
## Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination

Nathan Kallus, Xiaojie Mao, Angela Zhou

**To cite this article:**
Nathan Kallus, Xiaojie Mao, Angela Zhou (2021) Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination. Management Science

Published online in Articles in Advance 02 Apr 2021

. https://doi.org/10.1287/mnsc.2020.3850

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
*Articles in Advance*, pp. 1–23
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

# Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination

**Nathan Kallus,[a] Xiaojie Mao,[a] Angela Zhou[a]**

[a] Cornell University, Ithaca, New York 14850

**Contact:** kallus@cornell.edu, https://orcid.org/0000-0003-1672-0507 (NK); xm77@cornell.edu, https://orcid.org/0000-0003-2985-1741 (XM); az434@cornell.edu, https://orcid.org/0000-0003-2814-5693 (AZ)

**Received:** June 1, 2019
**Revised:** March 29, 2020; July 3, 2020
**Accepted:** August 25, 2020
**Published Online in Articles in Advance:** April 2, 2021

**https://doi.org/10.1287/mnsc.2020.3850**



**Abstract.** The increasing impact of algorithmic decisions on people's lives compels us to scrutinize their fairness and, in particular, the disparate impacts that ostensibly color-blind algorithms can have on different groups. Examples include credit decisioning, hiring, advertising, criminal justice, personalized medicine, and targeted policy making, where in some cases legislative or regulatory frameworks for fairness exist and define specific protected classes. In this paper we study a fundamental challenge to assessing disparate impacts in practice: protected class membership is often not observed in the data. This is particularly a problem in lending and healthcare. We consider the use of an auxiliary data set, such as the U.S. census, to construct models that predict the protected class from proxy variables, such as surname and geolocation. We show that even with such data, a variety of common disparity measures are generally unidentifiable, providing a new perspective on the documented biases of popular proxy-based methods. We provide exact characterizations of the tightest possible set of all possible true disparities that are consistent with the data (and possibly additional assumptions). We further provide optimization-based algorithms for computing and visualizing these sets and statistical tools to assess sampling uncertainty. Together, these enable reliable and robust assessments of disparities—an important tool when disparity assessment can have far-reaching policy implications. We demonstrate this in two case studies with real data: mortgage lending and personalized medicine dosing.

**History:** Accepted by Hamid Nazerzadeh, Guest Editor for the Special Issue on Data-Driven Prescriptive Analytics.
**Funding:** This material is based on work supported by the National Science Foundation [Grant 1939704].
**Supplemental Material:** Data and the online appendix are available at https://doi.org/10.1287/mnsc.2020.3850.

**Keywords:** disparate impact and algorithmic bias • partial identification • proxy variables • fractional optimization • Bayesian Improved Surname Geocoding

## 1. Introduction

The spread of prescriptive analytics and algorithmic decision making has given rise to urgent ethical and legal imperatives to avoid discrimination and guarantee fairness with respect to protected classes. In advertising, prescriptive algorithms target for maximal impact and revenue (Iyer et al. 2005, Goldfarb and Tucker 2011), but recent studies found gender-based discrimination in who receives ads for STEM (science, technology, engineering, and mathematics) careers (Lambrecht and Tucker 2019) and other worrying disparities (Sweeney 2013, Datta et al. 2015). In hiring, algorithms help employers efficiently screen applicants (Miller 2015), but in some cases this can have unintended biases—for example, against women and minorities (Dastin 2018). In criminal justice, algorithmic recidivism scores allow judges to assess risk (Monahan and Skeem 2016), whereas recent studies have revealed systematic race-based disparities in error rates (Angwin et al. 2016, Chouldechova 2017). In healthcare, algorithms that allocate resources such as care management have been shown to exhibit racial biases (Obermeyer and Mullainathan 2019), and personalized medicine algorithms can offer disparate benefits to different groups (Goodman et al. 2018, Rajkomar et al. 2018). In lending, prescriptive algorithms optimize credit decisions using predicted default risks, and their induced disparities are regulated by law (Comptroller of the Currency 2010), leading to legal cases against discriminatory lending (Consumer Financial Protection Bureau 2013).

For regulated decisions, there are two major legal theories of discrimination:

- *Disparate treatment* (Zimmer 1996): Informally, intentionally treating an individual differently on the basis of membership in a protected class.
- *Disparate impact* (Rutherglen 1987): Informally, adversely affecting members of one protected class

more than another, even if by an ostensibly neutral policy.

Thus, even prescriptive algorithms that do not take race, gender, or other sensitive attributes as an input may often satisfy equal treatment but may still induce disparate impact (Kleinberg et al. 2017). Indeed, many of the above-mentioned disparities take the form of *unintended disparate impact of ostensibly class-blind prescriptive algorithms*. Although our contextual discussion focuses on U.S. discrimination law and regulation, our methodology is a general one for assessing disparities with respect to protected class and may apply in many legal and regulatory contexts.[1]

In consequential decision-making contexts such as hiring or lending, assessing disparities is paramount for monitoring the potential harms of decision systems. Assessing disparities induced by a prescriptive algorithm involves evaluating the differences in the distributions of decision outcomes received by different groups, either marginally or conditional on some additional ground truth. We define precisely the disparity metrics of interest in Section 2.1 and discuss related work in Section 3. Although what size of disparity counts as unacceptable depends on the appropriate legal, ethical, and regulatory context, in any case, they must first be *measured*.

In this paper, we study a fundamental challenge to assessing the disparity induced by prescriptive algorithms in practice: *Protected class membership is often* not *observed in the data.*

There may be many reasons for this missingness in practice, both legal, operational, and behavioral. In the U.S. financial service industry, lenders are *not* permitted to collect race and ethnicity information on applicants for nonmortgage products[2] such as credit cards, auto loans, and student loans. This considerably hinders auditing fairness for nonmortgage loans, both by internal compliance officers and by regulators (Zhang 2016). Similarly, health plans and healthcare delivery entities lack race and ethnicity data on most of their enrollees and patients, as a consequence of high data-collection costs and people's reluctance to reveal their race information for fear of potential discrimination (Weissman and Hasnain-Wynia 2011). This data collection challenge makes monitoring of racial and ethnic differences in care impractical and impedes the progress of healthcare equity reforms (Gaffney and McCormick 2017).

To address this challenge, some methods heuristically use observed proxies to predict and impute unobserved protected class labels. The most (in)famous example is the Bayesian Improved Surname Geocoding (BISG) method. BISG estimates conditional race membership probabilities given a surname and geolocation (e.g., census tract, zip code, or county) using data from the U.S. decennial census, and then it imputes the race labels based on the estimated probabilities. Since its invention (Elliott et al. 2008, 2009), the BISG method has been widely used in assessing racial disparities in healthcare (e.g., Fremont et al. 2005, Ulmer et al. 2009, Weissman and Hasnain-Wynia 2011, and Brown et al. 2016), as well in the U.S. financial industry, where the Consumer Financial Protection Bureau (CFPB) used BISG to support analysis leading to a $98 million settlement against Ally Bank for harming minority borrowers for auto loans (Consumer Financial Protection Bureau 2013, 2014).

The validity of using proxies for the unobserved protected class for disparity assessment remains controversial, and relevant research is still limited. Although advanced proxy methods such as BISG outperform previous proxy methods, further research shows that it leads to biased disparity assessment (Baines and Courchane 2014, Zhang 2016). In particular, Chen et al. (2019) analyzed the underlying mechanism for the statistical bias of BISG's assessments as a result of the joint dependence among lending outcome, geolocation, and race.

However, a systematic understanding of the precise limitations of using proxy methods in disparity assessment in general, and possible remedies to the potential statistical biases, is still lacking.[3] Filling in this gap is an important and urgent need, especially given the wide use of proxy methods and the significant managerial and policy impacts of disparity assessment in the settings where they are used, which motivates our current work.

### 1.1. Contributions

In this paper, we demonstrate that it is generally *impossible* to identify disparities when only proxy information is available for protected class, and we instead study how to precisely and reliably characterize the *range* of all possible disparities that are consistent with all available data, known as the *partial identification set*. This perspective provides a principled approach to analyze the *fundamental limitations* of all previous proxy methods *regardless* of the actual point estimators they use. Specifically, we describe the set of point values for disparities that are consistent with the distribution of the observed data, which characterizes the fundamental identification uncertainty in disparity assessments that exists even in the absence of finite-sample uncertainty. To implement our work in practice, we, of course, need to estimate these sets from finite samples, which we show how to do. We further provide inferential methods to account for the finite-sample uncertainty in these estimates (which vanishes as we collect more data), in addition to the (nonvanishing) identification uncertainty.

We highlight our primary contributions in the following subsections.

**1.1.1. Problem Formulation.** To facilitate a principled analysis of (partial) identifiability, we formulate disparity assessment with proxies as a *data combination problem* with two data sets:

—A *primary data set* with the decision outcomes, (potentially) true outcomes, and proxy variables, but where the protected class labels are *missing*; and

—An *auxiliary* data set with proxy variables and protected class labels, but without outcomes.

**1.1.2. Identification Conditions.** We prove tight necessary and sufficient conditions for the identifiability of disparity measures in this setting. In the absence of these (unrealistically strong) conditions, disparities are *necessarily* unidentifiable from the two data sets. That is, the partial identification set of all disparity measure values consistent with the data-generating processes of the two data sets is not a singleton.

**1.1.3. Characterizing and Computing the Partial Identification Set.** We exactly characterize the partial identification sets of a variety of disparity measures under data combination—that is, the smallest set containing all possible values that disparity measures may simultaneously take while still agreeing with the data. Our characterization is *sharp* in that it is *equal* to this set rather than merely containing it. We provide closed-form formulations of partial identification sets for binary comparisons. And we provide optimization algorithms to compute partial identification sets when we incorporate additional mild smoothness assumptions that reduce ambiguity or when we consider simultaneous comparisons across more than two protected classes. In the latter case, we compute the *support function* of the partial identification set.

**1.1.4. Estimation and Inference.** We study the additional *sampling uncertainty* of our proposals when given finite observations from each data set. Specifically, we prove consistency guarantees when one plugs in estimates of probability and conditional probability models. To enable inference (i.e., constructing confidence intervals on top of the estimated partial identification intervals), we propose a debiased estimation approach that is invariant to the estimation of the conditional probability models.

**1.1.5. Robust Auditing.** Together, these tools facilitate robust and reliable fairness auditing. Because the sets we describe are sharp in that they are the tightest possible characterization of disparity given the data, their size generally captures the amount of *ambiguity* that remains in evaluating disparity when the protected class is unobserved and only proxies are available. When the observed data are very informative about the disparity measures, the set tends to be small and may still lead to meaningful conclusions regarding the sign and magnitudes of disparity, despite unidentifiability. By contrast, when the observed data are insufficient, the set tends to be large and gives a valuable warning about the risk of drawing conclusions from the fundamentally limited observed data.

**1.1.6. Empirical Analysis.** We apply our approach in two real case studies: evaluating the racial disparities (1) in mortgage lending decisions and (2) in personalized warfarin dosing. We demonstrate how adding extra assumptions may decrease the size of partial identification sets of disparity measures and illustrate how stronger proxies—either for race or for outcomes—can lead to smaller partial identification sets and more informative conclusions on disparities.

### 1.2. Practical Implications

Collecting sensitive attributes such as protected class membership remains a serious challenge in practice that may be ultimately insurmountable. For example, Holstein et al. (2019) surveyed industry practitioners in machine learning and found that many practitioners do not have access to protected attributes. Bogen et al. (2020) identified challenges for sensitive attribute collection in both traditionally regulated sectors such as credit and employment, voluntary efforts in healthcare, as well as efforts to audit and/or promote equity by technology companies including Airbnb, Facebook, and LinkedIn. Because, as we show, disparities are unidentifiable when we only have proxies for such attributes, any single point estimate of disparities, such as those given by many current methods, is fundamentally spurious, and any conclusion drawn from it is vulnerable to criticism and may mislead decision making. This is a grave and real concern in the above-mentioned applications where disparate impact assessments can have far-reaching policy implications.

By contrast, by conducting inference on the *range* of possible disparities based on the data available, our proposed methods can support credible, principled conclusions about disparities. This can be relevant, for example, if a decision maker is concerned with auditing disparities and is choosing between different algorithms based on performance disparities with respect to unobserved protected attributes, or even choosing between investments in different auxiliary proxy data to better estimate disparities. Our approach is relevant for partial identification of disparities and informing these decisions. If the partial identification sets are small, they provide a statistical test certifying the presence of disparities independent of untestable assumptions and

can inform credible comparisons between algorithms. If, instead, the partial identification sets are large, this highlights both the limitations of using the available data to draw credible conclusions about disparities and the value of more informative proxy variables or assumptions.

## 2. Problem Setup

We mainly consider four types of relevant variables:

- The *decision outcome*, $\hat{Y} \in \{0,1\}$, is the prescription by either human decision makers or machine learning algorithms. For example, $\hat{Y}=1$ represents the approval of a loan application, which is often based on some prediction of default risk. We call $\hat{Y}=1$ the *positive* decision, even if is not favorable in terms of utility (e.g., high medicine dosage in Section 8.2).
- The *true outcome*, $Y \in \{0,1\}$, is a target variable that justifies an optimal decision. $\hat{Y}$ is often based on imperfect predictions of $Y$. In the lending example (Section 8.1), we denote $Y=1$ for loan applicants who would not default on loan payment if the loan application were approved. Note that $Y$ is not known to decision makers at the time of decision making.
- The *protected attribute*, $A \in \mathcal{A}$, is a categorical variable (e.g., race or gender). Our convention is to let $A=a$ be a group understood to be generally *advantaged* and $A=b$ *disadvantaged*.
- The *proxy variables*, $Z \in \mathcal{Z}$, are a set of additional observed covariates. In proxy methods, these are used to predict $A$. In the BISG example (Section 8.1), $Z$ stands for surname and geolocation. The proxy variables can be categorical, continuous, or mixed.

In this paper, we mainly focus on binary outcomes (true outcome and decision outcome), but our results can be straightforwardly extended to multileveled outcomes.

We formulate the problem of using proxy methods from a *data combination* perspective. Specifically, we assume we have two data sets: the *main* data set with observations of $(\hat{Y}, Y, Z)$ and the *auxiliary* data set with observations of $(A, Z)$. Figure 1 is an illustration of these two data sets in the example of the BISG proxy method (Section 8.1).

**Assumption 1.** *The primary and auxiliary data sets both consist of independent and identically distribution (i.i.d.) draws, each from the respective marginalization of a common joint distribution.*

Therefore, the information from observing these two separate data sets can be characterized by $\mathbb{P}(\hat{Y}, Y, Z)$ and $\mathbb{P}(A, Z)$, respectively, each being a marginalization of a common larger joint distribution $\mathbb{P}(A, \hat{Y}, Y, Z)$.[4] However, we cannot simply join these two data sets directly for many possible reasons. For example, no unique identifier for individuals (e.g., social security number) exists in both data sets. Thus we *cannot* learn the combined joint distribution $\mathbb{P}(A, \hat{Y}, Y, Z)$ from these two separate, unconnected data sets.

### 2.1. Disparity Measures

In this paper, we focus on assessing the disparity in the decision $\hat{Y}$ with respect to the protected attribute $A$, as well as possibly with respect to true outcome labels $Y$. We illustrate our method with widely used disparity measures that are a measure of class-conditional classification error, and if we were given observations of true class labels, they could be computed from a $2 \times 2 \times |\mathcal{A}|$ within-class confusion matrix of the decision and true outcome.

Specifically, we consider the following disparities.

- *Demographic disparity* (DD): $\delta_{\mathrm{DD}}(a,b) = \mathbb{P}(\hat{Y}=1 \mid A=a) - \mathbb{P}(\hat{Y}=1 \mid A=b)$.
- *True-positive rate disparity* (TPRD): $\delta_{\mathrm{TPRD}}(a,b) = \mathbb{P}(\hat{Y}=1 \mid A=a, Y=1) - \mathbb{P}(\hat{Y}=1 \mid A=b, Y=1)$.
- *True-negative rate disparity* (TNRD): $\delta_{\mathrm{TNRD}}(a,b) = \mathbb{P}(\hat{Y}=0 \mid A=a, Y=0) - \mathbb{P}(\hat{Y}=0 \mid A=b, Y=0)$.
- *Positive predictive value disparity* (PPVD): $\delta_{\mathrm{PPVD}}(a,b) = \mathbb{P}(Y=1 \mid A=a, \hat{Y}=1) - \mathbb{P}(Y=1 \mid A=b, \hat{Y}=1)$.
- *Negative predictive value disparity* (NPVD): $\delta_{\mathrm{NPVD}}(a,b) = \mathbb{P}(Y=0 \mid A=a, \hat{Y}=0) - \mathbb{P}(Y=0 \mid A=b, \hat{Y}=0)$.

To illustrate, we interpret these disparity measures using the running example of making lending decisions. DD measures the disparity in within-class average loan approval rate.[5] TPRD (respectively, TNRD) measures the disparity in the proportions of people who *correctly* get approved (respectively, rejected) in loan applications between two classes, given their true nondefault or default outcome. Compared with DD, TPRD and TNRD only measure the disparity that is unmediated by existing base disparities in true outcome $Y$ and is considered more relevant for classification settings when concerned with disparities in allocation of a positive outcome in view of qualifying characteristics such as creditworthiness (Hardt et al. 2016). Such disparities can be interpreted as "disparate opportunity"

**Figure 1.** Illustration of the Two Observed Data Sets for Assessing Lending Disparity with Unobserved Race Labels

Primary dataset

| $Z_s$ Surname | $Z_g$ ZIP code | ⋯ | $\hat{Y}$ Approval | $Y$ Non-default |
|---|---|---|---|---|
| Jones | 94122 | ⋯ | Y | N |
| ⋮ | ⋮ | | ⋮ | ⋮ |

Auxiliary dataset

| $Z_s$ Surname | $Z_g$ ZIP code | White % | ⋯ | API % |
|---|---|---|---|---|
| Jones | 94122 | 47% | ⋯ | 31% |
| ⋮ | ⋮ | ⋮ | | ⋮ |

to equally qualified individuals from different groups. PPVD (respectively, NPVD) measures the disparity in the proportions of approved applicants who pay back their loan (respectively, rejected applicants who default) between two classes. Such disparities can be interpreted as "disparate benefit of the doubt" in an individual having the positive label.

