy labels, with the role of Y_i and Y reversed (Table 1). For instance, satisfaction of sufficiency with respect to a graphically unbiased proxy Y_i implies sufficiency is satisfied with respect to Y, but further faithfulness assumptions are required for violation of sufficiency with respect to Y_i to imply sufficiency violation with respect to Y. Furthermore, when the proxy Y_i is graphically biased, due to the presence of an effect between A and Y unmediated by Y_i, a faithfulness assumption is required for sufficiency violation with respect to Y_i to imply sufficiency violation with respect to Y.

4 RELATED WORK

4.1 Fairness and proxy label bias

Prior work has investigated the fairness implications of biased proxy labels in different settings. This challenge has been extensively documented in judicial [14, 29], child welfare [18, 48] and hiring [16] settings. For example, in a hiring setting, past performance reviews have been considered as a proxy for future job performance [59], while defendant re-arrest may be considered as a proxy for recidivism risk in criminal justice settings [8, 30]. Furthermore, differential selection bias and censoring across subgroups are well-documented phenomena in these settings and carry similar implications as label bias [47].

One approach to studying bias in proxy labels is to use causal directed acyclic graphs as a form of measurement model to encode assumptions about the relationship between the proxy and the true label [43]. Guerdan et al. [37] provides a review of this approach and presents a generic causal framework that can be used to reason about several forms of bias relevant to use of proxy labels for decision-making. Our work is complementary to that of Guerdan et al. [37] in that we study fairness properties implied by the "group-dependent error" setting discussed in that work. Our work is further related to that of Guerdan et al. [36], as they study counterfactual prediction with outcome measurement error and include a re-analysis of the data of Obermeyer et al. [56] to study differences in outcome measurement error between the enrolled and unenrolled populations. Our analysis of Obermeyer et al. [56] differs as we primarily contextualize the implications of label bias in terms of its impact on the sufficiency fairness criterion with respect to a true label and proxy variables.

Several studies propose approaches to correct for label bias during model development that are well-motivated when the structure of the bias is known. For example, Menon et al. [52] and Wang et al. [74] show that it is possible to learn an unbiased estimator for the true label given assumptions on the measurement error mechanism. In a more general setting, De-Arteaga et al. [25] propose to address label bias by learning from observed outcomes with deferral to historical expert decisions when they display certainty and consistency. Jiang and Nachum [44] and Blum and Stangl [13] take a different approach, and use fairness constraints to recover from data biases.

Table 1: Relationship between sufficiency with respect to a binary proxy label ($Y_i \perp\!\!\!\perp A \mid h(Z)$) and a true binary outcome of interest ($Y \perp\!\!\!\perp A \mid h(Z)$). We show cases where the proxy is downstream ($Y \rightarrow Y_i$) and upstream of the true outcome ($Y_i \rightarrow Y$). Downstream and upstream proxies are graphically biased when $Y_i \not\perp\!\!\!\perp A \mid Y$ and $Y \not\perp\!\!\!\perp A \mid Y_i$, respectively, within the class of causal DAGs studied.

Setting	Property satisfied in general	Requires further assumptions
$Y \rightarrow Y_i$ and $Y_i \perp\!\!\!\perp A \mid Y$	$Y \perp\!\!\!\perp A \mid h(Z) \implies Y_i \perp\!\!\!\perp A \mid h(Z)$	$Y_i \perp\!\!\!\perp A \mid h(Z) \implies Y \perp\!\!\!\perp A \mid h(Z)$
	$Y_i \not\perp\!\!\!\perp A \mid h(Z) \implies Y \not\perp\!\!\!\perp A \mid h(Z)$	$Y \not\perp\!\!\!\perp A \mid h(Z) \implies Y_i \not\perp\!\!\!\perp A \mid h(Z)$
$Y \rightarrow Y_i$ and $Y_i \not\perp\!\!\!\perp A \mid Y$		$Y \perp\!\!\!\perp A \mid h(Z) \implies Y_i \not\perp\!\!\!\perp A \mid h(Z)$
		$Y_i \perp\!\!\!\perp A \mid h(Z) \implies Y \not\perp\!\!\!\perp A \mid h(Z)$
$Y_i \rightarrow Y$ and $Y \perp\!\!\!\perp A \mid Y_i$	$Y_i \perp\!\!\!\perp A \mid h(Z) \implies Y \perp\!\!\!\perp A \mid h(Z)$	$Y \perp\!\!\!\perp A \mid h(Z) \implies Y_i \perp\!\!\!\perp A \mid h(Z)$
	$Y \not\perp\!\!\!\perp A \mid h(Z) \implies Y_i \not\perp\!\!\!\perp A \mid h(Z)$	$Y_i \not\perp\!\!\!\perp A \mid h(Z) \implies Y \not\perp\!\!\!\perp A \mid h(Z)$
$Y_i \rightarrow Y$ and $Y \not\perp\!\!\!\perp A \mid Y_i$		$Y_i \perp\!\!\!\perp A \mid h(Z) \implies Y \not\perp\!\!\!\perp A \mid h(Z)$
		$Y \not\perp\!\!\!\perp A \mid h(Z) \implies Y_i \perp\!\!\!\perp A \mid h(Z)$

4.2 Proxy Fairness

While in this study we focus on fairness assessment using proxies for a true label, prior work has also studied fairness assessment where the true sensitive attribute is unobserved. In these cases, it is common to use a "proxy model" to predict the unobserved sensitive attribute from covariates, as a proxy of the true sensitive attribute. For example, Diana et al. [26], study the problem of training a model that obeys fairness constraints when the sensitive attributes are not available at training time. They highlight if the proxy model is unable to accurately learn the unobserved sensitive attribute, downstream fairness implications with respect to the proxy of the sensitive attribute may not hold. Gupta et al. [38] highlight that the fairness evaluation depends on not just how well the proxy group aligns with the true groups based on the sensitive attribute but also on the choice of the fairness metric. Zhu et al. [81] analyzes this issue further by proposing a theoretical framework that shows that directly using proxies for the sensitive attribute can give a false sense of (un)fairness with respect to the true unobserved sensitive attributes. In this setting, there are additional challenges when the non-sensitive covariates are highly correlated with the sensitive attributes, affecting the proxy model. Hajian and Domingo-Ferrer [39] study the issue of indirect discrimination resulting from a high-degree of correlation between the non-sensitive attributes and the unobserved sensitive attribute.