We will present our results in terms of DD, TPRD, and TNRD. Indeed, by swapping the roles of $Y$ and $\hat{Y}$ in TPRD and TNRD, all our results can straightforwardly be extended to PPVD and NPVD, respectively. Similarly, disparities based on a false-negative rate and false-positive rate simply differ with TPRD and TNRD by a minus sign (i.e., are given by swapping $a$ and $b$). Because the true positive rate (TPR) and true negative rate (TNR) characterize the receiver operating characteristic curve, our results can also be extended to assessing bipartite ranking scores (Kallus and Zhou 2019b).

To streamline the presentation, we typically use $\alpha$, $z$, $\hat{y}$, and $y$ as generic values of the random variables $A$, $Z$, $\hat{Y}$, and $Y$, respectively. We also use $a$ and $b$ as additional generic values for $A$, where $a$ is generally understood to be a majority or advantaged class label. We further define the outcome probabilities for protected class $\alpha$ as $\mu(\alpha) := \mathbb{P}(\hat{Y} = 1 \mid A = \alpha)$ and $\mu_{\hat{y}y}(\alpha) := \mathbb{P}(\hat{Y} = \hat{y} \mid A = \alpha, Y = y)$, so that $\delta_{\mathrm{DD}}(a,b) = \mu(a) - \mu(b)$, $\delta_{\mathrm{TPRD}}(a,b) = \mu_{11}(a) - \mu_{11}(b)$, and $\delta_{\mathrm{TNRD}}(a,b) = \mu_{00}(a) - \mu_{00}(b)$. Throughout this paper, we use $\mathbb{E}$ to denote expectation with respect to the target distribution $\mathbb{P}$.

## 3. Related Literature

### 3.1. Proxy Methods

The validity of proxy methods for disparity assessment depends not only on the statistical estimation of, for example, $\mathbb{P}(A = \alpha \mid Z = z)$, but also the specific procedure with which this is combined with other information. Although BISG has been shown to outperform previous proxies (surname-only and geolocation-only analysis), these evaluations (Consumer Financial Protection Bureau 2014, Imai and Khanna 2016, Dembosky et al. 2019) focus on classification accuracy, which is never perfect, and do not consider impact on downstream disparity assessment, mostly because this is usually unknowable. By contrast, Baines and Courchane (2014) and Zhang (2016) assessed disparity on a mortgage data set, and they found that using imputed race tends to overestimate the true disparity. Chen et al. (2019) provided a full analysis of this bias and developed sufficient conditions to determine its direction and found that disparity estimation methods using imputed race are very sensitive to arbitrary tuning parameters such as imputation threshold. As we show in Section 4, disparity is generally *unidentifiable* from proxies when protected class is unobserved; consequently, all previous point estimators are generally biased unless very strong assumptions are satisfied.

### 3.2. Algorithmic Fairness

In this paper, we consider auditing two measures of fairness that have received considerable attention in the fair machine learning community: *demographic* (*dis*) *parity* and *classification* (*dis*) *parity*, which we outlined in Section 2.1. Many other "fairness metrics" have been proposed to facilitate risk assessment for algorithmic decision making in different contexts (Narayanan 2018, Verma and Rubin 2018); for a more comprehensive discussion, we refer to Barocas et al. (2018). We emphasize that we focus on auditing, not adjusting, disparity measures. Whether observed disparities warrant adjustments depends on the legal, ethical, and regulatory context.[6] Several works have considered limitations to measuring these fairness metrics in practice (Kallus and Zhou 2018b, 2019a; Chen et al. 2019).

### 3.3. Partial Identification and Data Combination

There is an extensive literature on partial identification of unidentifiable parameters (e.g., Manski 2003 and Beresteanu et al. 2011). There are many reasons parameters may be unidentifiable, including confounding (e.g., Kallus and Zhou 2018a and Kallus et al. 2019), missingness (e.g., Manski 2005), and multiple equilibria (e.g., Ciliberto and Tamer 2009). One prominent example is data combination, also termed the "ecological inference problem," where joint distributions must be reconstructed from observation of marginal distributions (Freedman 1999, Schuessler 1999, Wakefield 2004, Jiang et al. 2018). One key tool for studying this problem is the Fréchet–Hoeffding inequalities, which give sharp bounds on joint cumulative distributions and superadditive expectations given marginals (Cambanis et al. 1976, Ridder and Moffitt 2007, Fan et al. 2014). Such tools are also used in risk analysis in finance to assess risk without knowledge of copulas (Rüschendorf 2013). In contrast to much of the above-mentioned work, we focus on assessing *nonlinear* functionals of partially identified distributions—namely, true positive and negative rates, as well as on leveraging conditional information to integrate marginal information across proxy-value levels with possible smoothness constraints.

## 4. Unidentifiability of Disparity Measures Under Data Combination

In this section we study the fundamental limits of the two separate data sets to identify (i.e., pinpoint)

the disparity measures of interest. We first introduce the concept of *identification* (Lewbel 2018). We call a quantity of interest (either finite-dimensional or infinite-dimensional) *identifiable* if it can be uniquely determined by (i.e., is a function of) the probability distribution function of the data. Conversely, it is *unidentifiable* if multiple different values of this quantity all simultaneously agree with the distribution of observed data. This is motivated by the fact that, in the i.i.d. setting, the distribution of the data (equivalently, the distribution of any single data point) is the most we can hope to learn from any number of observations, even infinitely many.

The disparity measures of interest in Section 2 are all functions of the full joint distribution $\mathbb{P}(A, \hat{Y}, Y, Z)$ and are clearly identifiable *if* we observed the full data $(A, \hat{Y}, Y, Z)$. We will show in Section 4.1 that disparity measures are generally unidentifiable from two separate data sets, because the corresponding marginal distributions $\mathbb{P}(\hat{Y}, Y, Z)$ and $\mathbb{P}(A, Z)$ are insufficient to uniquely determine the full joint distribution and, in particular, the disparity measures.

Analyzing the identifiability of disparity measures is crucial because unidentifiability implies that learning disparities based on the observed data alone is fundamentally ambiguous: it is *impossible*, even with an infinite amount of observed data, to pin down the exact values of the disparity measures. Consequently, any point estimate is in some sense *spurious*: biased and potentially sensitive to ad hoc modeling specifications (D'Amour 2019). In this case, generally one must be very cautious about drawing any substantive conclusions based on point estimates of disparity measures.

Because identifiability and partial identification sets are properties of *distributions* (i.e., are *population* quantities), we focus for the time being on the consequences of fully knowing the marginals $\mathbb{P}(\hat{Y}, Y, Z)$ and $\mathbb{P}(A, Z)$, which can be learned from the two data sets given sufficient data. This captures the *identification uncertainty* involved in disparity assessments using data combination. We revisit the assumption of full knowledge of marginals in Section 7, where we discuss how the partial identification sets can be *estimated* from the data and how to construct *confidence intervals* on these. This captures the *sampling uncertainty* involved in only having finite data sets.

### 4.1. Unidentifiability of Disparities

Because the disparity measures are functions of the full joint distribution $\mathbb{P}(A, \hat{Y}, Y, Z)$, to prove the unidentifiability of the disparity measures, we show that there generally exist multiple *valid* full joint distributions that give rise to different disparities but at the same time agree with the marginal joint distributions $\mathbb{P}(\hat{Y}, Y, Z)$ and $\mathbb{P}(A, Z)$, which characterize the primary data set and the auxiliary data set, respectively. To formalize the validity of full joint distributions, we introduce the *coupling* of two marginal distributions (Villani 2008). Because outcomes and protected classes are discrete, we focus on couplings of discrete distributions.[7]

**Definition 1** (Coupling Sets)**.** Given two discrete probability spaces $(\mathcal{S}, \sigma)$ and $(\mathcal{T}, \tau)$ (i.e., $\sigma(s) \geq 0, \tau(t) \geq 0$, $\sum_{s\in\mathcal{S}} \sigma(s) = 1, \sum_{t\in\mathcal{T}} \tau(t) = 1$), a distribution $\pi$ over $\mathcal{S} \times \mathcal{T}$ is a *coupling* of $(\sigma, \tau)$ if the marginal distributions of $\pi$ coincide with $\sigma$, $\tau$. The set of all possible couplings is denoted by

$$\Pi(\sigma, \tau) = \left\{ \pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{T}} : \sum_{t\in\mathcal{T}} \pi(s,t) = \sigma(s), \sum_{s\in\mathcal{S}} \pi(s,t) = \tau(t),\ 0 \leq \pi(s,t) \leq 1, \ \forall s \in \mathcal{S}, t \in \mathcal{T} \right\}. \tag{1}$$

Definition 1 gives the set of all possible valid joint distributions that agree with given marginals. It states that any joint distribution is valid as long as it satisfies the *law of total probability* with respect to the fixed marginals. The classical Frećhet–Hoeffding inequality provides bounds on the possible values of these joint distributions with knowledge of the fixed marginals (Cambanis et al. 1976, Ridder and Moffitt 2007, Fan et al. 2014): this characterization informs our discussion of the size of partial identification sets in Section 4.2 and our derivation of closed-form partial identification sets in Section 5.

**Proposition 1** (Fréchet–Hoeffding)**.** *The coupling set is equivalently given by*

$$\Pi(\sigma, \tau) = \left\{ \pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{T}} : \sum_{t\in\mathcal{T}} \pi(s,t) = \sigma(s), \sum_{s\in\mathcal{S}} \pi(s,t) = \tau(t), \max\{\sigma(s) + \tau(t) - 1, 0\} \leq \pi(s,t) \leq \min\{\sigma(s), \tau(t)\},\ \forall s \in \mathcal{S}, t \in \mathcal{T} \right\}. \tag{2}$$

We let $\mathcal{P}_D$ denote the set of all valid full joint distributions of $(\hat{Y}, Y, A, Z)$ that agree with the marginal distributions $\mathbb{P}(\hat{Y}, Y, Z)$ and $\mathbb{P}(A, Z)$, as characterized by this definition of couplings:

$$\mathcal{P}_D = \left\{ \mathbb{P}' : \mathbb{P}'(Z) = \mathbb{P}(Z), \mathbb{P}'(\hat{Y}, Y, A \mid Z) \in \Pi\left(\mathbb{P}(\hat{Y}, Y \mid Z), \mathbb{P}(A \mid Z)\right) \right\} \tag{3}$$

**Figure 2.** Unidentifiability of Joint Distributions Given Marginal Distributions

| | $A=a$ | $A=b$ | |
|---|---|---|---|
| $\hat{Y}=0$ | $\mathbb{P}(A=a,\hat{Y}=0 \mid Z=z)$ | $\mathbb{P}(A=b,\hat{Y}=0 \mid Z=z)$ | $\mathbb{P}(\hat{Y}=0 \mid Z=z)$ |
| $\hat{Y}=1$ | $\mathbb{P}(A=a,\hat{Y}=1 \mid Z=z)$ | $\mathbb{P}(A=b,\hat{Y}=1 \mid Z=z)$ | $\mathbb{P}(\hat{Y}=1 \mid Z=z)$ |
| | $\mathbb{P}(A=a \mid Z=z)$ | $\mathbb{P}(A=b \mid Z=z)$ | 1 |

*Notes.* The gray region denotes unknown joint probabilities. Row and column sums are known. Even with binary protected class and outcome, this leaves one degree of freedom in the unknowns, unless one of the marginals is degenerate.

The set $\mathcal{P}_D$ generally contains multiple elements, because the joint dependence structure can be arbitrary so long as the marginals are compatible (characterized by either Definition 1 or Equation (2)). We illustrate this for $\Pi(\mathbb{P}(A \mid Z), \mathbb{P}(\hat{Y} \mid Z))$ in Figure 2. With binary protected class and outcomes, marginal information provides only three independent constraints on four unknowns, so that the joint distribution cannot be uniquely determined. This also extends to $\Pi(\mathbb{P}(A \mid Z), \mathbb{P}(\hat{Y}, Y \mid Z))$.

We next show that in addition to the full joint distribution, the disparities, which are differences of nonlinear functionals of the full joint distribution, in particular cannot be uniquely identified.

**Proposition 2.** *Let $\mathcal{A} = \{a, b\}$, and let marginal distributions $\mathbb{P}(\hat{Y}, Y, Z)$ and $\mathbb{P}(A, Z)$ be given.*

i. *If there exists a set of $z$'s with positive probability such that $0 < \mathbb{P}(\hat{Y} = \hat{y} \mid Z = z) < 1$ and $0 < \mathbb{P}(A = \alpha \mid Z = z) < 1$ for $\hat{y} \in \{0, 1\}$ and $\alpha \in \mathcal{A}$, then $\delta_{\mathrm{DD}}(a, b)$ is unidentifiable without further conditions. That is, there exist two different joint distributions of $(A, \hat{Y}, Y, Z)$ that agree with these marginals but give rise to different values of $\delta_{\mathrm{DD}}(a, b)$.*

ii. *If there exists a set of $z$'s with positive probability such that $0 < \mathbb{P}(\hat{Y} = \hat{y}, Y = y \mid Z = z) < 1$ and $0 < \mathbb{P}(A = \alpha \mid Z = z) < 1$ for $\hat{y}, y \in \{0, 1\}$ and $\alpha \in \mathcal{A}$, then both $\delta_{\mathrm{TPRD}}(a, b)$ and $\delta_{\mathrm{TNRD}}(a, b)$ are unidentifiable without further conditions.*

Proposition 2 shows that as long as the proxies $Z$ cannot perfectly predict the protected class $A$ or outcomes $\hat{Y}, Y$, then DD, TPRD, and TNRD are unidentifiable from the observed data information alone. This holds for *any* given pair of marginal distributions. We can also prove the same conclusion for PPVD and NPVD by exchanging $\hat{Y}$ and $Y$. To prove Proposition 2 we show that we can always construct different feasible couplings of the given marginals (i.e., different feasible elements in $\mathcal{P}_D$) that lead to different values of the disparities. Because disparities are differences of nonlinear functionals of the coupling, we need to construct the couplings very carefully to achieve unambiguously different disparity values. See Online Appendix B.1 for the proof.

### 4.2. Partial Identification Set of Disparities

In the last section we showed that DD, TPRD, and TNRD (and symmetrically, also PPVD and NPVD) are generally *not* identifiable from the two separate data sets. Next, we will characterize exactly how identifiable or unidentifiable they are by characterizing the partial identification set of all disparity values that agree with the observed data and possibly additional assumptions that reflect prior knowledge.

Each disparity measure in Section 2.1 can be viewed as a function of the true distribution of $(\hat{Y}, Y, A, Z)$, so we generically denote it as $\delta(a, b; \mathbb{P})$. The *partial identification set* of this disparity measure of interest given observed data information (encoded by $\mathcal{P}_D$, defined in Equation (3)) and extra assumptions (encoded by $\mathcal{P}_A$)[8] is defined as follows:

$$\Delta(\mathcal{P}_D \cap \mathcal{P}_A) = \{\delta(a, b; \mathbb{P}') : \mathbb{P}' \in \mathcal{P}_D \cap \mathcal{P}_A\}. \quad (4)$$

We will add subscripts such as DD or TPRD to indicate the set for a particular disparity measure. The partial identification set in Equation (4) is the smallest set containing all possible values of the disparity measures that agree with both the observed data and possibly extra assumptions. Each disparity value in this set is given by one valid full joint distribution that is compatible with the observed data and extra assumptions, and any disparity value outside this set is ruled out by either the observed data or the assumptions. A natural question is, when are these smallest possible sets also actually small? We next discuss different scenarios where the sets can be small or large.

#### 4.2.1. Informative Proxies.

If the proxies are very predictive, then the observed data alone may be informative enough to sufficiently pin down the disparity measures. At the extreme, if proxies are perfectly predictive of either the outcomes or the protected class, then the partial identification sets collapse into singletons; that is, the disparity measures are uniquely identified from the observed data. We formalize this in the following proposition.

**Proposition 3.** *Given marginal distributions $\mathbb{P}(\hat{Y}, Y, Z)$ and $\mathbb{P}(A, Z)$, if the assumptions of Proposition 2(ii) are not*

*satisfied—that is, for almost all $z$, either $\mathbb{P}(\hat{Y}=\hat{y}, Y=y \mid Z=z) \in \{0,1\}$ for $\hat{y}, y \in \{0,1\}$ or $\mathbb{P}(A=\alpha \mid Z=z) \in \{0,1\}$ for $\alpha \in \mathcal{A}$—then $\mathcal{P}_D$ is a singleton, and hence $\Delta(\mathcal{P}_D)$ for any disparity measure in Section 2 is also a singleton.*

**Proof.** According to the Fréchet–Hoeffding inequality in Proposition 1, any valid full joint distribution agreeing with the observed data (i.e., an element $\mathbb{P}' \in \mathcal{P}_D$) has to satisfy that

$$\begin{aligned}&\mathbb{P}'(A=\alpha, \hat{Y}=\hat{y}, Y=y \mid Z)\\ &\quad \le \min\{\mathbb{P}(\hat{Y}=\hat{y}, Y=y \mid Z), \mathbb{P}(A=\alpha \mid Z)\},\end{aligned} \tag{5}$$

$$\begin{aligned}&\mathbb{P}'(A=\alpha, \hat{Y}=\hat{y}, Y=y \mid Z)\\ &\quad \ge \max\{\mathbb{P}(\hat{Y}=\hat{y}, Y=y \mid Z)+\mathbb{P}(A=\alpha \mid Z)-1, 0\}.\end{aligned} \tag{6}$$

Under the stated assumptions, the right-hand sides of Equations (5) and (6) are equal. Thus, the full joint distribution is uniquely determined by the marginals, and $\mathcal{P}_D$ is a singleton.

This shows that the conditions of Proposition 2 are the tight necessary *and* sufficient conditions for identifiability from marginals alone. If proxies are not perfect but are very predictive, either of protected class or of outcomes, or both, then the endpoints of Fréchet–Hoeffding inequality (i.e., right-hand sides of Equations (5) and (6)) are not exactly equal but are still close. Consequently, the partial identification sets will be small. This is the case we observe in Section 8.2 when using very informative genetic proxies for race.

#### 4.2.2. Strong Assumptions on the Joint Distributions.

For the case of DD, Chen et al. (2019) discussed a conditional independence assumption that admits an unbiased proxy-based estimator. Actually, this assumption is sufficient for the identifiability of disparities more generally.

**Proposition 4.** *If we assume that $Y, \hat{Y} \perp\!\!\!\perp A \mid Z$, which is to say*

$$\begin{aligned}\mathcal{P}_A = \{&\mathbb{P}' : \mathbb{P}'(\hat{Y}=\hat{y}, Y=y, A=\alpha \mid Z=z)\\ &= \mathbb{P}'(\hat{Y}=\hat{y}, Y=y \mid Z=z)\mathbb{P}'(A=\alpha \mid Z=z),\\ &\forall \alpha, \hat{y}, y, z\},\end{aligned}$$

*then $\mathcal{P}_D \cap \mathcal{P}_A$ is a singleton, and hence $\Delta(\mathcal{P}_D \cap \mathcal{P}_A)$ for any disparity measure is also a singleton.*

**Proof.** Any $\mathbb{P}' \in \mathcal{P}_D \cap \mathcal{P}_A$ satisfies for $z \in \mathcal{Z}, \hat{y}, y \in \{0,1\}$, $\alpha \in \mathcal{A}$, $\mathbb{P}'(\hat{Y}=\hat{y}, Y=y, A=\alpha \mid Z=z) = \mathbb{P}(\hat{Y}=\hat{y}, Y=y \mid Z=z)\mathbb{P}(A=\alpha \mid Z=z)$. Because this is uniquely determined by the marginals, $\mathcal{P}_D \cap \mathcal{P}_A$ contains only a single element.

Although the conditional independence assumption is indeed very informative, it may be too unrealistic in practice. Indeed, the proxies $Z$ that can be observed on both data sets are usually low dimensional (e.g., surname and geolocation), so they are unlikely to capture all dependence between the outcomes and the protected class. Therefore, although imposing strong assumptions such as this may help identification, it can also result in misleading conclusions, considering that these assumptions are often wrong in reality.

#### 4.2.3. Uninformative or Weakly Informative Proxies and No or Weak Assumptions.

If the observed data sets alone are not highly informative, and we are not willing to impose overly stringent assumptions on the unknown joint distribution, we generally end up with partial identification sets with nontrivial size. For example, in Section 8.1, we find that using geolocation and income as proxies results in quite large partial identification sets in a lending example. Imposing additional mild smoothness assumptions (Section 6.1) narrowed down the set based on income proxies only slightly. In this case, the size of partial identification sets exactly captures the ambiguity in learning disparity measures based on the observed data and imposed assumptions. Large sets are not meaningless: they serve as an important warning about drawing any conclusions from highly flawed data. And even large sets may be informative of the presence of disparities when they are well separated from zero.

## 5. Closed-Form Partial Identification Sets of Disparities for Binary Protected Class Attribute

In this section, we show that the partial identification set in Equation (4) has closed-form solutions when we consider a binary protected class (i.e., $\mathcal{A}=\{a,b\}$) without imposing any additional assumption (i.e., $\mathcal{P}_A$ does not impose any constraint, and $\mathcal{P}_D \cap \mathcal{P}_A = \mathcal{P}_D$).

We first reformulate the partial identification set in Equation (4) for different disparity measures in terms of weighted representations that conveniently separate the identifiable and *un*identifiable parts of the disparities. For any functions $w_\alpha(\hat{y}, z)$ and $\tilde{w}_\alpha(\hat{y}, y, z)$, we define, respectively,

$$\mu(\alpha; w) := \frac{\mathbb{E}\left[w_\alpha\left(\hat{Y}, Z\right)\hat{Y}\right]}{\mathbb{P}(A=\alpha)}, \tag{7}$$

$$\begin{aligned}\mu_{\hat{y}y}(\alpha; \tilde{w}) := {}& \mathbb{E}[\tilde{w}_\alpha(\hat{Y}, Y, Z)\mathbb{I}(Y=y)\mathbb{I}(\hat{Y}=\hat{y})]\\ &\times \mathbb{E}\big[\tilde{w}_\alpha(\hat{Y}, Y, Z)\mathbb{I}(Y=y)\mathbb{I}(\hat{Y}=\hat{y})\\ &+ \tilde{w}_\alpha(\hat{Y}, Y, Z)\mathbb{I}(Y=y)\mathbb{I}(\hat{Y}\neq\hat{y})\big]^{-1},\end{aligned} \tag{8}$$

Furthermore, we define $w^*_\alpha(\hat{y},z), \tilde{w}^*_\alpha(\hat{y},y,z)$ as the conditional probabilities of protected class given outcomes and proxies:

$$w^*_\alpha(\hat{y},z) := \mathbb{P}\big(A=\alpha \mid \hat{Y}=\hat{y}, Z=z\big),$$
$$\tilde{w}^*_\alpha(\hat{y},y,z) := \mathbb{P}\big(A=\alpha \mid \hat{Y}=\hat{y}, Y=y, Z=z\big),$$

such that DD, TPRD, and TNRD satisfy $\delta_{\mathrm{DD}}(a,b) = \mu(a;w^*) - \mu(b;w^*)$, $\delta_{\mathrm{TPRD}}(a,b) = \mu_{11}(a;\tilde{w}^*) - \mu_{11}(b;\tilde{w}^*)$, and $\delta_{\mathrm{TNRD}}(a,b) = \mu_{00}(a;\tilde{w}^*) - \mu_{00}(b;\tilde{w}^*)$, respectively. Note that for any fixed functions $w, \tilde{w}$, both $\mu(\alpha;w)$ and $\mu_{\hat{y}y}(\alpha;\tilde{w})$ are *identifiable* from just the marginal distribution $\mathbb{P}(\hat{Y},Y,Z)$ because every term is just an expectation over this distribution. On the other hand, $w^*, \tilde{w}^*$, which depend on the unidentifiable full joint distribution $\mathbb{P}(A,\hat{Y},Y,Z)$, are themselves unidentifiable and therefore render the disparities, which depend on them, unidentifiable.

Although the true $w^*, \tilde{w}^*$ are unidentifiable, we can construct the set of all possible values of these unknown conditional probabilities that agree with the observed data. For any set $\mathcal{P}$ of joint distributions, define

$$\mathcal{W}(\mathcal{P}) = \big\{w : w_\alpha(\hat{y},z) = \mathbb{P}'\big(A=\alpha \mid \hat{Y}=\hat{y}, Z=z\big), \ \forall \hat{y},y,z,\ \mathbb{P}' \in \mathcal{P}\big\}, \tag{9}$$

$$\tilde{\mathcal{W}}(\mathcal{P}) = \big\{\tilde{w} : \tilde{w}_\alpha(\hat{y},y,z) = \mathbb{P}'\big(A=\alpha \mid \hat{Y}=\hat{y}, Y=y, Z=z\big), \ \forall \hat{y},y,z,\ \mathbb{P}' \in \mathcal{P}\big\}. \tag{10}$$

Then, we can characterize the partial identification sets of disparities simply by these sets of conditional probabilities.

**Proposition 5.** *For any set $\mathcal{P}$ of joint distributions on $(A,\hat{Y},Y,Z)$, we have $\Delta_{\mathrm{DD}}(\mathcal{P}) = \{\mu(a;w) - \mu(b;w) : w \in \mathcal{W}(\mathcal{P})\}$, $\Delta_{\mathrm{TPRD}}(\mathcal{P}) = \{\mu_{11}(a;\tilde{w}) - \mu_{11}(b;\tilde{w}) : \tilde{w} \in \tilde{\mathcal{W}}(\mathcal{P})\}$, and $\Delta_{\mathrm{TNRD}}(\mathcal{P}) = \{\mu_{00}(a;\tilde{w}) - \mu_{00}(b;\tilde{w}) : \tilde{w} \in \tilde{\mathcal{W}}(\mathcal{P})\}$.*

In particular, Proposition 5 holds for $\mathcal{P}_D$. In the following proposition, we give explicit formulae for $\mathcal{W}(\mathcal{P}_D)$, $\tilde{\mathcal{W}}(\mathcal{P}_D)$ in terms of the law of total probability (LTP) constraints as in Definition 1.

**Proposition 6.** *Given marginals $\mathbb{P}(A \mid Z)$, $\mathbb{P}(\hat{Y},Y,Z)$, we have*

$$\mathcal{W}(\mathcal{P}_D) = \left\{ w : \begin{array}{c} \sum_{\hat{y}\in\{0,1\}} w_\alpha(\hat{y},z)\mathbb{P}(\hat{Y}=\hat{y}\mid Z=z) = \mathbb{P}(A=\alpha\mid Z=z), \\ \sum_{\alpha\in\mathcal{A}} w_\alpha(\hat{y},z) = 1,\ 0 \le w_\alpha(\hat{y},z) \le 1, \text{ for any } \alpha,z,\hat{y}, \end{array} \right\}$$

$$\tilde{\mathcal{W}}(\mathcal{P}_D) = \left\{ \tilde{w} : \begin{array}{c} \sum_{\hat{y},y\in\{0,1\}} \tilde{w}_\alpha(\hat{y},y,z)\,\mathbb{P}(\hat{Y}=\hat{y},Y=y\mid Z=z) \\ = \mathbb{P}(A=\alpha\mid Z=z), \\ \sum_{\alpha\in\mathcal{A}} \tilde{w}_\alpha(\hat{y},y,z) = 1,\ 0 \le \tilde{w}_\alpha(\hat{y},y,z) \le 1, \\ \text{for any } \alpha,z,\hat{y},y. \end{array} \right\}$$

From Propositions 5 and 6, we show in Propositions 7 and 8 that the partial identification sets of DD and TPRD/TNRD for a binary protected class, without imposing extra assumptions, actually have closed-form solutions.

**Proposition 7** (Closed-Form Set for DD)**.** *Let*

$$w^L_\alpha(\hat{y},z) = \max\left\{0, 1 + \frac{\mathbb{P}(A=\alpha\mid Z=z)-1}{\mathbb{P}(\hat{Y}=\hat{y}\mid Z=z)}\right\},$$
$$w^U_\alpha(\hat{y},z) = \min\left\{1, \frac{\mathbb{P}(A=\alpha\mid Z=z)}{\mathbb{P}(\hat{Y}=\hat{y}\mid Z=z)}\right\}.$$

Then,

$$\Delta_{\mathrm{DD}}(\mathcal{P}_D) = \big[\mu(a;w^L) - \mu(b;w^U), \mu(a;w^U) - \mu(b;w^L)\big]. \tag{11}$$

**Proof.** Notice that $[\mathbb{P}(\hat{Y}=\hat{y}\mid Z=z)w^L_\alpha(\hat{y},z), \mathbb{P}(\hat{Y}=\hat{y}\mid Z=z)\, w^U_\alpha(\hat{y},z)]$ are exactly the endpoints of the Fréchet–Hoeffding inequalities in Equation (2) for the coupling set $\Pi(\mathbb{P}(\hat{Y}\mid Z=z), \mathbb{P}(A\mid Z=z))$. According to Propositions 1 and 6, the set $\mathcal{W}(\mathcal{P}_D)$ has the following equivalent formulation:

$$\mathcal{W}(\mathcal{P}_D) = \left\{ w : \begin{array}{c} \sum_{\hat{y}\in\{0,1\}} w_\alpha(\hat{y},z)\mathbb{P}(\hat{Y}=\hat{y}\mid Z=z) = \mathbb{P}(A=\alpha\mid Z=z), \\ \sum_{\alpha\in\mathcal{A}} w_\alpha(\hat{y},z) = 1, w^L_\alpha(\hat{y},z) \le w_\alpha(\hat{y},z) \le w^U_\alpha(\hat{y},z), \\ \text{for any } \alpha,z,\hat{y}. \end{array} \right\} \tag{12}$$

Notice that $\mathcal{W}(\mathcal{P}_D)$ is compact and connected in $L_\infty$ and that the function $\mu(\alpha,w)$ is continuous in $w$ for $\alpha = a,b$. Thus, by Propositions 5 and 6, the partial identification set is an interval:

$$\Delta_{\mathrm{DD}}(\mathcal{P}_D) = \Big[\min_{w\in\mathcal{W}(\mathcal{P}_D)} \mu(a,w) - \mu(b,w), \ \max_{w\in\mathcal{W}(\mathcal{P}_D)} \mu(a,w) - \mu(b,w)\Big]$$

We derive the lower bound as an example, and the upper bound can be derived analogously. According to Equation (7),

$$\mu(a,w) - \mu(b,w) = \frac{\mathbb{E}\big[w_a(\hat{Y},Z)\ \hat{Y}\big]}{\mathbb{P}(A=a)} - \frac{\mathbb{E}\big[w_b(\hat{Y},Z)\ \hat{Y}\big]}{\mathbb{P}(A=b)}. \tag{13}$$

Because $\mu(a,w) - \mu(b,w)$ is increasing in $w_a$ and decreasing in $w_b$, $\min_{w\in\mathcal{W}(\mathcal{P}_D)}(\mu(a,w) - \mu(b,w)) \ge \min_{w\in\mathcal{W}(\mathcal{P}_D)} \mu(a,w) - \max_{w\in\mathcal{W}(\mathcal{P}_D)} \mu(b,w) = \mu(a;w^L) - \mu(b;w^U)$.

Moreover, it is easy to verify that $w^\dagger = (w^L_a, w^U_b)$ satisfies the law of total probability constraints and is feasible in $\mathcal{P}_D$.

The partial identification set given in Proposition 7 concretely illustrates the general unidentifiability of demographic disparity under data combination: any element within the interval in Proposition 11 is a valid disparity value that agrees with the observed data information. In the unrealistically ideal case, if the proxy variables $Z$ are perfectly predictive of either $\hat{Y}$ or $A$, then we can verify that $w^L = w^U$, and the two interval endpoints in Proposition 11 are equal.

We next show that the partial identification sets corresponding to TPRD and TNRD disparity for a binary-valued protected class also have closed-form solutions.

**Proposition 8** (Closed-Form Sets of TPRD and TNRD)**.** *Let*

$$\mu'_{\hat{y}y}(\alpha;\tilde{w},\tilde{w}') := \mathbb{E}[\tilde{w}_\alpha(\hat{Y},Y,Z)\mathbb{I}(Y=y)\mathbb{I}(\hat{Y}=\hat{y})] \times \left(\mathbb{E}\left[\tilde{w}_\alpha(\hat{Y},Y,Z)\mathbb{I}(Y=y)\mathbb{I}(\hat{Y}=\hat{y})\right] + \mathbb{E}\left[\tilde{w}'_\alpha(\hat{Y},Y,Z)\mathbb{I}(Y=y)\mathbb{I}(\hat{Y}\neq\hat{y})\right]\right)^{-1},$$

$$\tilde{w}^L_\alpha(\hat{y},y,z) = \max\left\{0, 1+\frac{\mathbb{P}(A=\alpha|Z=z)-1}{\mathbb{P}(\hat{Y}=\hat{y},\ Y=y|Z=z)}\right\},$$

$$\tilde{w}^U_\alpha(\hat{y},y,z) = \max\left\{1, \frac{\mathbb{P}(A=\alpha|Z=z)}{\mathbb{P}(\hat{Y}=\hat{y},\ Y=y|Z=z)}\right\}.$$

Then,

$$\Delta_{\mathrm{TPRD}}(\mathcal{P}_D) = \left[\mu'_{11}(a;\tilde{w}^L,\tilde{w}^U) - \mu'_{11}(b;\tilde{w}^U,\tilde{w}^L),\ \mu'_{11}(a;\tilde{w}^U,\tilde{w}^L) - \mu'_{11}(b;\tilde{w}^L,\tilde{w}^U)\right], \tag{14}$$

$$\Delta_{\mathrm{TNRD}}(\mathcal{P}_D) = \left[\mu'_{00}(a;\tilde{w}^L,\tilde{w}^U) - \mu'_{00}(b;\tilde{w}^U,\tilde{w}^L),\ \mu'_{00}(a;\tilde{w}^U,\tilde{w}^L) - \mu'_{00}(b;\tilde{w}^L,\tilde{w}^U)\right]. \tag{15}$$

Proposition 8 can be proved by following similar procedures in the proof of Proposition 7. We again leverage a reformulation of $\tilde{\mathcal{W}}(\mathcal{P}_D)$ in terms of Fréchet–Hoeffding inequalities with $\tilde{w}^L$ and $\tilde{w}^U$ as extremal weights. Then $\mu'_{\hat{y}y}(\alpha;\tilde{w},\tilde{w}')$ is continuous in $(\tilde{w},\tilde{w}')$ and is increasing in $\tilde{w}$ but decreasing in $\tilde{w}'$,[9] which would imply that the interval endpoints in Equations (14) and (15) indeed bracket the partial identification sets. It remains to verify that the extremal weights are simultaneously feasible in $\tilde{\mathcal{W}}(\mathcal{P}_D)$ so that the interval endpoints are attained. See Online Appendix B.4 for details. Again, when the proxy variables $Z$ can predict either $(\hat{Y},Y)$ or $A$ perfectly, we can easily verify that $\tilde{w}^L = \tilde{w}^U$, so the intervals in Equations (14) and (15) also collapse into singletons, but this is unrealistic.