Approaches for alleviating concerns with the use of proxy models for fairness assessments have focused on causal assumptions between the proxy model and the sensitive attribute. In the context of rankings, Ghazimatin et al. [34] propose causal conditions that result in fair assessments of rankings with proxies of sensitive attributes. Specifically, if the ranking score is conditionally independent of the unobserved sensitive attribute given its proxy, the rankings would be fair. While this approach relates to the causal assumptions we propose between the proxy of the unobserved true outcome and the sensitive attribute, our assumptions about the proxies are different than those of Ghazimatin et al. [34], instead focusing on proxies of the unobserved true label. Kilbertus et al. [49] also study fairness assessment in settings with proxies of sensitive attributes, where the proxies are causal descendants of the sensitive attribute.

4.3 Surrogate Outcomes

In the causal inference literature, it is common to reason about estimation of the effect of a treatment on a long-term outcome using a short-term "surrogate" outcome [1–3, 11, 22, 24, 28, 35, 42, 46, 63, 64, 71, 75, 77, 79]. A key causal assumption required for the use of the surrogate in place of the true unobserved outcome, *i.e.*, for a variable to be a valid surrogate, is that the long-term outcome is independent of treatment conditional on the surrogate, which is often called as the surrogacy assumption [11, 32]. This is related to our graphical notion of unbiasedness in the case of a proxy label that is causally-upstream of a true outcome of interest.

Freedman et al. [33] highlight a major concern regarding the use of surrogates, where biased estimates of the effect of the treatment on the true outcome may result if the full effect of the treatment on the true outcome is not completely mediated by the surrogate. Furthermore, Frangakis and Rubin [32], Joffe and Greene [45], Rosenbaum [66], VanderWeele [73] postulate that if there is unmeasured confounding between the surrogate and the true outcome, the surrogacy assumption would still be invalid even if the treatment has no direct effect on the true outcome and the entire effect of the treatment on the true outcome is completely mediated by the surrogate. Considering the issues that can manifest from using a single surrogate, Athey et al. [4] consider the use of multiple proxies or surrogates in place of the true outcome. They propose learning a surrogate index based on multiple surrogates, showing that the average treatment effect on the surrogate index equals the treatment effect on the long-term outcome under the assumption that the long-term outcome is independent of the treatment conditional on the surrogate index. Furthermore, Athey et al. [4] utilize the methodology of Bibaut et al. [12] that further allows to proxy for effects not perfectly mediated by the surrogates, to handle both confounding and effect leakage as violations of standard statistical surrogacy conditions.

Our assumptions about the conditional independence between the true outcome and the sensitive attribute are also related to the literature on mediation and missing data. The mediation literature [10, 41, 70, 72, 73, 80] decomposes the average treatment effect into the direct effect of a treatment on an outcome and indirect effects that flow through a mediator. In the case of surrogate outcomes, the surrogate plays the role of the mediator and aligns with the conditional independence assumptions we consider between the true unobserved outcome, the surrogate (proxy label) and the sensitive attribute. Moreover, rather than focusing on merely the absence of the direct effect of the treatment on the true outcome to ensure that the surrogate proxy is unbiased, we also study settings analogous to the case where the true outcome and the treatment could be confounded.

5 EXPERIMENTS

5.1 Simulation study

To verify the properties studied in Section 3, we conduct a simulation study to investigate how the causal structure of the data generating processes influences the relationship between sufficiency assessed with respect to proxy labels and a true label of interest for subgroup-specific predictive models. These settings correspond to cases where (a) all proxies are downstream and we observe both biased and unbiased proxies (Figure 1d); (b) an upstream proxy is biased and a downstream proxy is unbiased (Figure 2d); (c) an upstream proxy is unbiased and a downstream proxy is biased (Figure 2f); and (d) both upstream and downstream proxies are biased (Figure 2h). The data generating processes are illustrated in Figure 3.

For each data generating process, we samples 10,000 instances and use half for training and half for evaluation. In all cases, we fit separate models for each target variable $\{Y_1, Y_2, Y_3, Y\}$ using both X and A as inputs. We use the Scikit-learn v1.4.0 [58] implementation of gradient boosting (`HistGradientBoostingClassifier`) on the training data with default hyperparameters. For evaluation, we generate calibration curves via logistic regression fit on the held-out test set to estimate the probability of each target variable conditioned on the logit of the model output.

For the case where the proxies are downstream of the true outcome (Figure 3a), Y_1 and Y_2 are unbiased proxies of Y, with Y_2 being a noisier proxy (*i.e.*, with a structural equation with an error term with a greater standard deviation) and Y_3 is a biased proxy of Y with an additional effect of A on Y_3 that is not mediated by Y ($Y_3 \not\perp A \mid Y$), consistent with Figure 1d. The results of this evaluation on the simulated data are shown in Figure 4 and are generally consistent with the analytical relationships reported in Table 1. We find that the model trained on the biased proxy Y_3 satisfies sufficiency with respect to Y_3 but violates it with respect to the true outcome Y and the unbiased proxies Y_1 and Y_2. The magnitude of the sufficiency violation is larger for the less noisy proxy Y_1 than it is for Y_2. Training on Y results in sufficiency satisfaction for the unbiased proxies Y_1 and Y_2. Training on Y or either of unbiased proxies results in sufficiency violation with respect to the biased proxy Y_3.

We also assess various settings with multiple proxy labels where one of the proxy labels is upstream of the true outcome while the other proxy label is downstream of the true label. First, we generate simulated data according to the data-generating process outlined in Figure 3b, for the controlled experiment where the proxy label upstream of the true outcome is a biased proxy whereas the proxy label downstream of the true outcome is an unbiased proxy. Here, Y_1, which is upstream of Y, is a biased proxy of Y, whereas Y_2, which is downstream of Y, is an unbiased proxy. Figure 5 illustrates the results for this evaluation. The results are generally consistent with the analytical relationships reported in Table 1. The model trained on the biased proxy Y_1 violates sufficiency with respect to the true outcome Y and the unbiased proxy Y_2, while satisfying sufficiency with respect to Y_1. Training on Y results in sufficiency satisfaction with respect to the unbiased proxy Y_2 but violates sufficiency with respect to the biased proxy Y_1. Similarly, training on the unbiased proxy Y_2 results in sufficiency satisfaction with respect to the true outcome Y but not with respect to the biased proxy Y_1.

We also assess a complimentary setting to the previous simulation, where the proxy label upstream of the true outcome, Y_1, is the unbiased proxy while the proxy label downstream of the true outcome, Y_3, is biased, corresponding to the causal graph in Figure 2f. The data-generating process for this setup is described in Figure 3c. The model trained on the unbiased proxy Y_1 satisfies sufficiency with respect to the true outcome Y but violates sufficiency with respect to the biased proxy Y_2, as shown in Supplementary Figure A1. Training on Y results in sufficiency satisfaction with respect to the unbiased proxy Y_1 but violates sufficiency with respect to the biased proxy, Y_2. Whereas, training on Y_2 results is sufficiency violation with respect to both Y_1 and Y.

Finally, we also assess a setting where both proxy labels, upstream and downstream of the true outcome are biased. The simulated data is generated according to the process described in Figure 3(d) corresponding to the causal graph in Figure 2(h). In this setup, training on Y results in sufficiency satisfaction only with respect to Y but violation with respect to both biased proxy labels, Y_1 and Y_2 as illustrated in Supplementary Figure A2. Moreover, training on either of the biased proxy results in sufficiency violation with respect to the true outcome of interest.

Overall, we verify in simulation that a model that is apparently fair with respect to a biased proxy may not be fair with respect to the true outcome. Furthermore, we find that unbiased proxy labels sampled from these data generating processes generally reproduce the sufficiency characteristics of the true label of interest, such that violation of sufficiency with respect to an unbiased proxy provides evidence of sufficiency violation with respect to the true outcome of interest.

5.2 Label bias with synthetic health insurance data

Here, we conduct experiments with publicly-available synthetic health insurance data released by Obermeyer et al. [56]. This data matches the size and structure of the data used in the original study, such that claims from the original work can be reproduced using the synthetic data. The synthetic data represents data from 48,784 patient-years with 160 pre-processed variables indicating demographics, comorbidities, medications, laboratory orders and results, and healthcare expenditures. Following the formulation of the Obermeyer et al. [56] study, we consider "medical expenditure" to be a biased proxy and "chronic conditions" to be an unbiased proxy of true label of interest, "health status". In this setting, the 97th percentile of the risk score is used for direct enrollment in a care management program and the 55th percentile is used for potential enrollment following a consultation between the clinician

$$\begin{aligned}
&A \sim \text{Bernoulli}\,(0.5)\\
&X \sim \mathcal{N}(0,3)\\
&Y \sim \text{Bernoulli}\,(\sigma\,((1-2A)X + (2A-1)))\\
&Y_1 = \text{Bernoulli}\,(\sigma\,(Y + \mathcal{N}(0,0.5))\\
&Y_2 = \text{Bernoulli}\,(\sigma\,(0.5Y + \mathcal{N}(0,4))\\
&Y_3 = \text{Bernoulli}\,(\sigma\,(1.5Y + \tan(-2A)))\\
&\sigma(x) = 1/(1+\exp(-x))
\end{aligned}$$

(a) Data generating process based on Figure 1d, where Y_1 and Y_2 are downstream unbiased proxies and Y_3 is a downstream biased proxy.

$$\begin{aligned}
&A \sim \text{Bernoulli}\,(0.5)\\
&X \sim \mathcal{N}(0,3)\\
&Y_1 \sim \text{Bernoulli}\,(\sigma\,((1-2A)X + (2A-1)))\\
&Y = \text{Bernoulli}\,(\sigma\,(1.5Y_1 + \tan(-2A)))\\
&Y_2 = \text{Bernoulli}\,(\sigma\,(1.5Y))\\
&\sigma(x) = 1/(1+\exp(-x))
\end{aligned}$$

(b) Data generating process based on Figure 2d, where Y_1 is an upstream biased surrogate proxy and Y_2 is a downstream unbiased proxy.

$$\begin{aligned}
&A \sim \text{Bernoulli}\,(0.5)\\
&X \sim \mathcal{N}(0,3)\\
&Y_1 \sim \text{Bernoulli}\,(\sigma\,((1-2A)X + (2A-1)))\\
&Y = \text{Bernoulli}\,(\sigma\,(1.5Y_1))\\
&Y_2 = \text{Bernoulli}\,(\sigma\,(1.5Y + \tan(-2A)))\\
&\sigma(x) = 1/(1+\exp(-x))
\end{aligned}$$