## 6. Extensions for General Partial Identification Sets

In this section, we discuss *general* partial identification sets, allowing additional structural assumptions, such as *smoothness restrictions*, and accommodating *multiple-level protected class*.

### 6.1. Additional Smoothness Assumptions

We first introduce smoothness restrictions to illustrate possible additional structural knowledge that can be used to restrict the partial identification sets. One might expect that, for two similar values $z, z'$, the two true joint distributions $\mathbb{P}(A,Y\mid Z=z), \mathbb{P}(A,Y\mid Z=z')$ are also similar (some limited amount of similarity is already implied by the law of total probability when the given marginals are themselves smooth). There is no way to verify this from the separate data sets only, but such an assumption may be defensible based on domain knowledge and can help narrow down the possible values disparities may take. We therefore further consider partial identification sets of disparities when we impose the following additional assumptions:

$$\mathbb{P}(A=\alpha\mid Y=y,Z=z) - \mathbb{P}(A=\alpha\mid Y=y,Z=z') \le d(z,z')\quad \forall\alpha,y,z,z', \tag{16}$$

$$\mathbb{P}(A=\alpha\mid \hat{Y}=\hat{y},Y=y,Z=z) - \mathbb{P}(A=\alpha\mid \hat{Y}=\hat{y},Y=y,Z=z') \le d(z,z')\quad \forall\alpha,\hat{y},y,z,z', \tag{17}$$

where $d(z,z')$ is a given metric. In particular, we encode the implicit Lipschitz constant by scaling the metric $d$ itself. We can then let $\mathcal{P}_{\mathrm{Lip}}$ be the set of all joints that satisfy Equations (16) and (17).

Equations (16) and (17) imply that the weight constraints $\mathcal{W}(\mathcal{P}_D\cap\mathcal{P}_{\mathrm{Lip}})$ and $\tilde{\mathcal{W}}(\mathcal{P}_D\cap\mathcal{P}_{\mathrm{Lip}})$ corresponding to the Lipschitz assumption take the following respective forms:

$$\mathcal{W}(\mathcal{P}_D\cap\mathcal{P}_{\mathrm{Lip}}) = \mathcal{W}(\mathcal{P}_D)\cap\mathcal{W}_{\mathrm{Lip}},\ \text{where } \mathcal{W}_{\mathrm{Lip}} := \{w : w_\alpha(\hat{y},z) - w_\alpha(\hat{y},z') \le d(z,z')\ \forall z,z',\hat{y}\};$$

$$\tilde{\mathcal{W}}(\mathcal{P}_D\cap\mathcal{P}_{\mathrm{Lip}}) = \tilde{\mathcal{W}}(\mathcal{P}_D)\cap\tilde{\mathcal{W}}_{\mathrm{Lip}},\ \text{where } \tilde{\mathcal{W}}_{\mathrm{Lip}} := \{\tilde{w} : \tilde{w}_\alpha(\hat{y},y,z) - \tilde{w}_\alpha(\hat{y},y,z') \le d(z,z')\ \forall z,z',\hat{y},y\}.$$

Leveraging Proposition 5, we can translate this to the partial identification sets for DD, TPRD, and TNRD when we assume Equations (16) and (17). In particular, Proposition 6 and the preceding provide an explicit form for these sets. To actually compute their endpoints we now need to solve an optimization problem. For generality, we consider this optimization problem in the context of a multiple-level protected class attribute, which we study next.

### 6.2. Multiple-Level Protected Class Attribute

We now consider the most general case and study the partial identification set of all *simultaneously achievable* disparities for multiple groups, potentially imposing additional assumptions such as smoothness. Without loss of generality, we will evaluate disparities for multiple levels by designating some element $a$ of $\mathcal{A}$ to be the reference class. Specifically, letting $\mathcal{A}_0 := \mathcal{A}\setminus\{a\}$

and $\delta(a,b;\mathbb{P})$ be any of the disparities defined in Section 2.1, we consider the *multivariate* partial identification of all pairwise disparities:

$$\Delta(\mathcal{P}_D \cap \mathcal{P}_A) = \left\{(\delta(a,b;\mathbb{P}'))_{b\in\mathcal{A}_0} : \mathbb{P}' \in \mathcal{P}_D \cap \mathcal{P}_A\right\} \subset \mathbb{R}^{|\mathcal{A}|-1}. \tag{18}$$

Note that for any $b, b'$, $\delta(b,b';\mathbb{P}') = \delta(a,b';\mathbb{P}') - \delta(a,b;\mathbb{P}')$, so that the above-mentioned set characterizes all simultaneously achievable pairwise disparities, regardless of the choice of reference class $a$. We can also extend the general approach to linear combinations of multiple disparity measures at the same time.

Next, we note that with $\mathcal{W}(\mathcal{P}), \tilde{\mathcal{W}}(\mathcal{P})$ as defined in Equations (9) and (10), we have the following generalization of Proposition 5.

**Proposition 9.** *For any set $\mathcal{P}$ of joint distributions on $(A, \hat{Y}, Y, Z)$, we have $\Delta_{\mathrm{DD}}(\mathcal{P}) = \{(\mu(a;w) - \mu(b;w))_{b\in\mathcal{A}_0} : w \in \mathcal{W}(\mathcal{P})\}$, $\Delta_{\mathrm{TPRD}}(\mathcal{P}) = \{(\mu_{11}(a;\tilde{w}) - \mu_{11}(b;\tilde{w}))_{b\in\mathcal{A}_0} : \tilde{w} \in \tilde{\mathcal{W}}(\mathcal{P})\}$, and $\Delta_{\mathrm{TNRD}}(\mathcal{P}) = \{(\mu_{00}(a;\tilde{w}) - \mu_{00}(b;\tilde{w}))_{b\in\mathcal{A}_0} : \tilde{w} \in \tilde{\mathcal{W}}(\mathcal{P})\}$.*

Because these sets are multivariate, they have more than just two "endpoints." In particular, we characterize these sets by computing their *support functions*. Given a set $\Theta \subseteq \mathbb{R}^d$, its support function is given by $h_\Theta(\rho) = \sup_{\theta\in\Theta} \rho^\top\theta$. First, the support function provides the maximal and minimal contrasts achieved over a set: for example, setting $\rho_b = 1$, $\rho_{b'} = -1$, and $\rho_\alpha = 0$ for $\alpha \neq b, b'$ gives the maximal disparity between groups $b'$ and $b$. Moreover, the support function also characterizes the set itself via its (closed) convex hull (Rockafellar 2015). Specifically,[10]

$$\begin{aligned}\mathrm{Conv}\,(\Theta) := \mathrm{Closure}\left(\left\{\sum_{j=1}^m \lambda_j\theta_j : m\in\mathbb{N}, \theta_j\in\Theta, \right.\right.\\ \left.\left. \lambda_j \geq 0, \sum_{j=1}^m \lambda_j = 1\right\}\right) \\ = \left\{\theta : \rho^\top\theta \leq h_\Theta(\rho),\ \forall \rho \text{ s.t. } \|\rho\| = 1\right\}.\end{aligned} \tag{19}$$

In the following, we characterize the support functions. In Section 7.2, we discuss their computation and estimation from data and how to use this to visualize the partial identification set.

#### 6.2.1. Demographic Disparity.

We first consider the simpler case of demographic disparity.

**Proposition 10.** *Let $\mathcal{P}_A$ be given. Then*

$$h_{\Delta_{\mathrm{DD}}(\mathcal{P}_D\cap\mathcal{P}_A)}(\rho) = \max_{w\in\mathcal{W}(\mathcal{P}_D)\cap\mathcal{W}(\mathcal{P}_A)} \sum_{b\in\mathcal{A}_0} \rho_b\left(\frac{\mathbb{E}[w_a(\hat{Y},Z)\hat{Y}]}{\mathbb{P}(A=a)} - \frac{\mathbb{E}[w_b(\hat{Y},Z)\hat{Y}]}{\mathbb{P}(A=b)}\right).$$

Proposition 10 follows immediately from Proposition 9 and Equation (13). When either $\mathcal{P}_A$ imposes no restrictions or $\mathcal{P}_A = \mathcal{P}_{\mathrm{Lip}}$, the preceding gives an infinite linear program because both the law of total probability constraint $\mathcal{W}(\mathcal{P}_D)$ and the Lipschitz constraint $\mathcal{W}_{\mathrm{Lip}}$ are linear in $w$.

#### 6.2.2. Classification Disparity.

We next consider the case of classification disparities. For a concise and clear exposition, we focus on the case of TPRD. Note that $\Delta_{\mathrm{TPRD}}(\mathcal{P}_D \cap \mathcal{P}_A)$ is generally a nonconvex set. The case of TNRD can be symmetrically handled. Using Equation (8), the support function for multiple-leveled protected class attribute can be written as follows:

$$h_{\Delta_{\mathrm{TPRD}}(\mathcal{P}_D\cap\mathcal{P}_A)}(\rho) = \sup_{\tilde{w}\in\tilde{\mathcal{W}}(\mathcal{P}_D\cap\mathcal{P}_A)} \sum_{b\in\mathcal{A}_0} \rho_b\left(\frac{\mathbb{E}[\tilde{w}_a(\hat{Y},Y,Z)Y\hat{Y}]}{\mathbb{E}[\tilde{w}_a(\hat{Y},Y,Z)Y]} - \frac{\mathbb{E}[\tilde{w}_b(\hat{Y},Y,Z)Y\hat{Y}]}{\mathbb{E}[\tilde{w}_b(\hat{Y},Y,Z)Y]}\right). \tag{20}$$

This optimization is the sum of linear-fractional functions, and optimizing it is generally intractable. We next provide a reformulation as an optimization problem that is a linear program once we fix some parameters, so that the problem is reduced to a search over these parameters.

**Proposition 11.** *Let $\mathcal{P}_A$ be given. Then*

$$\begin{aligned} h_{\Delta_{\mathrm{TPRD}}(\mathcal{P}_D\cap\mathcal{P}_A)}(\rho) &= \max\nolimits_{t\in\mathbb{R}^{\mathcal{A}}:t\geq 1} \phi(\rho;t),\\ \phi(\rho;t) = \max_{\tilde{u}} \sum_{b\in\mathcal{A}_0} \rho_b\Big(&\mathbb{E}[\tilde{u}_a(\hat{Y},Y,Z)Y\hat{Y}] \\ &- \mathbb{E}[\tilde{u}_b(\hat{Y},Y,Z)Y\hat{Y}]\Big)\\ \text{s.t. } &\mathbb{E}[\tilde{u}_\alpha(\hat{Y},Y,Z)Y] = 1,\ \forall\alpha\in\mathcal{A},\\ &\sum_{\alpha\in\mathcal{A}} \frac{\tilde{u}_\alpha(\hat{y},y,z)}{t_\alpha} = 1, \forall\hat{y}\in\{0,1\}, y\in\{0,1\}, z\in\mathcal{Z},\\ \sum_{\hat{y},y\in\{0,1\}} &\tilde{u}_\alpha(\hat{y},y,z)\mathbb{P}(\hat{Y}=\hat{y},Y=y|Z=z) = \mathbb{P}(A=\alpha|Z=z)\ t_\alpha\\ &\qquad\qquad \forall\alpha\in\mathcal{A}, z\in\mathcal{Z},\\ \tilde{u}_\alpha(\hat{y},y,z) \geq 0, &\qquad \forall\alpha\in\mathcal{A},\hat{y}\in\{0,1\},y\in\{0,1\},z\in\mathcal{Z},\\ &\qquad (\tilde{u}_\alpha/t_\alpha)_{\alpha\in\mathcal{A}}\in\tilde{\mathcal{W}}(\mathcal{P}_A). \end{aligned} \tag{21}$$

To obtain this reformulation, we apply a Charnes–Cooper transformation (Charnes and Cooper 1962) to linearize each linear-fractional term in Equation (20). Specifically, for each $\alpha\in\mathcal{A}$, we optimize over the transformed variables, $t_\alpha = 1/\mathbb{E}[\tilde{w}_a(\hat{Y},Y,Z)Y]$, $\tilde{u}_\alpha(\hat{y},y,z) = t_\alpha\tilde{w}_\alpha(\hat{y},y,z)$. See Online Appendix B.5 for the detailed proof. Although Equation (21) is generally a nonconvex optimization problem, the inner problem $\phi(\rho,t)$ is a linear program whenever $\tilde{\mathcal{W}}(\mathcal{P}_A)$ is the product of polyhedra over $\alpha\in\mathcal{A}$, such as $\tilde{\mathcal{W}}(\mathcal{P}_A) = \tilde{\mathcal{W}}_{\mathrm{Lip}}$, or when $\mathcal{P}_A$ is unrestricted (in which case we may omit the last constraint in $\phi(\rho,t)$).

## 7. Implementation, Estimation, and Inference

In this section, we discuss how to implement our approach in practice in order to go from actual data to

assessments of disparities. Specifically, in previous sections, we characterized the partial identification sets for disparity measures in terms of the two population distributions $\mathbb{P}(A, Z)$, $\mathbb{P}(\hat{Y}, Y, Z)$: these sets are *deterministic* population objects that reflect the *intrinsic ambiguity* of disparities given only marginal information. In practice, we are given *data* rather than marginal distributions. The question we address in this section is how to *estimate* the partial identification sets from data. We further discuss the consistency of our estimates and inferential procedures for constructing confidence intervals that characterize the *additional* uncertainty as a result of finite-sample variability.

In this section, instead of assuming access to the population-level marginal distributions, we assume we are given finite-sample data sets. Let $n_{\text{pri}}$ and $n_{\text{aux}}$ denote the sample size of the *primary* and *auxiliary* data sets, respectively. Our combined data set is

$$\left\{\left(\hat{Y}_i, Y_i, Z_i\right)_{i=1}^{n_{\text{pri}}}, \left(A_i, Z_i\right)_{i=n_{\text{pri}}+1}^{n}\right\}, \text{ with total sample size } n = n_{\text{pri}} + n_{\text{aux}},$$

where the first $n_{\text{pri}}$ units form the primary data set, and the latter $n_{\text{aux}}$ units form the auxiliary data set. We suppose the data satisfy Assumption 1 and that observations in these two data sets are independent. We assume that as $n$ grows to infinity, the proportion of the primary data set $r_n = n_{\text{pri}}/n$ converges to a limiting proportion $r$ (i.e., $r_n \to r$). Because the primary data are typically more expensive to acquire than the auxiliary data, we focus on the setting where the primary data set is asymptotically of comparable or smaller size (i.e., $0 \le r < 1$).[11]

According to Sections 5 and 6, the partial identification sets of disparity measures involve the conditional probabilities of the protected class $A$ and the outcomes $\hat{Y}, Y$ given proxies $Z$. We denote these conditional probabilities by the following shorthand notations:

$$\begin{aligned}
\eta_{\text{aux}}(\alpha, z) &:= \mathbb{P}(A = \alpha \mid Z = z),\\
\eta_{\text{pri}}(\hat{y}, z) &:= \mathbb{P}(\hat{Y} = \hat{y} \mid Z = z),\\
\tilde{\eta}_{\text{pri}}(\hat{y}, y, z) &:= \mathbb{P}(\hat{Y} = \hat{y}, Y = y \mid Z = z).
\end{aligned}$$

Along with $\mathbb{P}(Z)$, these specify the marginals $\mathbb{P}(A, Z)$, $\mathbb{P}(\hat{Y}, Y, Z)$. In practice, these conditional probabilities are usually unknown and need to be estimated from the primary and auxiliary data sets, respectively. Because $\eta_{\text{aux}}, \eta_{\text{pri}}, \tilde{\eta}_{\text{pri}}$ are discrete regression functions with features $Z$ (or probabilistic classification models), they can each be learned using supervised learning on each of the data sets. For example, in Section 8, we use logistic and multinomial logistic regression. Other options include random forests or neural networks. Because we are primarily interested in estimating the partial identification sets rather than these conditional probabilities, we refer to these conditional probabilities as *nuisance parameters* and estimators for them as *nuisance estimators*.

### 7.1. Debiased Estimation and Inference for the Case of Binary Protected Class

In Propositions 7 and 8, we prove that the partial identification sets of DD, TPRD, and TNRD for binary protected class are intervals with closed-form endpoints. Therefore, estimating these two endpoints is enough to characterize the whole partial identification set. In this section we study how to estimate and conduct inference on these. In particular, any estimate, even if consistent, still has sampling uncertainty, and so to construct an estimated partial identification set that has guarantees on containing the true disparity measures, we may wish to add sampling-uncertainty confidence intervals *on top* of the estimated endpoints. We here propose a novel debiased estimator with a sampling variance that is easily estimable and can be used to construct confidence intervals. We start with motivating our approach by explaining why a simple plug-in approach would be insufficient, then we present a reformulation of the endpoints that is useful for our estimation approach, then we present our estimator, and finally we present our confidence interval. For simplicity, we only present the estimation and inference results for the partial identification set of DD. The results for TPRD and TNRD are analogous but require more involved notation so we defer them to Online Appendix A.2.

**7.1.1. Motivation.** Proposition 7 characterize the endpoints of $\Delta_{\text{DD}}(\mathcal{P}_D)$ in terms of expectations involving the nuisance functions $\eta_{\text{aux}}, \eta_{\text{pri}}$ in Equation (11). Therefore, one simple approach to estimating these endpoints would be to estimate these nuisance functions and then replace nuisances with estimated nuisances and all expectations with empirical averages in Equation (11). Then, under standard assumptions of consistency of our nuisance estimates and that the set of possible estimates is a Glivenko–Cantelli class (e.g., $\hat{\eta}_{\text{aux}} \in \mathcal{F}_{\text{aux}}$ almost surely and $\mathcal{F}_{\text{aux}}$ is $\mathbb{P}$-Glivenko–Cantelli), the simple plug-in estimator would be consistent; that is, it converges to the true endpoints given in Proposition 7. If we further assume that the set of possible estimates is a Donsker class (e.g, $\mathcal{F}_{\text{aux}}$ is $\mathbb{P}$-Donsker) and certain regularity conditions (e.g., condition (ii) in Theorem 1), then the estimate would also be asymptotically normal, which we might hope to use to construct confidence intervals.