(c) Data generating process based on Figure 2f, where Y_1 is an upstream unbiased surrogate proxy and Y_2 is a downstream biased proxy.

$$\begin{aligned}
&A \sim \text{Bernoulli}\,(0.5)\\
&X \sim \mathcal{N}(0,3)\\
&Y_1 \sim \text{Bernoulli}\,(\sigma\,((1-2A)X + (2A-1)))\\
&Y = \text{Bernoulli}\,(\sigma\,(1.5Y_1 + \tan(-2A)))\\
&Y_2 = \text{Bernoulli}\,(\sigma\,(1.5Y + \tan(-2A)))\\
&\sigma(x) = 1/(1+\exp(-x))
\end{aligned}$$

(d) Data generating process based on Figure 2h, where Y_1 is an upstream biased proxy and Y_2 is a downstream biased proxy.

Figure 3: Data generating process based on the causal graphs from Figure 1d in (a), Figure 2d in (b), Figure 2f in (c), and Figure 2h in (d).

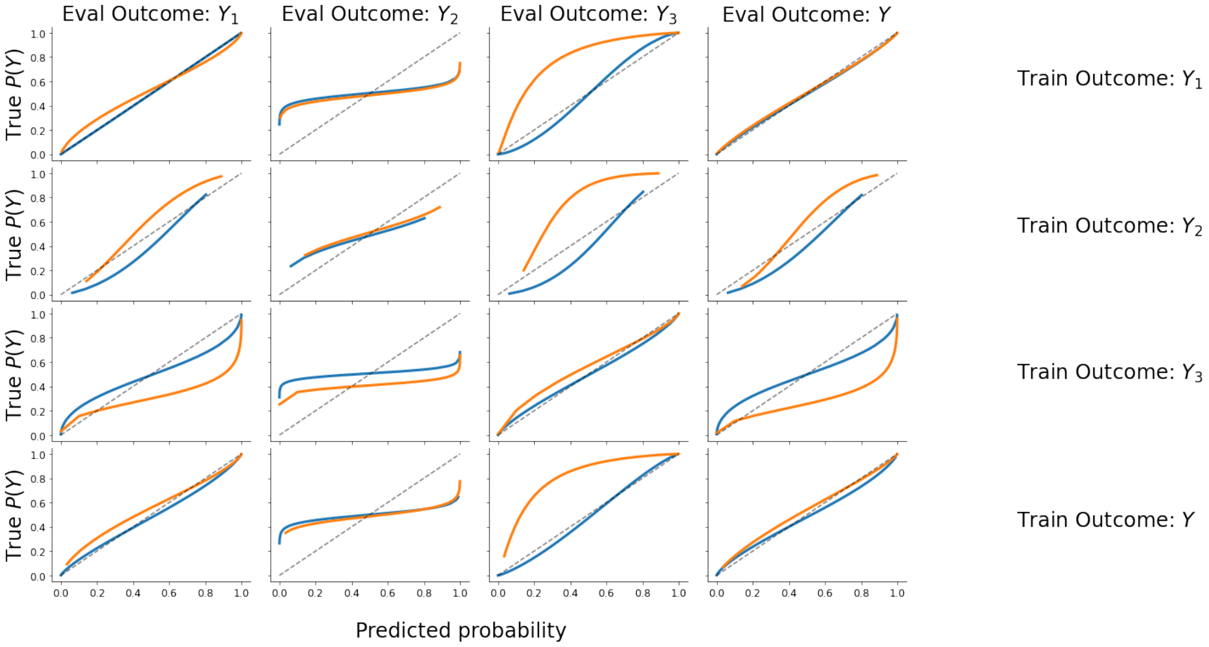


Figure 4: Sufficiency characteristics of models trained on proxies and true outcomes generated according to the data generating process described in Figure 1d and Figure 3a, evaluated across two subgroups represented by orange and blue lines. In this setting, all proxies are downstream of the outcome Y, where Y_1 and Y_2 are unbiased proxies, and Y_3 is a biased proxy.

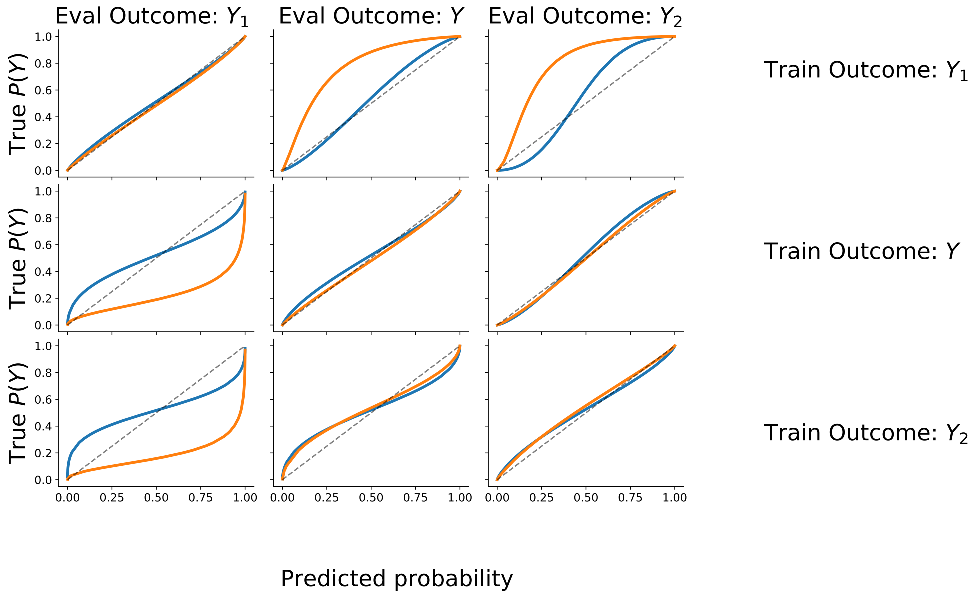


Figure 5: Sufficiency characteristics for models trained with proxies and true outcome generated according to the data generating process described in Figure 2d and Figure 3b, evaluated across two subgroups represented by orange and blue lines. In this setting, the upstream proxy Y_1 is biased and the downstream proxy Y_2 is unbiased.

and patient. Here, we use the 55th percentile as the threshold of interest.

In order to study the effect of training on the biased proxy, as is done in the original [56] study, we follow a similar training strategy involving a `lasso` model trained on the "medical expenditure" proxy label. The original training algorithm did not include the subgroup attribute, 'race'; that is replicated here as well. Accordingly, we evaluate the bias of a race-agnostic model on both medical expenditure and chronic conditions. Here, sufficiency is satisfied with respect to neither of the proxies.

Figure 6a presents the total medical expenditure values as a function of the percentile of predicted algorithmic risk score. For a particular percentile of the algorithmic risk score (*e.g.*, consider the 55th percentile) the medical expenditure is higher for the White patients in comparison to Black patients. When the same model is evaluated on the unbiased proxy, "chronic conditions", Black patients have a greater number of chronic conditions in comparison to the White patients at the same predicted risk score (Figure 6b).

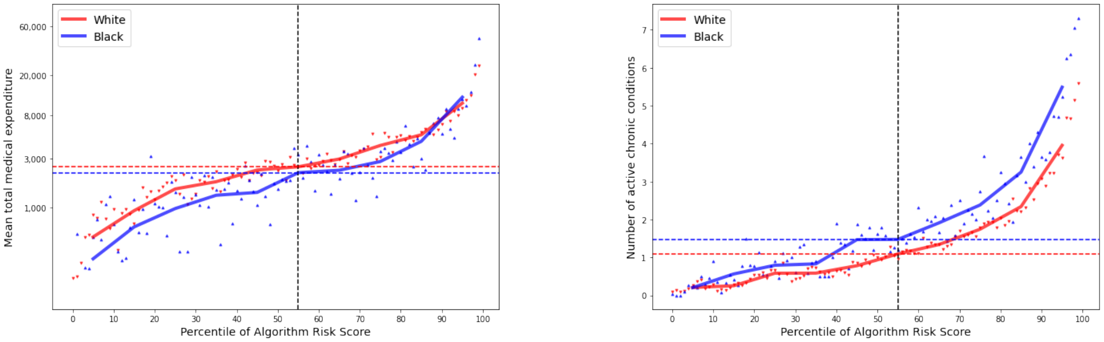
In the absence of awareness that healthcare expenditure is a biased proxy, it may be of interest to mitigate sufficiency violation with respect to it by incorporating the subgroup attribute into the predictor. Accordingly, we also evaluate the sufficiency characteristics of a model that includes the race attribute. Figure 7 presents the average medical expenditure conditioned on the percentile of the predicted risk score produced by such a model. At the 55th percentile, the medical expenditure for White patients is equal to that of Black patients, which may tempt one to conclude that the predicted risk score will lead to downstream fair decision making. However, when we evaluate the same model for the unbiased proxy, chronic conditions, we observe that sufficiency is violated because the number of chronic conditions are not the same across Black and White patients conditioned on the risk score. Furthermore, when we compare the sufficiency characteristics of the race-agnostic and race-dependent models the gap between the number of chronic conditions for the two groups is wider when the model is race-dependent.

Overall, our experiments with synthetic health insurance data are consistent with the theoretical causal perspective on label bias we develop in section 3 and with the simulation results. We also observe that mitigating apparent fairness violation with respect to the biased proxy (medical expenditure) introduces further fairness violation with respect to the unbiased proxy (chronic conditions).

6 DISCUSSION

Label bias is a critical issue for fairness assessments with proxy labels. Our theoretical construction includes candidate graphical criteria for bias in proxy labels that depend on the causal structure of data generating process, *i.e.*, whether the proxy is upstream or downstream of the true label of interest, and connects them to conditions under which violation or satisfaction of sufficiency in unbiased and biased proxies that are upstream or downstream of the true label may be used to reason about sufficiency with respect to the true label.

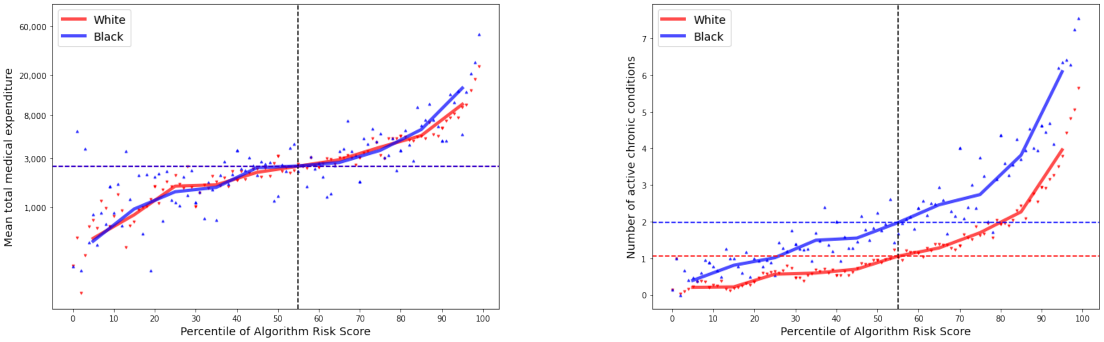
Our theoretical presentation serves to potentially explain empirical phenomena related to the model calibration and sufficiency fairness properties in settings with label bias, such as those reported in Obermeyer et al. [56]. In our simulation study, we observe trends that qualitatively match those implied by our theory. The results of our replication of the Obermeyer et al. [56] analysis are further consistent with what would be expected in a setting with both