However, there are two crucial concerns with this plug-in approach. The first is that the requirements on the class of possible nuisance estimates can be very

restrictive and exclude the use of flexible nonparametric methods to fit nuisances, such as random forests or kernel estimators. Second, even if we use highly regular nuisance estimators that satisfy the Donsker condition (e.g., logistic regression), estimating the asymptotic variance (e.g., using the delta method) can be very intractable as the gradient of the estimand in the nuisance parameters is nonsmooth, let alone that we need to estimate the variance of the estimated nuisance parameters themselves. Instead, we propose a debiased approach that is slightly more complicated to explain but actually leads to a much simpler inferential algorithm, as the asymptotic distribution of our debiased estimator is independent of how we estimate the nuisances. Additional detail on the failure of the simple plug-in estimator is given in Online Appendix A.1.

**7.1.2. Reformulation of the Estimand.** According to Proposition 7, estimating the partial identification sets of demographic disparity only requires estimating the bounds $\mu(a;w^L)-\mu(b;w^U)$ and $\mu(a;w^U)-\mu(b;w^L)$. In the following lemma, we consider a reformulation of $\mu(\alpha;w^L)$ and $\mu(\alpha;w^U)$ that will be useful for constructing estimators for them.

**Lemma 1.** *For* $\mu(\alpha,\cdot)$ *given in* Equation (7), *and* $w^L$ *and* $w^U$ *given in* Proposition 7,

$$\mu(\alpha,w^L)=\frac{1}{p_\alpha}\Big\{\mathbb{E}\Big[\lambda_\alpha^L(Z;\eta)\Big]+\mathbb{E}\Big[\xi_\alpha^L(A,Z;\eta)\Big]+\mathbb{E}\Big[\gamma_\alpha^L\big(\hat{Y},Z;\eta\big)\Big]\Big\}, \tag{22}$$

$$\mu(\alpha,w^U)=\frac{1}{p_\alpha}\Big\{\mathbb{E}\Big[\lambda_\alpha^U(Z;\eta)\Big]+\mathbb{E}\Big[\xi_\alpha^U(A,Z;\eta)\Big]+\mathbb{E}\Big[\gamma_\alpha^U\Big(\hat{Y},Z;\eta\Big)\Big]\Big\}, \tag{23}$$

*where* $p_\alpha:=\mathbb{P}(A=\alpha)$, $\eta=(\eta_{\mathrm{pri}},\eta_{\mathrm{aux}})$, *and*

$$\begin{aligned}
\lambda_\alpha^L(z;\eta)&=I_\alpha^L(z)\big(\eta_{\mathrm{pri}}(1,z)+\eta_{\mathrm{aux}}(\alpha,z)-1\big),\\
\xi_\alpha^L(A,z;\eta)&:=I_\alpha^L(z)\big(\mathbb{I}(A=\alpha)-\eta_{\mathrm{aux}}(\alpha,z)\big),\\
\gamma_\alpha^L\big(\hat{Y},z;\eta\big)&:=I_\alpha^L(z)\big(\hat{Y}-\eta_{\mathrm{pri}}(1,z)\big),\ \textit{with}\\
I_\alpha^L(z)&:=\mathbb{I}\big(\eta_{\mathrm{pri}}(1,z)+\eta_{\mathrm{aux}}(\alpha,z)-1\ge 0\big),\\
\lambda_\alpha^U(z;\eta)&=I_\alpha^U(z)\big(\eta_{\mathrm{pri}}(1,z)-\eta_{\mathrm{aux}}(\alpha,z)\big)+\eta_{\mathrm{aux}}(\alpha,z),\\
\xi_\alpha^U(A,z;\eta)&:=\big(1-I_\alpha^U(z)\big)\big(\mathbb{I}(A=\alpha)-\eta_{\mathrm{aux}}(\alpha,z)\big),\\
\gamma_\alpha^U\big(\hat{Y},z;\eta\big)&:=I_\alpha^U(z)\big(\hat{Y}-\eta_{\mathrm{pri}}(1,z)\big),\ \textit{with}\\
I_\alpha^U(z)&:=\mathbb{I}\big(\eta_{\mathrm{pri}}(1,z)-\eta_{\mathrm{aux}}(\alpha,z)\le 0\big).
\end{aligned}$$

The variables $I_\alpha^L, I_\alpha^U$ indicate which branch of the max/min we take in the definitions of $w^L, w^U$ in Proposition 7, where the max/min, in turn, arise from the endpoints of the Fréchet–Hoeffding bounds. The variables $\xi_\alpha^L, \gamma_\alpha^L, \xi_\alpha^U, \gamma_\alpha^U$ are residuals of the regressions defining the nuisance functions. It is straightforward to verify that $\mathbb{E}[\xi_\alpha^L(A,Z;\eta)+\gamma_\alpha^L(\hat{Y},Z;\eta)]=\mathbb{E}[\xi_\alpha^U(A,Z;\eta)+\gamma_\alpha^U(\hat{Y},Z;\eta)]=0$ by iterated expectations. So, in a sense, these are not necessary for characterizing $\mu(\alpha,w^L)$ and $\mu(\alpha,w^U)$ and indeed do not appear in Proposition 7. However, incorporating these augmentation terms can debias errors in the main $\lambda_\alpha^L, \lambda_\alpha^U$ terms. Specifically, when we use estimated values of the nuisance parameters $\eta$ instead of their unknown true values, these augmentation terms cancel out first-order bias terms so that estimation errors of $\eta$ only have negligible effect. In Online Appendix A.1, we illustrate that estimators without these augmentation terms generally have intractable asymptotic distributions.

**Algorithm 1** (Estimation of $\Delta_{\mathrm{DD}}(\mathcal{P}_D)$ for a binary-valued protected class attribute)

1: Input: number of folds $K$, nuisance estimation procedures
2: Randomly partition the two data sets into $K$ disjoint even folds:
$\mathcal{I}_{\mathrm{pri}}=\{1,\dots,n_{\mathrm{pri}}\}=\mathcal{I}_{1,\mathrm{pri}}\cup\cdots\cup\mathcal{I}_{K,\mathrm{pri}}$,
$\|\mathcal{I}_{k,\mathrm{pri}}|-n_{\mathrm{pri}}/K|\le 1$,
$\mathcal{I}_{\mathrm{aux}}=\{n_{\mathrm{pri}}+1,\dots,n\}=\mathcal{I}_{1,\mathrm{aux}}\cup\cdots\cup\mathcal{I}_{K,\mathrm{aux}}$,
$\|\mathcal{I}_{k,\mathrm{aux}}|-n_{\mathrm{aux}}/K|\le 1$.
3: Set $\mathcal{I}_k=\mathcal{I}_{k,\mathrm{pri}}\cup\mathcal{I}_{k,\mathrm{aux}}$, and let $\hat{\mathbb{E}}_k$, $\hat{\mathbb{E}}_{k,\mathrm{pri}}$, and $\hat{\mathbb{E}}_{k,\mathrm{aux}}$ be the sample averages over the $k$th fold in the combined, primary, and auxiliary data sets, respectively. For example, $\hat{\mathbb{E}}_k\lambda_\alpha^L(Z;\eta)=\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\lambda_\alpha^L(Z_i;\eta)$.
4: Set $\hat{p}_\alpha=\frac{1}{K}\sum_{k=1}^K\hat{\mathbb{E}}_{k,\mathrm{aux}}[\mathbb{I}(A=\alpha)]$.
5: **for** $k=1,\dots,K$ **do**:
6: Train $\hat{\eta}_{\mathrm{pri}}^{-k}$ on $\{(\hat{Y}_i,Z_i):i\in\mathcal{I}_{\mathrm{pri}}\setminus\mathcal{I}_{k,\mathrm{pri}}\}$.
7: Train $\hat{\eta}_{\mathrm{aux}}^{-k}$ on $\{(A_i,Z_i):i\in\mathcal{I}_{\mathrm{aux}}\setminus\mathcal{I}_{k,\mathrm{aux}}\}$.
8: Set $\hat{\eta}^{-k}=(\hat{\eta}_{\mathrm{pri}}^{-k},\hat{\eta}_{\mathrm{aux}}^{-k})$.
9: **for** $\alpha\in\mathcal{A}$ **do**: compute

$$\hat{\mu}(\alpha,w^L)=\frac{1}{\hat{p}_\alpha K}\sum_{k=1}^K\Big\{\hat{\mathbb{E}}_k\big[\lambda_\alpha^L(Z;\hat{\eta}^{-k})\big]+\hat{\mathbb{E}}_{k,\mathrm{aux}}\big[\xi_\alpha^L(A,Z;\hat{\eta}^{-k})\big]+\hat{\mathbb{E}}_{k,\mathrm{pri}}\big[\gamma_\alpha^L\big(\hat{Y},Z;\hat{\eta}^{-k}\big)\big]\Big\}, \tag{24}$$

$$\hat{\mu}(\alpha,w^U)=\frac{1}{\hat{p}_\alpha K}\sum_{k=1}^K\Big\{\hat{\mathbb{E}}_k\big[\lambda_\alpha^U(Z;\hat{\eta}^{-k})\big]+\hat{\mathbb{E}}_{k,\mathrm{aux}}\big[\xi_\alpha^U(A,Z;\hat{\eta}^{-k})\big]+\hat{\mathbb{E}}_{k,\mathrm{pri}}\big[\gamma_\alpha^U\big(\hat{Y},Z;\hat{\eta}^{-k}\big)\big]\Big\}. \tag{25}$$

10: Return the estimated partial identification set

$$\hat{\Delta}_{\mathrm{DD}}(\mathcal{P}_D)=\big[\hat{\mu}(a,w^L)-\hat{\mu}(b,w^U),\ \hat{\mu}(a,w^U)-\hat{\mu}(b,w^L)\big]. \tag{26}$$

**7.1.3. The Estimator.** Our estimator for the partial identification set is given in Algorithm 1.

Our estimates for $\hat{\mu}(\alpha,w^L)$ and $\hat{\mu}(\alpha,w^U)$ are based on Equations (22) and (23) and a cross-fitting strategy:

the nuisance estimator $\hat{\eta}^{-k}$ is only applied to data in the $k$th fold (i.e., data not used to train $\hat{\eta}^{-k}$). This prevents the nuisance estimators from overfitting to the data where they are evaluated (Chernozhukov et al. 2018).

**7.1.4. Inference.** We next prove that the estimated endpoints in Equation (26) are asymptotically normal with closed-form asymptotic variance. This allows us to construct confidence intervals. It also shows that we are largely invariant to how one fits $\eta$ and that no conditions except for a slow convergence rate are needed, which is appealing when one uses machine learning methods for this task.

**Theorem 1.** *Suppose that the nuisance estimators converge at the following rate:*

$$\left|\hat{\eta}^{-k}_{\text{pri}}(1,Z)-\eta_{\text{pri}}(1,Z)\right| = O_p\left(\kappa_{n_{\text{pri}},\hat{Y}}\right),$$
$$\left|\hat{\eta}^{-k}_{\text{aux}}(\alpha,Z)-\eta_{\text{aux}}(\alpha,Z)\right| = O_p\left(\kappa_{n_{\text{aux}},A}\right),$$
$$\alpha = a,b,\ k = 1,\dots,K.$$

Assume the following conditions: for $\alpha = a, b$,

i. $p_\alpha > 0$;

ii. *there exist positive constants* $m_1, m_2, c_1, c_2$ *such that for* $\alpha \in \mathcal{A}$ *and any* $p \geq 0$,

$$\mathbb{P}\left(0 \leq \left|\eta_{\text{pri}}(1,Z)+\eta_{\text{aux}}(\alpha,Z)-1\right| \leq p\right) \leq c_1 p^{m_1},$$
$$\mathbb{P}\left(0 \leq \left|\eta_{\text{pri}}(1,Z)-\eta_{\text{aux}}(\alpha,Z)\right| \leq p\right) \leq c_2 p^{m_2};$$

iii. $\max\{\kappa_{n_{\text{aux}},A},\ \kappa_{n_{\text{pri}},\hat{Y}}\} = o(n_{\text{pri}}^{-1/(2+2m_1)})$, *and* $\max\{\kappa_{n_{\text{aux}},A}, \kappa_{n_{\text{pri}},\hat{Y}}\} = o(n_{\text{pri}}^{-1/(2+2m_2)})$;

iv. $|r-r_n|\kappa_{n_{\text{pri}},\hat{Y}} = o(n_{\text{pri}}^{-1/2})$, *and* $|r-r_n|\kappa_{n_{\text{aux}},A} = o(n_{\text{pri}}^{-1/2})$.

Then, as $n \to \infty$, the lower bound and upper bound estimators for demographic disparity with binary protected class are asymptotically normal:

$$\sqrt{n_{\text{pri}}}\left\{\left(\hat{\mu}(a,w^L)-\hat{\mu}(b,w^U)\right)-\left(\mu(a,w^L)-\mu(b,w^U)\right)\right\} \xrightarrow{d} \mathcal{N}(0,V_L), \tag{27}$$

$$\sqrt{n_{\text{pri}}}\left\{\left(\hat{\mu}(a,w^U)-\hat{\mu}(b,w^L)\right)-\left(\mu(a,w^U)-\mu(b,w^L)\right)\right\} \xrightarrow{d} \mathcal{N}(0,V_U), \tag{28}$$

where

$$\begin{aligned} V_L = {} & r\mathbb{E}\left[\lambda^L_a(Z;\eta)/p_a - \lambda^U_b(Z;\eta)/p_b \right.\\ & \left. -\left(\mu(a,w^L)-\mu(b,w^U)\right)\right]^2 \\ & + \mathbb{E}\left[\gamma^L_a(\hat{Y},Z;\eta)/p_a - \gamma^U_b(\hat{Y},Z;\eta)/p_b\right]^2 \\ & + \frac{r}{1-r}\mathbb{E}\left[\xi^L_a(A,Z;\eta)/p_a - \xi^U_b(A,Z;\eta)/p_b\right]^2, \\ V_U = {} & r\mathbb{E}\left[\lambda^U_a(Z;\eta)/p_a - \lambda^L_b(Z;\eta)/p_b \right.\\ & \left. -\left(\mu(a,w^U)-\mu(b,w^L)\right)\right]^2 \\ & + \mathbb{E}\left[\gamma^U_a(\hat{Y},Z;\eta)/p_a - \gamma^L_b(\hat{Y},Z;\eta)/p_b\right]^2 \\ & + \frac{r}{1-r}\mathbb{E}\left[\xi^U_a(A,Z;\eta)/p_a - \xi^L_b(A,Z;\eta)/p_b\right]^2. \end{aligned}$$

Condition (i) is needed for the problem to be well defined: both classes need to be present to compare them. Condition (ii) is a margin condition (Audibert and Tsybakov 2007) that characterizes the probability mass near the nondifferentiable boundary. In particular, for $p = 0$, it implies that $\eta_{\text{pri}}(1,Z) + \eta_{\text{aux}}(\alpha,Z) - 1 \neq 0$ and $\eta_{\text{pri}}(1,Z) - \eta_{\text{aux}}(\alpha,Z) \neq 0$ almost surely, which is trivially satisfied if $Z$ includes continuous variables. This ensures that even though $w^L$ and $w^U$ depend on nonsmooth max and min operators, respectively, $\mu(\alpha,w^L)$ and $\mu(\alpha,w^U)$ are still smooth functionals of the conditional probabilities $\eta$. Otherwise, statistical inference for nonsmooth functionals is a notoriously difficult nonregular problem, and it is well known that no estimator with well-behaved asymptotic distribution exits in this case (e.g., Hirano and Porter 2012 and Laber et al. 2014). Similar regularity conditions also appear in other partial identification literature to circumvent nonsmoothness (e.g., Kennedy et al. 2018 and Bonvini and Kennedy 2019). Condition (iii) requires that our nuisance estimators are consistent but only requires a slow, nonparametric rate (i.e., slower than $n_{\text{pri}}^{-1/2}$). For example, if $m_1, m_2 \geq 1$, then condition (iii) is satisfied if $\kappa_{n_{\text{aux}},A} = o_p(n_{\text{pri}}^{-1/4})$ and $\kappa_{n_{\text{aux}},\hat{Y}} = o_p(n_{\text{pri}}^{-1/4})$. This slow rate together with no other assumptions on our nuisance estimators means that the theorem holds even when we use flexible machine learning models to estimate nuisances (e.g., random forest, gradient boosting tree, neural networks with many neurons relative to $n_{\text{pri}}$). Finally, condition (iv) requires that the observed ratio of primary to auxiliary data, $r_n$, is sufficiently similar to the asymptotic ratio. It is trivially satisfied if $r_n = r$ or $r_n - r = O(n_{\text{pri}}^{-1/2})$, such as would be the case if $n_{\text{pri}} \sim \text{Binomial}(n,r)$.

In the proof (in Online Appendix B.6), we show that the asymptotic distributions in Equations (27) and (28) are actually the same as distributions of the infeasible oracle estimators where we use the true values of nuisances, $\eta$. In other words, using the estimated value $\hat{\eta}$ instead of the unknown true value $\eta$ does not inflate the variance of our estimates. This is possible mainly because of the augmented formulation we derive in Lemma 1 (see Online Appendix A.1).