(a)

(b)

Figure 6: Average total medical expenditure (A) and number of active chronic conditions (B) vs. the percentile of the algorithm risk score for h_{θ_A}, the model that does not incorporate subgroup membership to predict medical expenditure. Conditioned on the risk score, mean medical expenditure is higher for white patients than for black patients, and the number of active chronic conditions is higher for black patients than for white patients.



(a)

(b)

Figure 7: Average total medical expenditure (A) and number of active chronic conditions (B) vs. the percentile of the algorithm risk score for h_{θ_A}, the model incorporate subgroup membership to predict medical expenditure. Sufficiency is approximately satisfied with respect to medical expenditure, but violated for the number of active chronic conditions, with a larger gap as compared to the case where subgroup membership is not used for prediction (Figure 6B).

unbiased and biased proxies. Overall, our empirical results are consistent with expectations. We find that evaluation of fairness with respect to biased proxy labels masks fairness violation with respect to the true label of interest. Furthermore, in cases where a model is apparently unfair with respect to unbiased proxy label, we find that mitigation of sufficiency fairness violation through incorporation of subgroup information as a predictor worsens fairness violation with respect to the true label, and that this effect can be measured through an unbiased proxy of the true label. This highlights that the appropriateness of incorporating subgroup information into the model may depend on the presence of bias in the proxy label used for prediction.

This work opens up important directions for future work. It is an open question as to how to identify whether a proxy is biased in the absence of the domain knowledge required to define an appropriate causal graph, which may be challenging to obtain in scenarios with insufficient domain expertise available, or when the causal graph is large and complex. Our approach further requires the availability of at least one unbiased proxy in order to identify a biased proxy, and it is unclear whether and how biased proxies can be identified in cases where no auxiliary unbiased proxy is available. Furthermore, our work introduces opportunities to further formalize the necessary assumptions and conditions that allow for meaningful inference regarding the label of interest using proxy labels.

ACKNOWLEDGMENTS

This study is funded by Google.

REFERENCES

[1] Odd O Aalen. 2004. Discussion of Causality. *Scandinavian Journal of Statistics* 31, 2 (2004), 193–195.
[2] Ariel Alonso, Geert Molenberghs, Helena Geys, Marc Buyse, and Tony Vangeneugden. 2006. A unifying approach for surrogate marker validation based on Prentice's criteria. *Statistics in medicine* 25, 2 (2006), 205–221.
[3] Susan Athey, Raj Chetty, and Guido Imbens. 2020. Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv*

preprint arXiv:2006.09676 (2020).

[4] Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2019. *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely.* Technical Report. National Bureau of Economic Research.

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

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

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

[8] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498* (2021).

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities.* MIT Press.

[10] Reuben M Baron and David A Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* 51, 6 (1986), 1173.

[11] Colin B Begg and Denis HY Leung. 2000. On the use of surrogate end points in randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 163, 1 (2000), 15–28.

[12] Aurélien Bibaut, Nathan Kallus, Simon Ejdemyr, and Michael Zhao. 2023. Long-Term Causal Inference with Imperfect Surrogates using Many Weak Experiments, Proxies, and Cross-Fold Moments. *arXiv preprint arXiv:2311.04657* (2023).

[13] Avrim Blum and Kevin Stangl. 2019. Recovering from biased data: Can fairness constraints improve accuracy? *arXiv preprint arXiv:1912.01094* (2019).

[14] Bradley Butcher, Chris Robinson, Miri Zilka, Riccardo Fogliato, Carolyn Ashurst, and Adrian Weller. 2022. Racial Disparities in the Enforcement of Marijuana Violations in the US. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society.* 130–143.

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

[16] Aaron Chalfin, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. Productivity and selection of human capital with machine learning. *American Economic Review* 106, 5 (2016), 124–127.

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

[18] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How child welfare workers reduce racial disparities in algorithmic decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–22.

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

[20] Sam Corbett-Davies, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2023. The measure and mismeasure of fairness. *The Journal of Machine Learning Research* 24, 1 (2023), 14730–14846.

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

[22] Ralph B D'Agostino, Michael J Campbell, and Joel B Greenhouse. 2006. Surrogate markers: back to the future: Special Papers for the 25th Anniversary of Statistics in Medicine. *Statistics in Medicine* 25, 2 (2006), 181–182.

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

[24] Nicholas E Day and Stephen W Duffy. 1996. Trial design based on surrogate end points—application to comparison of different breast screening frequencies. *Journal of the Royal Statistical Society Series A: Statistics in Society* 159, 1 (1996), 49–60.

[25] Maria De-Arteaga, Vincent Jeanselme, Artur Dubrawski, and Alexandra Chouldechova. 2021. Leveraging expert consistency to improve algorithmic decision support. *arXiv preprint arXiv:2101.09648* (2021).

[26] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2022. Multiaccurate proxies for downstream fairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* 1207–1239.

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

[28] Thomas R Fleming and David L DeMets. 1996. Surrogate end points in clinical trials: are we being misled? *Annals of internal medicine* 125, 7 (1996), 605–613.