The closed-form asymptotic variances in Theorem 1 suggest the following variance estimators:

$$\begin{aligned} \hat{V}_L = {} & \tfrac{r_n}{K}\sum_{k=1}^{K}\hat{\mathbb{E}}_k\left[\lambda^L_a(Z;\hat{\eta}^{-k})/\hat{p}_a - \lambda^U_b(Z;\hat{\eta}^{-k})/\hat{p}_b \right.\\ & \left. -\left(\hat{\mu}(a,w^L)-\hat{\mu}(b,w^U)\right)\right]^2 \\ & + \frac{1}{K}\sum_{k=1}^{K}\hat{\mathbb{E}}_{k,\text{pri}}\left[\gamma^L_a(\hat{Y},Z;\hat{\eta}^{-k})/\hat{p}_a - \gamma^U_b(\hat{Y},Z;\hat{\eta}^{-k})/\hat{p}_b\right]^2 \\ & + \tfrac{r_n}{1-r_n}\tfrac{1}{K}\sum_{k=1}^{K}\hat{\mathbb{E}}_{k,\text{aux}}\left[\xi^L_a(A,Z;\hat{\eta}^{-k})/\hat{p}_a - \xi^U_b(A,Z;\hat{\eta}^{-k})/\hat{p}_b\right]^2, \end{aligned} \tag{29}$$

and $\hat{V}_U$ is similarly defined by swapping $L$ and $U$ everywhere above.

We further prove in the following theorem that these variance estimators are consistent, and they can be used to construct confidence intervals for the partial identification sets.

**Theorem 2.** *Under the assumptions of* Theorem 1, $\hat{V}_L, \hat{V}_U$ *are consistent: as* $n_{\text{pri}} \to \infty$,

$$\hat{V}_L \xrightarrow{p} V_L, \quad \hat{V}_U \xrightarrow{p} V_U.$$

Therefore, we can construct the following $(1-\beta)\times 100\%$ confidence interval:

$$\text{CI} = \Big[\hat{\mu}(a, w^L) - \hat{\mu}(b, w^U) - \Phi^{-1}(1-\beta/2)\hat{V}_L^{1/2}/n_{\text{pri}}^{1/2},\ \hat{\mu}(a, w^U) - \hat{\mu}(b, w^L) + \Phi^{-1}(1-\beta/2)\hat{V}_U^{1/2}/n_{\text{pri}}^{1/2}\Big]$$

where $\Phi^{-1}$ is the quantile function of standard normal distribution. This confidence interval asymptotically covers the partial identification set of DD with probability at least $1-\beta$:

$$\liminf_{n_{\text{pri}}\to\infty} \mathbb{P}(\Delta_{\text{DD}}(\mathcal{P}_D) \subseteq \text{CI}) \geq 1-\beta.$$

In Section 8 (Figures 4 and 8), we illustrate how to use these confidence intervals to test whether a given disparity value (or a range) is compatible with the observed data information and thus belongs to the corresponding partial identification set.

Note that the above-mentioned confidence interval is *conservative* in that its asymptotic coverage may *exceed* $1-\beta$. In Online Appendix A.4, we present a calibrated confidence interval with asymptotic coverage *exactly* $1-\beta$, albeit having a more complicated form.

## 7.2. General Partial Identification Sets

We next discuss finite-sample estimation of general partial identification sets given in Section 6. That is, we discuss how we obtain a representation of the partially identified sets $\Delta(\mathcal{P}_D \cap \mathcal{P}_A)$ when we either consider a multiple-level protected class attribute or impose smoothness restrictions in $\mathcal{P}_A$, or both. We propose an estimator for the support function using a linear program and prove it is statistically consistent.[12] We then describe how to use these support function estimates to visualize $\text{Conv}(\Delta(\mathcal{P}_D \cap \mathcal{P}_A))$. For this section, we employ a simpler plug-in estimator based on nuisance estimators constructed on the whole primary and auxiliary data sets, respectively.

### 7.2.1. Demographic Disparity.

We first introduce the support function estimator for the case of demographic disparity, $h_{\Delta_{\text{DD}}(\mathcal{P}_D\cap\mathcal{P}_A)}(\rho)$. The estimator applies for the case of multiple-leveled protected attributes with any linearly representable additional constraints $\mathcal{W}(\mathcal{P}_A)$, such as none or $\mathcal{W}_{\text{Lip}}$. Given nuisance estimators $\hat{\eta}_{\text{aux}}, \hat{\eta}_{\text{pri}}$ and letting $\hat{\mathbb{E}}_p$ denote computing sample averages over the primary data set, we define our estimator as the following linear program:

$$\begin{aligned}
&\hat{h}_{\Delta_{\text{DD}}(\mathcal{P}_D\cap\mathcal{P}_A)}(\rho)\\
&\quad= \max_{w} \sum_{b\in\mathcal{A}_0} \rho_b \left(\frac{\hat{\mathbb{E}}_p[w_a(\hat{Y},Z)\hat{Y}]}{\hat{\mathbb{E}}_p[\hat{\eta}_{\text{aux}}(a,Z)]} - \frac{\hat{\mathbb{E}}_p[w_b(\hat{Y},Z)\hat{Y}]}{\hat{\mathbb{E}}_p[\hat{\eta}_{\text{aux}}(b,Z)]}\right)\\
&\qquad\text{s.t.}\quad 0 \leq w_\alpha(\hat{y},z) \leq 1,\ \forall \alpha\in\mathcal{A}, \hat{y}\in\{0,1\},\\
&\qquad\qquad\qquad\qquad y\in\{0,1\}, z\in\{Z_i\}_{i=1}^n\\
&\qquad\qquad \sum_{\hat{y}\in\{0,1\}} w_\alpha(\hat{y},z)\hat{\eta}_{\text{pri}}(\hat{y},z) = \hat{\eta}_{\text{aux}}(\alpha,z),\\
&\qquad\qquad \sum_{\alpha\in\mathcal{A}} w_\alpha(\hat{y},z) = 1, \quad w\in\mathcal{W}(\mathcal{P}_A).
\end{aligned}$$

We next show that the estimator is consistent.

**Theorem 3.** *Assume that*

i. $\sup_{\hat{y}\in\{0,1\},z\in\mathcal{Z}} |\hat{\eta}_{\text{pri}}(\hat{y},z) - \eta_{\text{pri}}(\hat{y},z)| = o_p(1)$ *and* $\sup_{\alpha\in\mathcal{A},z\in\mathcal{Z}} |\hat{\eta}_{\text{aux}}(\alpha,z) - \eta_{\text{aux}}(\alpha,z)| = o_p(1)$, *and*

ii. $Z$ *has finite support* (*i.e.*, $|\mathcal{Z}|$ *is finite*).

*Then, for any* $\rho$,

$$\hat{h}_{\Delta_{\text{DD}}(\mathcal{P}_D\cap\mathcal{P}_A)}(\rho) - h_{\Delta_{\text{DD}}(\mathcal{P}_D\cap\mathcal{P}_A)}(\rho) \xrightarrow{p} 0.$$

Proving Theorem 3 uses a stability analysis attributable to Robinson (1975) to bound the deviation of a linear program under stochastic perturbations to coefficients of the constraint matrix that arise from the estimation errors of the nuisance functions $\hat{\eta}_{\text{aux}}(\alpha,z), \hat{\eta}_{\text{pri}}(\hat{y},z)$. In Proposition EC.1 of the online appendix, we discuss how to additionally obtain the asymptotic distribution of $\hat{h}_{\Delta_{\text{DD}}(\mathcal{P}_D\cap\mathcal{P}_A)}$, under the assumption of unique primal and dual solutions.

### 7.2.2. Classification Disparity.

We next handle the general case for TPRD (TNRD is handled symmetrically). Estimating the support function of $\Delta_{\text{TPRD}}$ introduces additional challenges because the optimization problem that defines it is generally nonconvex (see Proposition 11). We instead leverage the fact that it is the maximum of linear programs if $\tilde{\mathcal{W}}(\mathcal{P}_A)$ is linearly representable.

The estimator for the support function, which computes the sample-level subproblem $\hat{\phi}(\rho;t)$ for a collection of values of $t$, $\mathcal{T}\subseteq\mathbb{R}^{|\mathcal{A}|}$, and the nuisance estimators $\hat{\tilde{\eta}}_{\text{pri}}(\hat{y},y,z), \hat{\eta}_{\text{aux}}(\alpha,z)$ is

$$\hat{h}_{\Delta_{\text{TPRD}}(\mathcal{P}_D\cap\mathcal{P}_A)}(\rho;\mathcal{T}) = \max_{t\in\mathcal{T}} \hat{\phi}(\rho;t), \tag{30}$$

$$\hat{\phi}(\rho;t) = \max_{\tilde{u}} \sum_{b\in\mathcal{A}_0} \rho_b\big(\hat{\mathbb{E}}_p\big[\tilde{u}_a\big(\hat{Y},Y,Z\big)Y\hat{Y}\big] - \hat{\mathbb{E}}_p\big[\tilde{u}_b\big(\hat{Y},Y,Z\big)Y\hat{Y}\big]\big)$$

$$\text{s.t. } \forall \alpha,\hat{y},y,z\in\{Z_i\}_{i=1}^n, \hat{\mathbb{E}}_p\big[\tilde{u}_\alpha\big(\hat{Y},Y,Z\big)Y\big]=1,$$

$$(\tilde{u}_\alpha/t_\alpha)_{\alpha\in\mathcal{A}}\in\tilde{\mathcal{W}}(\mathcal{P}_A),$$

$$\sum_{\hat{y},y\in\{0,1\}} \tilde{u}_\alpha\big(\hat{y},y,z\big)\hat{\tilde{\eta}}_{\text{pri}}(\hat{y},y,z) = t_\alpha\hat{\eta}_{\text{aux}}(\alpha,z),$$

$$\sum_{\alpha\in\mathcal{A}} \frac{\tilde{u}_\alpha\big(\hat{y},y,z\big)}{t_\alpha}=1, \tag{31}$$

$$\tilde{u}_\alpha\big(\hat{y},y,z\big)\geq 0. \tag{32}$$

In the following theorem, we show that the proposed support function estimator is pointwise consistent if $Z$ has only finitely many values and the nuisance estimators are uniformly consistent.

**Theorem 4.** *Assume that*

i. $\sup_{\hat{y}\in\{0,1\},y\in\{0,1\}z\in\mathcal{Z}} |\hat{\tilde{\eta}}_{\text{pri}}(\hat{y},y,z) - \tilde{\eta}_{\text{pri}}(\hat{y},y,z)| \xrightarrow{p} 0$ *and* $\sup_{\alpha\in\mathcal{A},z\in\mathcal{Z}} |\hat{\eta}_{\text{aux}}(\alpha,z) - \eta_{\text{aux}}(\alpha,z)| \xrightarrow{p} 0$;

ii. *there exists a positive constant* $\nu$ *such that* $\mathbb{P}(A=\alpha, Y=1)\geq\nu$, $\forall\alpha\in\mathcal{A}$;

iii. $\mathcal{T}$ *is the componentwise inverse of a set* $\mathcal{T}^{-1}$, *where* $\mathcal{T}^{-1}$ *is an* $\epsilon_{n_{\text{pri}}}$*-covering of* $\mathcal{T}_0^{-1} := \{\tau\in\mathbb{R}^{|\mathcal{A}|}: \sum_{\alpha\in\mathcal{A}}\tau_\alpha = \hat{\mathbb{E}}_p[Y];\ \nu\leq\tau_\alpha\leq 1,\ \alpha\in\mathcal{A}\}$ (*i.e.*, $\min_{\tau'\in\mathcal{T}^{-1}}\|\tau-\tau'\|_1 \leq \epsilon_{n_{\text{pri}}}$ *for any* $\tau\in\mathcal{T}_0^{-1}$);

iv. $Z$ *has finite support* (*i.e.*, $|\mathcal{Z}|$ *is finite*); *and*

v. $\epsilon_{n_{\text{pri}}}\to 0$ *as* $n_{\text{pri}}\to\infty$.

*Then, for any* $\rho$,

$$\hat{h}_{\Delta_{\text{TPRD}}(\mathcal{P}_D\cap\mathcal{P}_A)}(\rho;\mathcal{T}) - h_{\Delta_{\text{TPRD}}(\mathcal{P}_D\cap\mathcal{P}_A)}(\rho)\xrightarrow{p}0.$$

The proof of Theorem 4 is similar to that of Theorem 3 but also shows that the optimization problem is stable under approximation errors from the discretization, $\mathcal{T}$. Condition (ii) ensures that we may restrict attention to a compact range for $t$. Condition (v) ensures consistency as we consider a sequence of finer $t$-grids.

**Algorithm 2** (Estimation of Conv $(\Delta)$ from support function estimates)

1: Input: Support function estimator $\hat{h}_\Delta(\rho)$, contrast sample size $N_\rho$
2: Sample contrast vectors, $\rho_1,\ldots,\rho_{N_\rho}$, uniformly from the $(|\mathcal{A}|-1)$-dimensional unit sphere.
3: **for** $j=1,\ldots,N_\rho$ **do**:
4: Solve $\hat{h}_\Delta(\rho_j)$, record the maximizer $\hat{\delta}_j\in\Delta$ such that $\hat{h}_\Delta(\rho_j)=\hat{\delta}_j^\top\rho_j$.
5: Return
$\hat{\Delta}_{\text{inner}} = \text{Conv}\ (\{\hat{\delta}_1,\ldots,\hat{\delta}_{N_\rho}\})$,
$\hat{\Delta}_{\text{outer}} = \{\delta\in\mathbb{R}^{\mathcal{A}_0}: \delta^\top\rho_j\leq\hat{h}_\Delta(\rho_j)\ \forall j=1,\ldots,N_\rho\}$.

**7.2.3. Estimating and Visualizing the Partial Identification Set.** The above-mentioned procedures estimate the support function of the partial identification set. It remains to actually estimate the partial identification set itself. Given a support function estimator, Algorithm 2 provides a procedure to obtain inner and outer approximations to the set (up to vanishing estimation errors in the support function) by sampling the contrast directions, $\rho$. Because the convex hull of a set is given by the hyperplanes defined by the support function in all directions (see Equation (19)), the outer approximation is given by considering a finite subset of directions. Moreover, because the points realizing the support function in these directions are all in the set, the inner approximation is given by considering only the convex hull of this finite subset. As the number of contrasts sampled increases, the inner and outer approximations become closer. Either set can be visualized using standard tools for plotting convex hulls and polyhedra. We recommend using the outer approximation because (up to vanishing estimation errors in the support function) it is guaranteed to contain the true partial identification set, and this is the set we use in Section 8.

## 8. Case Studies

In the subsequent sections we consider applying our results and methods in two different case studies: mortgage credit decisioning and personalized warfarin dosing. The replication code is available at https://github.com/CausalML/FairnessWithUnobservedProtectedClass.

### 8.1. Mortgage Credit Decisioning

We consider assessing demographic disparity—the simplest measure (see Sections 2.1 and 3 for others) that is relevant for the context of mortgage credit decisioning (Zhang 2016, Chen et al. 2019): here, it measures the discrepancy in marginal approval rates between different racial groups. For groups, we consider White, Black, and Asian and Pacific Islander (API).

**8.1.1. Data Set, Proxy Variables, and Nuisance Estimation.** We demonstrate the partial identification set of demographic disparity using the public HMDA (Home Mortgage Disclosure Act) data set for U.S. mortgage market. This data set contains self-reported race labels, and it has been used in the literature to evaluate proxy methods for race (Baines and Courchane 2014, Zhang 2016, Chen et al. 2019).[13] However, this data set is anonymized and does not include surname information, so we could not evaluate the popular BISG method exactly; it also does not contain default outcomes, so we only study demographic disparity in loan application approval.

We use a random 0.1% subsample containing 14,903 loan application records for White, Black, and API applicants with annual income no more than $100,000 during 2011–2012 as the primary data set and the full sample of all records in this population as the auxiliary data set. This mimics the fact that in BISG, the primary data set typically only contains information of a subset of units in the auxiliary data (decennial census data). We set $\hat{Y}=1$ if a loan application was approved or originated and $\hat{Y}=0$ if it was denied.

We consider three different set of proxy variables for race: only geolocation (county), only annual income, and both geolocation and annual income. The distribution of race/ethnicity by these proxies can both be estimated from public records. U.S. Census Summary File I (U.S. Census Bureau 2010) contains race distributions for different geolocation levels, and the Annual Population Survey (U.S. Census Bureau 2018) contains race distributions for different income brackets.

We estimate the conditional probabilities of race and decision outcome directly on the auxiliary data set. When only geolocation is used as the proxy variable, we use the within-county race proportions and average loan acceptance rate to estimate the conditional probabilities of race and loan acceptance, respectively. When only income is used as the proxy variable, we fit a logistic regression to estimate the conditional probability of loan acceptance and a multinomial logistic regression to estimate the conditional probabilities of races. When both income and geolocation are used, we fit the logistic and multinomial logistic regressions with respect to income within each county.

Recall that the size of the partial identification set depends on the informativeness of the proxies about both protected class and outcomes (Section 4.2). In Figure 3, we show the histograms of the conditional probabilities for each race and, separately, for the positive outcome. We also report the (negative) entropy, which summarizes how predictive the proxies are. For example, the entropy for race probabilities is $\mathbb{E}[\sum_{\alpha\in\mathcal{A}}\mathbb{P}(A=\alpha\mid Z)\log\mathbb{P}(A=\alpha\mid Z)]/|\mathcal{A}|$. Smaller entropy means that the race probabilities are more concentrated toward 0 or 1, which indicates more predictive proxies. We find that, in terms of outcome, all proxies are equally uninformative (the entropy without using any proxy is about 0.5). In terms of protected class, we find geolocation more informative than income and that combining them adds very little.