[29] Riccardo Fogliato, Alexandra Chouldechova, and Max G'Sell. 2020. Fairness evaluation in presence of biased noisy labels. In *International conference on artificial intelligence and statistics.* PMLR, 2325–2336.

[30] Riccardo Fogliato, Alice Xiang, Zachary Lipton, Daniel Nagin, and Alexandra Chouldechova. 2021. On the validity of arrest as a proxy for offense: Race and the likelihood of arrest for violent crimes. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* 100–111.

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

[32] Constantine E Frangakis and Donald B Rubin. 2002. Principal stratification in causal inference. *Biometrics* 58, 1 (2002), 21–29.

[33] Laurence S Freedman, Barry I Graubard, and Arthur Schatzkin. 1992. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in medicine* 11, 2 (1992), 167–178.

[34] Azin Ghazimatin, Matthaus Kleindessner, Chris Russell, Ziawasch Abedjan, and Jacek Golebiowski. 2022. Measuring fairness of rankings under noisy sensitive information. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* 2263–2279.

[35] Peter B Gilbert and Michael G Hudgens. 2008. Evaluating candidate principal surrogate endpoints. *Biometrics* 64, 4 (2008), 1146–1154.

[36] Luke Guerdan, Amanda Coston, Kenneth Holstein, and Zhiwei Steven Wu. 2023. Counterfactual Prediction Under Outcome Measurement Error. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.* 1584–1598.

[37] Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. 2023. 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.* 688–704.

[38] Maya R. Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Lutong Wang. 2018. Proxy Fairness. *ArXiv* abs/1806.11212 (2018).

[39] Sara Hajian and Josep Domingo-Ferrer. 2012. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering* 25, 7 (2012), 1445–1459.

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

[41] Kosuke Imai, Luke Keele, and Dustin Tingley. 2010. A general approach to causal mediation analysis. *Psychological methods* 15, 4 (2010), 309.

[42] Guido Imbens, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. 2022. Long-term causal inference under persistent confounding via data combination. *arXiv preprint arXiv:2202.07234* (2022).

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

[44] Heinrich Jiang and Ofir Nachum. 2020. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics.* PMLR, 702–712.

[45] Marshall M Joffe and Tom Greene. 2009. Related causal frameworks for surrogate outcomes. *Biometrics* 65, 2 (2009), 530–538.

[46] Nathan Kallus and Xiaojie Mao. 2020. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408* (2020).

[47] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning.* PMLR, 2439–2448.

[48] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving human-AI partnerships in child welfare: understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–18.

[49] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems* 30 (2017).

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

[51] Zea Malawa, Jenna Gaarde, and Solaire Spellen. 2021. Racism as a root cause approach: a new framework. *Pediatrics* 147, 1 (2021).

[52] Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. 2015. Learning from corrupted binary labels via class-probability estimation. In *International conference on machine learning.* PMLR, 125–134.

[53] Vishwali Mhasawade and Rumi Chunara. 2021. Causal multi-level fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 784–794.

[54] Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. 2021. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence* 3, 8 (2021), 659–666.

[55] Sendhil Mullainathan and Ziad Obermeyer. 2021. On the inequity of predicting a while hoping for B. In *AEA Papers and Proceedings*, Vol. 111. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 37–42.

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

[57] Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge University-Press* (2000).

[58] 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. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[59] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What you see is what you get? the impact of representation criteria on human bias in hiring. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 125–134.

[60] Stephen Pfohl, Yizhe Xu, Agata Foryciarz, Nikolaos Ignatiadis, Julian Genkins, and Nigam Shah. 2022. 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*. 1039–1052.

[61] Stephen Robert Pfohl, Natalie Harris, Chirag Nagpal, David Madras, Vishwali Mhasawade, Olawale Elijah Salaudeen, Katherine A Heller, Sanmi Koyejo, and Alexander Nicholas D'Amour. 2023. Understanding subgroup performance differences of fair predictors using causal models. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*. https://openreview.net/forum?id=Fd00jISBD0

[62] Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. 2020. A large-scale analysis of racial disparities in police stops across the United States. *Nature human behaviour* 4, 7 (2020), 736–745.

[63] Ross L Prentice. 1989. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine* 8, 4 (1989), 431–440.

[64] Yongming Qu and Michael Case. 2006. Quantifying the indirect treatment effect via surrogate markers. *Statistics in medicine* 25, 2 (2006), 223–231.

[65] Eliane Röösli, Selen Bozkurt, and Tina Hernandez-Boussard. 2022. Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model. *Scientific Data* 9, 1 (2022), 24.

[66] Paul R Rosenbaum. 1984. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society Series A: Statistics in Society* 147, 5 (1984), 656–666.

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

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

[69] Peter Spirtes, Clark N Glymour, and Richard Scheines. 2000. *Causation, prediction, and search*. MIT press.

[70] Eric J Tchetgen Tchetgen and Ilya Shpitser. 2014. Estimation of a semiparametric natural direct effect model incorporating baseline covariates. *Biometrika* 101, 4 (2014), 849–864.

[71] Allen Tran, Aurélien Bibaut, and Nathan Kallus. 2023. Inferring the Long-Term Causal Effects of Long-Term Treatments from Short-Term Experiments. *arXiv preprint arXiv:2311.08527* (2023).

[72] Mark J van der Laan and Maya L Petersen. 2004. Estimation of direct and indirect causal effects in longitudinal studies. (2004).

[73] Tyler VanderWeele. 2015. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

[74] Jialu Wang, Yang Liu, and Caleb Levy. 2021. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 526–536.

[75] Christopher J Weir and Rosalind J Walley. 2006. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Statistics in medicine* 25, 2 (2006), 183–203.

[76] DR Williams and C Collins. 2001. Racial Residential Segregation: A Fundamental Cause of Racial Disparities in Health. *Public Health Reports* 116, September/October (2001).

[77] Jane Xu and Scott L Zeger. 2001. The evaluation of multiple surrogate endpoints. *Biometrics* 57, 1 (2001), 81–87.

[78] Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. 2022. Improving the fairness of chest x-ray classifiers. In *Conference on health, inference, and learning*. PMLR, 204–233.

[79] Vickie Zhang, Michael Zhao, Anh Le, and Nathan Kallus. 2023. Evaluating the Surrogate Index as a Decision-Making Tool Using 200 A/B Tests at Netflix. *arXiv preprint arXiv:2311.11922* (2023).

[80] Wenjing Zheng and Mark J van der Laan. 2012. Targeted maximum likelihood estimation of natural direct effects. *The international journal of biostatistics* 8, 1 (2012), 1–40.

[81] Zhaowei Zhu, Yuanshun Yao, Jiankai Sun, Hang Li, and Yang Liu. 2023. Weak proxies are sufficient and preferable for fairness with missing sensitive attributes. In *International Conference on Machine Learning*. PMLR, 43258–43288.

A APPENDIX

A.1 Additional simulation results

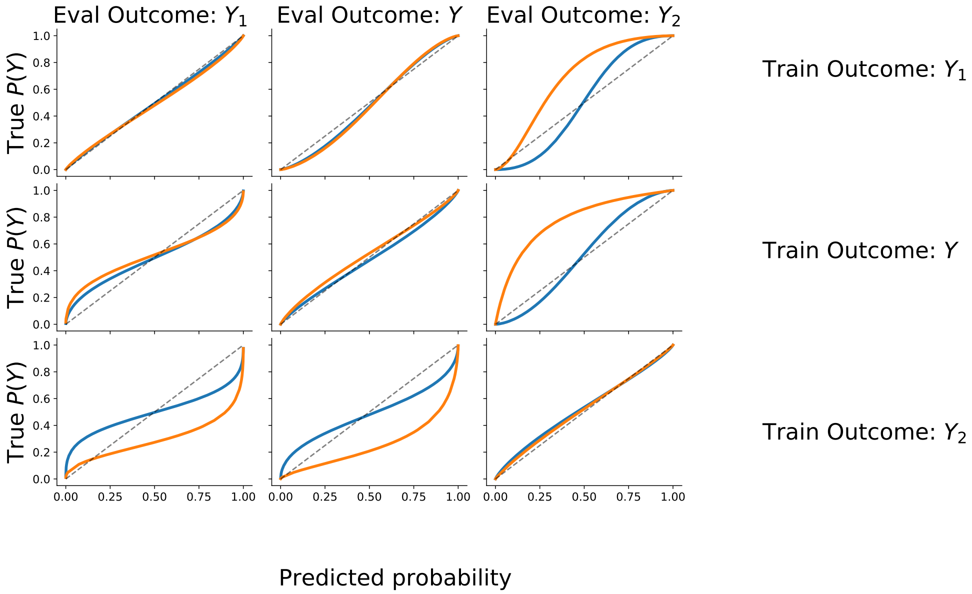


Figure A1: Sufficiency characteristics for models trained with proxies and true outcome generated according to the data generating process described in Figure 2f and Figure 3c, evaluated across two subgroups represented by orange and blue lines. In this setting, the upstream proxy Y_1 is unbiased and the downstream proxy Y_2 is biased.

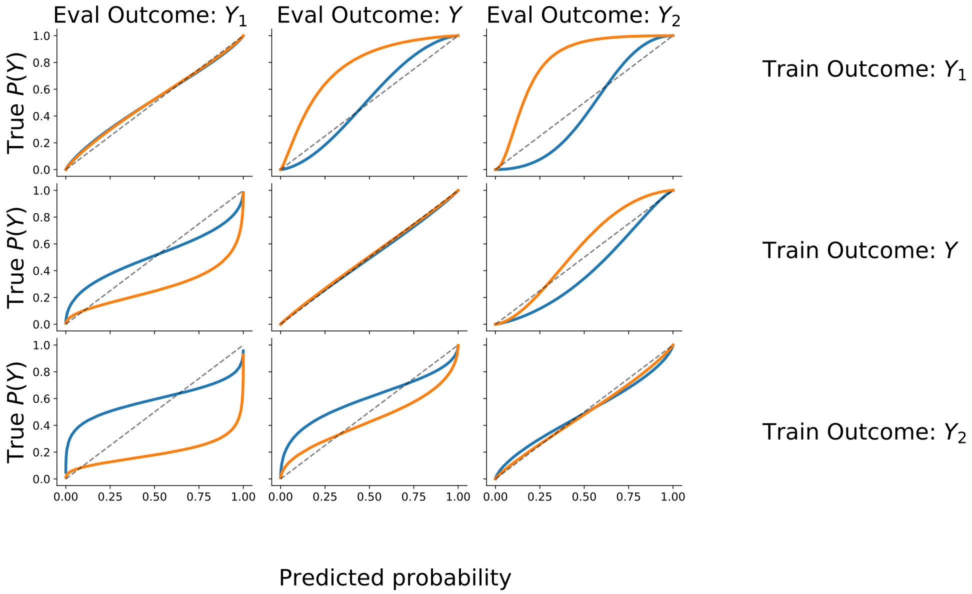


Figure A2: Sufficiency characteristics for models trained with proxies and true outcome generated according to the data generating process in Figure 2h, evaluated across two subgroups represented by orange and blue lines. In this setting, Y_1 is a biased upstream proxy and Y_2 is a biased downstream proxy.