**8.1.2. Binary Comparisons.** Figure 4 demonstrates estimates of closed-form bounds of demographic disparities of one race versus the others[14] without any extra assumptions (Proposition 7), as well as the associated confidence intervals. By recognizing that in case studies such as the BISG proxy, the auxiliary data set typically describes the whole population (e.g., the whole U.S. population in decennial census data), we use an alternative estimator and confidence interval in Online Appendix A.5 that assumes that the true

**Figure 3.** (Color online) Histograms of Conditional Probabilities of Outcomes (Upper Row) and Race (Lower Row) for Different Choices of Proxies in the HMDA Data Set, Along with the Resulting Entropy

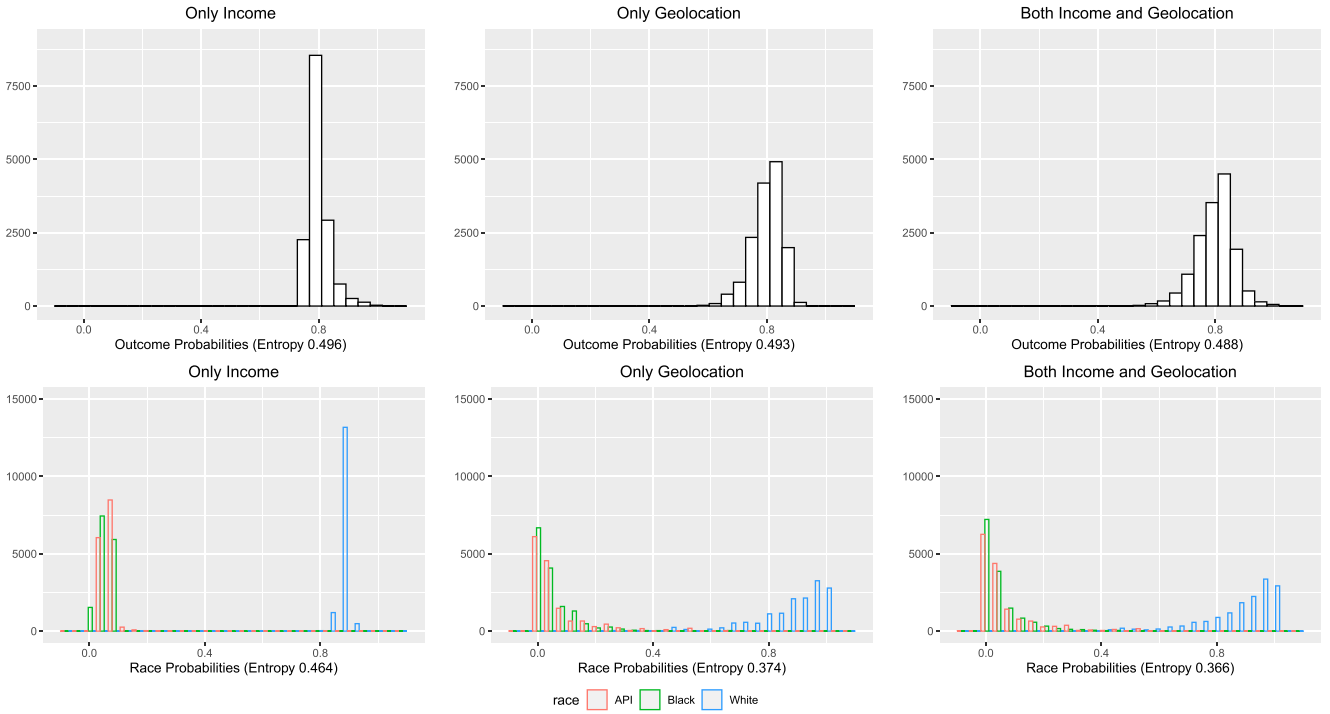
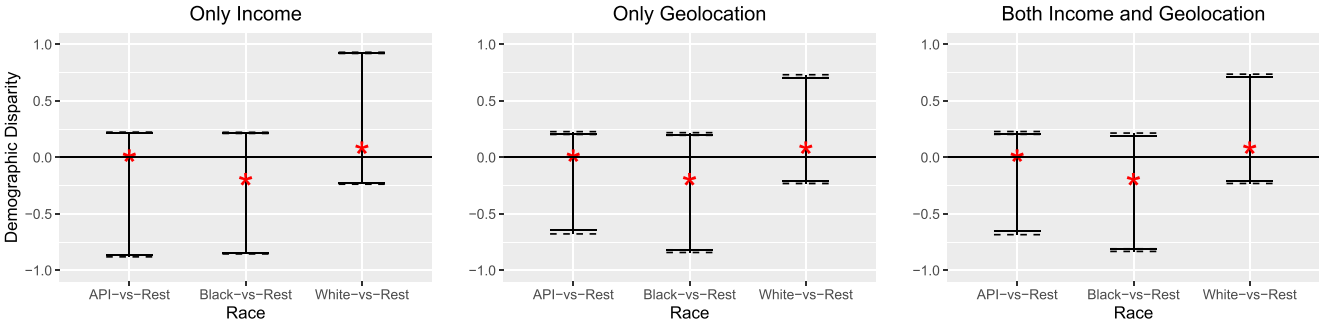


**Figure 4.** (Color online) Partial Identification Bounds of Demographic Disparity (Proposition 7) for Different Proxy Variables in the HMDA Data Set

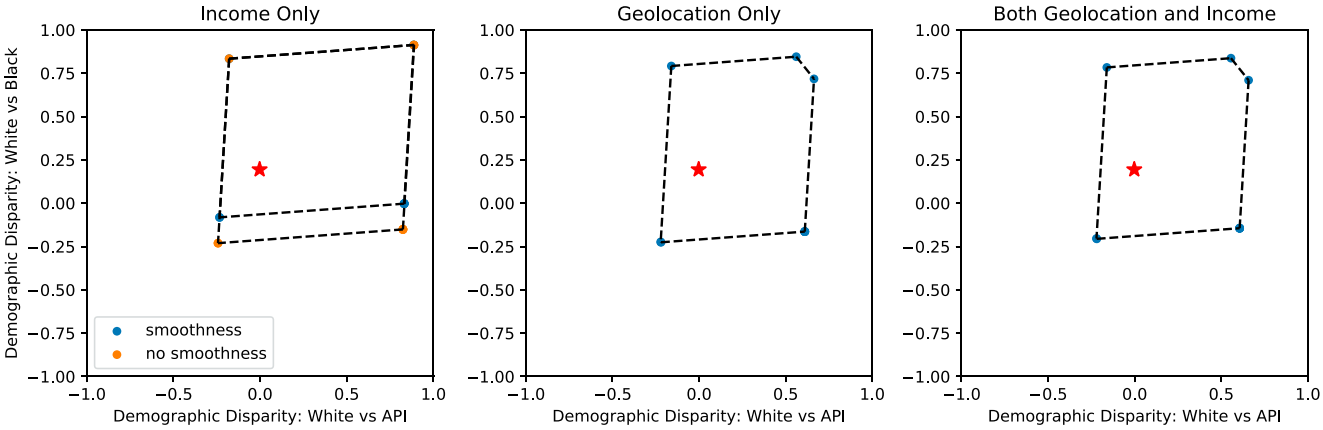



*Notes.* Solid bars represent the estimates of the bounds, and dashed bars indicate 95% confidence intervals. The true value based on self-reported race is shown as an asterisk.

conditional race probabilities (but not the conditional outcome probabilities) are exactly known from the auxiliary data set. This figure also shows the true demographic disparity computed based on the self-reported race using the full data directly. We can observe that overall all estimated partial identification intervals are fairly wide, and all of them correctly contain the ground-truth demographic disparity. Moreover, the finite-sample uncertainty of these estimates is quite small, and the confidence intervals show that at a 5% significance level, we cannot reject zero as a valid disparity value according to the observed data information. This also highlights that identification uncertainty, which does not vanish as we collect more data, generally dominates the sampling uncertainty, which does vanish.

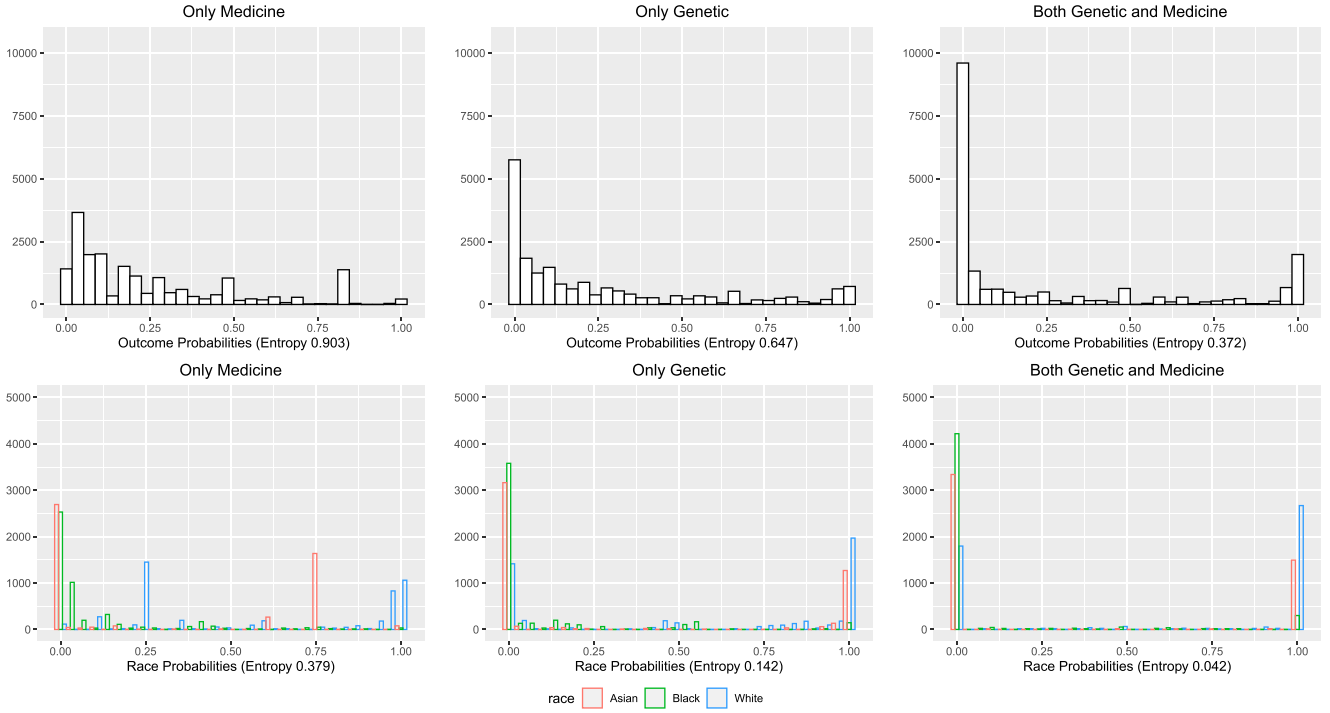
**8.1.3. Multiple-Level Protected Class and Extra Smoothness Assumption.** Figure 5 shows the estimated partial identification sets of the demographic disparities of White versus each other group. The sets are computed by the support function approach described in Section 7.2. For the income-only proxy, we show the partial identification sets both without the smoothness constraint and with the smoothness constraint, where the Lipschitz constant is set as the minimal one such that the constraint set $\mathcal{W}(\mathcal{P}_D) \cap \mathcal{W}(\mathcal{P}_A)$ is still feasible.[15] Smoothness constraints are implemented by enforcing the constraint of Equation (16) on the weight function, whereas the pairwise distance $d(z, z')$ can be computed efficiently for all observed values of the proxy variables. The figure shows that using income as the only proxy, without additional smoothness constraints, seems quite weak in terms of identifying the demographic disparity. Income-only proxy without smoothness results in the largest partial identification set, and using income on top of geolocation barely shrinks the partial identification set relative to the set from using only the geolocation proxy. Adding the smoothness constraint indeed shrinks the partial identification set of income-only proxy, and given that we are willing to assume smoothness, it shows that the White group

**Figure 5.** (Color online) The Outer Approximation of Partial Identification Set for Demographic Disparity in Loan Approval Rates in the HMDA Data Set as Determined by Different Proxies




*Notes.* Positive values correspond to disparity in favor of White. The true demographic disparity is shown as a star.

**Figure 6.** (Color online) Histograms of Conditional Probabilities of Outcomes (Upper Row) and Race (Lower Row) for Different Choices of Proxies in the Warfarin Data Set, Along with the Resulting Entropy




either has a higher approval rate than the Black group or has about roughly the same. However, the magnitude of a positive White-versus-Black disparity, and the direction of White-versus-API disparity, still remains very ambiguous.

Overall, the large size of all partial identification sets reflects the tremendous ambiguity in assessing lending disparities based on proxy variables such as geolocation and income. Thus it is nearly *impossible* to draw reliable conclusions about demographic disparity only according to the observed data. This conclusion is roughly in line with previous analyses of BISG (Chen et al. 2019) but provides a precise meaning to these limits.

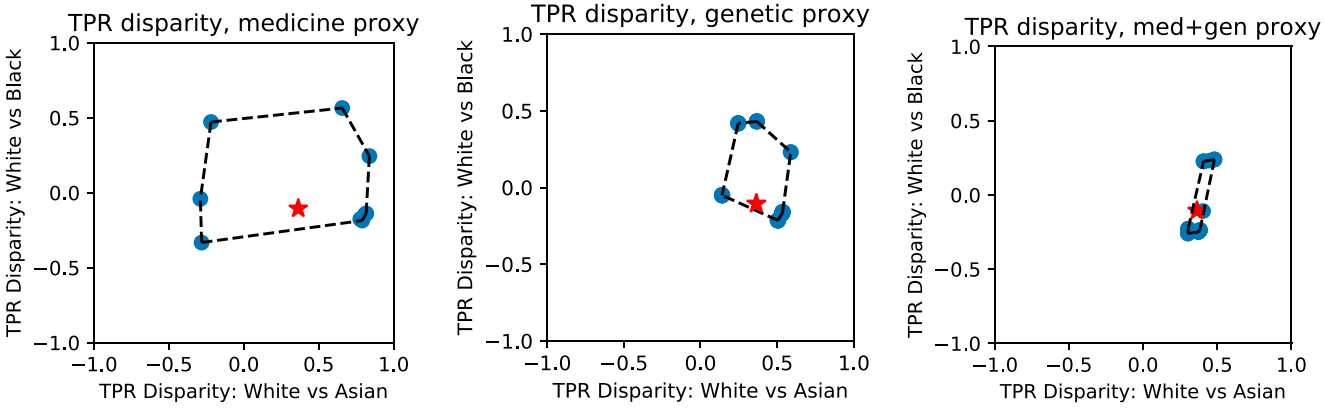
## 8.2. Personalized Warfarin Dosing

**8.2.1. Background.** Warfarin is the most commonly used oral anticoagulant agent worldwide (International Warfarin Pharmacogenetics Consortium 2009). Finding the appropriate warfarin dosage is very challenging and important, because it can vary drastically among patients, and an incorrect dose can possibly lead to serious adverse outcomes. This challenge attracts considerable interest in designing personalized warfarin dosage algorithms, including linear regression (International Warfarin Pharmacogenetics Consortium 2009), LASSO (Bastani and Bayati 2020), and decision trees (Kallus 2017). However, it was shown that the personalized dosing algorithms may show disparate performance for different ethnic groups (see, e.g., appendix 9 in International Warfarin Pharmacogenetics Consortium (2009)).

**8.2.2. Data Set, Proxy Variables, and Nuisance Estimation.** We use the PharmGKB data set[16] of 5,700 patients treated with warfarin. The data for each patient include demographics (sex, ethnicity, age, weight, height, and smoking status), reason for treatment (e.g., atrial fibrillation), current medications, comorbidities (e.g., diabetes), and genetic factors (presence of genotype variants of CYP2C9 and VKORC1). All of these variables are categorical, and we treat missing value of each variable as a separate value. Moreover, this data set contains the true patient-specific optimal warfarin doses determined by physicians' adjustment over a few weeks. We focus on the subsample of 4,891 White, Black, and Asian patients whose optimal warfarin doses are not missing. We dichotomize the optimal doses into high dosage (more than 35 mg/week, denoted as $Y = 1$) and low dosage (less than 35 mg/week, denoted as $Y = 0$). To develop a personalized dosage algorithm, we follow International Warfarin Pharmacogenetics Consortium (2009) and fit a linear regression to predict the optimal dosage based on all other variables, and we recommend high dosage if the predicted optimal dosage is more than 35 mg/week ($\hat{Y} = 1$) and recommend low dosage ($\hat{Y} = 0$) otherwise.

We randomly split the data set into two halves, with one half as the primary data set and the other as the auxiliary data set, so that the independence of two data sets assumed in Section 7 is satisfied. Our goal is to evaluate the partial identification sets for true-positive rate disparities of this personalized

**Figure 7.** (Color online) The Outer Approximation of Partial Identification Set for TPRD in Warfarin Dosing as Determined by Different Proxies




*Note.* The true disparity is shown as a star.

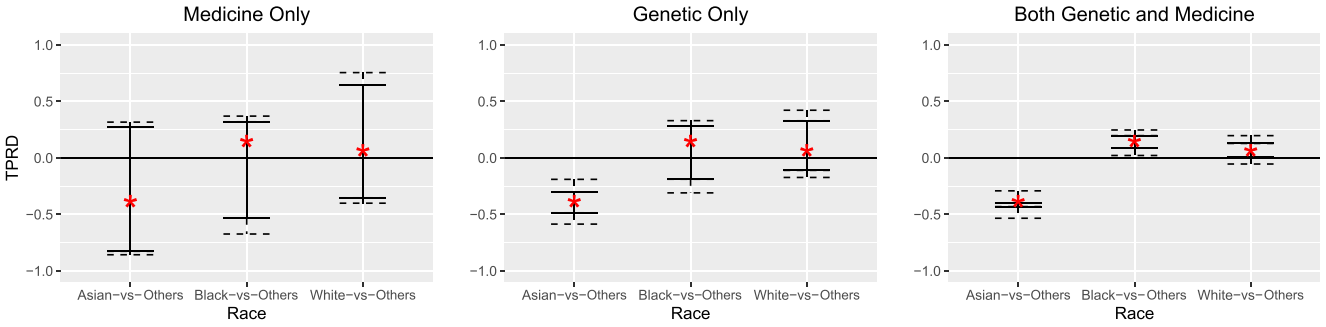
dosage algorithm. Positive disparities indicate that the personalized algorithm has a higher chance to correctly recommend a high dosage to one group than to another group.

We consider three sets of discrete proxy variables: only genetic factors, only current medications, and both genetic factors and current medications. Among the proxy variables, the genetic factors are particularly strong candidates because they are found to be highly predictive for the optimal warfarin dosage (International Warfarin Pharmacogenetics Consortium 2009). At the same time, genotype variants of CYP2C9 and VKORC1 are known to also be highly correlated with race. For example, International Warfarin Pharmacogenetics Consortium (2009) even recommended imputing missing values of the genotypes based on race labels.

The conditional probabilities of race, optimal dosage indicator $Y$, and recommended dosage indicator $\hat{Y}$, given these proxy variables, can be easily estimated by corresponding sample averages within each level of the proxy variables. In Figure 6 we display the histograms of the estimated conditional probabilities for both race and outcomes, for each proxy. For outcomes, we show probabilities of all four combinations of true outcome and decision outcome. For race, we separate the probabilities by label. We note that current medications and genetic factors together form a highly informative proxy, both for race and for outcomes.

**8.2.3. Binary Comparisons.** Figure 8 shows the estimates of closed-form bounds of TPRD for one race versus the other without any extra assumptions. The bound estimators for TPRD and associated confidence intervals are similar to those for DD in Section 7.1 (see Online Appendix A.2 for details). We first observe that using genetic factor proxies, whether in combination with current medication or not, provides clear evidence that the TPR disparity between Asian and other races is negative in disfavor of Asians. Although the directions of the Black-versus-Others TPRD and White-versus-TPRD are unclear when

**Figure 8.** (Color online) Partial Identification Bounds of Demographic Disparity (Proposition 8) for Different Proxy Variables in the Warfarin Dosing




*Notes.* Solid bars represent the estimates of the bounds, and dashed bars indicate 95% confidence intervals. The true value based on self-reported race is shown as an asterisk.

using either genetic factor proxies or current medication proxies alone, combining these two set of proxies considerably narrows the bounds of these two disparities, and the Black-versus-Others TPRD is positive at a 95% confidence level.

**8.2.4. Multiple-Level Protected Class.** Figure 7 shows the estimated partial identification sets of TPRD for the White group versus another group. The sets are computed by the support function approach described in Section 7.2. We observe that using genetic factor proxies provides clear evidence that the TPRD between White and Asian is positive in favor of White. Further adding medication proxies provides a very clear sense of the significant magnitude of the TPRD between White and Asian, not just its direction. However, in all cases, both the direction and magnitude of the disparity between White and Black is unclear.

Overall, our observations are consistent with the different quality of the proxies: although the genetic proxy is stronger than the medicine proxy, combining the proxies adds additional information that tightens the partially identified set. Studying the partially identified plots allows a practitioner to assess the value of additional information and, in some cases, the direction of disparities.

## 9. Conclusion

Assessing the fairness of algorithmic decisions is a fundamentally difficult task: it is now well understood that even when algorithms do not take sensitive information as an input they can still be biased in various worrisome ways, but what counts as "unfair" can be very context dependent. But any such adjudication and scrutiny must start from understanding how different groups are disparately impacted by such decisions. For example, disparate impact has been codified in U.S. law and regulation as evidentiary basis for closer review and even sanction. We here studied a further complication: membership in protected groups is usually not even recorded in the data, requiring the use of auxiliary data where such labels are present. This limitation hinders both fair lending and healthcare reforms, and it is important to address it.

We formulated this problem from the perspective of data combination and studied the fundamental limits of identification. This provided a new perspective on the commonplace usage of proxy models and a way to assess what can and cannot be learned from the data. The tools we developed allow one to compute exactly the tightest possible bounds on disparity that could possibly be learned from the data. We believe this is an invaluable tool given that disparate impact assessments can have far-reaching policy implications.

Beyond the specific tools we presented here, we also hope our work will inspire other researchers to consider fundamental statistical ambiguities in the measurement of fairness, beyond just the ambiguities between the different definitions. Given the sensitivity of such matters, truly understanding the limits of what cannot actually be measured—and what, on the other hand, can be said with certainty—is critical for any reliable assessment of the fairness of any decision-making algorithm.

## Acknowledgments

The authors thank the editorial team including three anonymous reviewers for the constructive comments on an earlier draft.

## Endnotes

[1] In the United States, the Fair Housing Act and Equal Credit Opportunity Act codify as protected attributes: age, race/ethnicity, disability, exercised rights under the Consumer Credit Protection Act of 1968, familial status (household composition), gender identity, marital status (single or married), national origin, race, recipient of public assistance, religion, and/or sex.

[2] The U.S. Home Mortgage Disclosure Act, or HMDA, authorizes lenders to collect such information for mortgage applicants and coapplicants.

[3] For clarity, we emphasize to the reader the difference between an algorithm's "bias" with respect to protected groups (e.g., as quantified by disparate impact) and the *statistical bias* of assessments of such disparities. In this paper, "bias" only ever refers to the latter statistical bias and "disparities" to systematic differences in algorithmic outputs.

[4] Assumption 1 can be relaxed by assuming instead that the distribution $\mathbb{P}_a$ of the auxiliary observations $(A, Z)$ satisfies $\mathbb{P}_a(A = a \mid Z) = \mathbb{P}(A = a \mid Z)$, with an *arbitrary* distribution $\mathbb{P}_a(Z)$ of proxy variables. This relaxation does not change any of our results in Sections 4–6, but it does change our estimators in Section 7, where we would need to account for this distributional shift in $Z$ across the data sets. We omit this straightforward extension for brevity.

[5] Strictly speaking, demographic disparity is not based on classification "error," but it can be also computed from the within-class confusion matrices.

[6] For example, as fairness criteria, both demographic and classification parities have been criticized for their *inframarginality*; that is, they average over individual risk far from the decision boundary (Corbett-Davies and Goel 2018). However, inframarginality may be unavoidable when outcomes are binary. There may be no true individual "risk," only the stratified frequencies of binary outcomes (default or recidivation) over strata defined by predictive features, which are, in turn, chosen by the decision maker.

[7] Proxies can still be continuous, which we will leverage when we impose extra smoothness assumptions in Section 6.1.

[8] If no extra assumption is imposed, then $\mathcal{P}_A$ is the set of all joint distributions, so that $\mathcal{P}_D \cap \mathcal{P}_A = \mathcal{P}_D$.

[9] Note that $\mu'_{\tilde{y}y}$ differs with $\mu_{\tilde{y}y}$ in Equation (8) only in the two separate arguments $\tilde{w}$ and $\tilde{w}'$ to explicitly characterize the monotonicity in two different directions, a property crucial for deriving the closed-form sets.

[10] Note that $\Delta_{\mathrm{DD}}(\mathcal{P})$ is equal to its closed convex hull if $\mathcal{W}(\mathcal{P})$ is closed convex, because $\mu(a, w)$ is affine in $w$. Both $\mathcal{W}(\mathcal{P}_D)$ and $\mathcal{W}(\mathcal{P}_D) \cap \mathcal{W}_{\mathrm{Lip}}$ are closed convex. On the other hand, $\Delta_{\mathrm{TPRD}}(\mathcal{P}), \Delta_{\mathrm{TNRD}}(\mathcal{P})$ are

generally not convex in the nonbinary setting, and taking their convex hull provides the smallest convex outer approximation to them.

[11] If, instead, the auxiliary data set is of smaller size, then we can focus on estimators that converge at rate of $O(\sqrt{n_{\text{aux}}})$. For example, in Equations (28) and (29) of Theorem 1, we can use a scaling factor of $\sqrt{n_{\text{aux}}}$ instead of $\sqrt{n_{\text{pri}}}$ to get a similar asymptotic normality result. These two different scaling factors are asymptotically equivalent when $n_{\text{aux}} \asymp n_{\text{pri}}$ but may differ when $r = 0$ or $r = 1$. For brevity, we only allow the first. The latter can be handled symmetrically.

[12] Unlike the case in Section 7.1, statistical inference (confidence intervals) for general multivariate sets characterized by estimated support functions is an active research area (Molinari 2019) and generally computationally burdensome, so we leave it for further research and focus on the consistency of our support function estimates.

[13] The data set can be downloaded from https://www.consumerfinance.gov/data-research/hmda/explore. This data set includes mortgage loan application records in the United States, which include self-reported race/ethnicity, loan origination outcome, geolocation (state, county, and census tract), annual income, and loan amount, among other variables.

[14] For example, the White-versus-Rest disparity is the demographic disparity of $a$ as White and $b$ as either API or Black.

[15] Restricting the conditional joint distribution to be any smoother can, in fact, be refuted from the data via infeasibility.

[16] The data set can be downloaded from https://www.pharmgkb.org/downloads.

## References

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. *Propublica* (May 23), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Audibert J-Y, Tsybakov AB (2007) Fast learning rates for plug-in classifiers. *Ann. Statist.* 35(2):608–633.

Baines AP, Courchane MJ (2014) Fair lending: Implications for the indirect auto finance market. Accessed May 31, 2019. https://www.crai.com/sites/default/files/publications/Fair-Lending-Implications-for-the-Indirect-Auto-Finance-Market.pdf.

Barocas S, Hardt M, Narayanan A (2018) *Fairness and Machine Learning: Limitations and Opportunities* (fairmlbook.org). http://www.fairmlbook.org.

Bastani H, Bayati M (2020) Online decision making with high-dimensional covariates. *Oper. Res.* 68(1):276–294.

Beresteanu A, Molchanov I, Molinari F (2011) Sharp identification regions in models with convex moment predictions. *Econometrica* 79(6):1785–1821.

Bogen M, Rieke A, Ahmed S (2020) Awareness in practice: Tensions in access to sensitive attribute data for antidiscrimination. *Proc. 2020 Conf. Fairness Accountability Transparency* (ACM, New York), 492–500.

Bonvini M, Kennedy EH (2019) Sensitivity analysis via the proportion of unmeasured confounding. Preprint, submitted December 5, https://arxiv.org/abs/1912.02793.

Brown DP, Knapp C, Baker K, Kaufmann M (2016) Using bayesian imputation to assess racial and ethnic disparities in pediatric performance measures. *Health Service Res.* 51(3):1095–1108.

Cambanis S, Simons G, Stout W (1976) Inequalities for E $k(X, Y)$ when the marginals are fixed. *Z. Wahrscheinlichkeitstheor. Verwandte Geb.* 36(4):285–294.

Charnes A, Cooper WW (1962) Programming with linear fractional functionals. *Naval Res. Logist. Quart.* 9(3–4):181–186.

Chen J, Kallus N, Mao X, Svacha G, Udell M (2019) Fairness under unawareness: Assessing disparity when protected class is unobserved. *Proc. Conf. Fairness Accountability Transparency* (ACM, New York), 339–348.

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/debiased machine learning for treatment and structural parameters. *Econometrics J.* 21(1): C1–C68.

Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5(2):153–163.

Ciliberto F, Tamer E (2009) Market structure and multiple equilibria in airline markets. *Econometrica* 77(6):1791–1828.

Comptroller of the Currency (2010) Fair lending: Comptroller's handbook. Accessed May 31, 2019. https://www.occ.treas.gov/publications/publications-by-type/comptrollers-handbook/fair-lending/pub-ch-fair-lending.pdf.

Consumer Financial Protection Bureau (2013) CFPB and DOJ order ally to pay $80 million to consumers harmed by discriminatory auto loan pricing. https://www.consumerfinance.gov/about-us/newsroom/cfpb-and-doj-order-ally-to-pay-80-million-to-consumers-harmed-by-discriminatory-auto-loan-pricing/.

Consumer Financial Protection Bureau (2014) Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment. Accessed May 31, 2019. https://www.consumerfinance.gov/data-research/research-reports/using-publicly-available-information-to-proxy-for-unidentified-race-and-ethnicity/.

Corbett-Davies S, Goel S (2018) The measure and mismeasure of fairness: A critical review of fair machine learning. Preprint, submitted July 31, https://arxiv.org/abs/1808.00023.

D'Amour A (2019) On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. Proc. 22nd Internat. Conf. Artificial Intelligence Statist. (PMLR), 3478–3486.

Dastin J (2018) Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (October 10), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Datta A, Tschantz MC, Datta A (2015) Automated experiments on ad privacy settings. *Proc. Privacy Enhancing Tech.* 2015(1):92–112.

Dembosky JW, Haviland AM, Haas A, Hambarsoomian K, Weech-Maldonado R, Wilson-Frederick SM, Gaillot S, Elliott MN (2019) Indirect estimation of race/ethnicity for survey respondents who do not report race/ethnicity. *Medical Care* 57(5):e28–e33.

Elliott MN, Fremont A, Morrison PA, Pantoja P, Lurie N (2008) A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Services Res.* 43(5, Part 1):1722–1736.

Elliott MN, Morrison PA, Fremont A, McCaffrey DF, Pantoja P, Lurie N (2009) Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services Outcomes Res. Methodol.* 9(2):69–83.

Fan Y, Sherman R, Shum M (2014) Identifying treatment effects under data combination. *Econometrica* 82(2):811–822.

Freedman DA (1999) Ecological inference and the ecological fallacy. *International Encyclopedia of the Social & Behavioral Sciences*, 6(7): 4027–4030.

Fremont AM, Bierman A, Wickstrom SL, Bird CE, Shah M, Escarce JJ, Horstman T, Rector T (2005) Use of geocoding in managed care settings to identify quality disparities. *Health Affairs* 24(2):516–526.

Gaffney A, McCormick D (2017) The Affordable Care Act: Implications for health-care equity. *Lancet* 389(10077):1442–1452.

Goldfarb A, Tucker C (2011) Online display advertising: Targeting and obtrusiveness. *Marketing Sci.* 30(3):389–404.

Goodman SN, Goel S, Cullen MR (2018) Machine learning, health disparities, and causal reasoning. *Ann. Internal Medicine* 169(12): 883–884.

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. Lee DD, Sugiyama M, Luxburg UV, Guyon I,

Garnett R, eds. *Advances in Neural Information Processing Systems*, Vol. 29 (Curran Associates, Red Hook, NY), 3315–3323.

Hirano K, Porter JR (2012) Impossibility results for nondifferentiable functionals. *Econometrica* 80(4):1769–1790.

Holstein K, Wortman Vaughan J, Daumé H III, Dudik M, Wallach H (2019) Improving fairness in machine learning systems: What do industry practitioners need? *Proc. 2019 CHI Conf. Human Factors Comput. Systems* (ACM, New York), 1–16.

Imai K, Khanna K (2016) Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Anal.* 24(2):263–272.

International Warfarin Pharmacogenetics Consortium (2009) Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England J. Medicine* 360(8):753–764.

Iyer G, Soberman D, Villas-Boas JM (2005) The targeting of advertising. *Marketing Sci.* 24(3):461–476.

Jiang W, King G, Schmaltz A, Tanner MA (2018) Ecological regression with partial identification. Preprint, submitted April 18, https://arxiv.org/abs/1804.05803.

Kallus N (2017) Recursive partitioning for personalization using observational data. *Proc. 34th Internat. Conf. Machine Learn.*, Vol. 70 (PMLR), 1789–1798.

Kallus N, Zhou A (2018a) Confounding-robust policy improvement. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *Advances in Neural Information Processing Systems*, Vol. 31 (Curran Associates, Red Hook, NY), 9269–9279.

Kallus N, Zhou A (2018b) Residual unfairness in fair machine learning from prejudiced data. *Proc. 35th Internat. Conf. Machine Learn.* (PMLR), 2444–2453.

Kallus N, Zhou A (2019a) Assessing disparate impact of personalized interventions: Identifiability and bounds. Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems*, Vol. 32 (Curran Associates, Red Hook, NY), 3426–3437.

Kallus N, Zhou A (2019b) The fairness of risk scores beyond classification: Bipartite ranking and the XAUC metric. Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems*, Vol. 32 (Curran Associates, Red Hook, NY), 3438–3448.

Kallus N, Mao X, Zhou A (2019) Interval estimation of individual-level causal effects under unobserved confounding. *Proc. 22nd Internat. Conf. Artificial Intelligence Statist.* (PMLR), 2281–2290.

Kennedy EH, BalakrishnanS G'Sell M (2018) Sharp instruments for classifying compliers and generalizing causal effects. Preprint, submitted January 11, https://arxiv.org/abs/1801.03635.

Kleinberg J, Mullainathan S, Raghavan M (2017) Inherent trade-offs in the fair determination of risk scores. Papadimitriou CH, ed. 8th Innov. Theoret. Comput. Sci. Conf. (Dagstuhl Publishing, Saarbrücken/Wadern, Germany), 43:1–43:23.

Laber EB, Lizotte DJ, Qian M, Pelham WE, Murphy SA (2014) Dynamic treatment regimes: Technical challenges and applications. *Electronic J. Statist.* 8(1):1225–1272.

Lambrecht A, Tucker C (2019) Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Sci.* 65(7):2966–2981.

Lewbel A (2018) The identification zoo: Meanings of identification in econometrics. *J. Econom. Lit.* 57(4):835–903.

Manski CF (2003) *Partial Identification of Probability Distributions* (Springer Science & Business Media, New York).

Manski CF (2005) Partial identification with missing data: Concepts and findings. *Internat. J. Approximate Reasoning* 39(2–3):151–165.

Miller CC (2015) Can an algorithm hire better than a human? *New York Times* (June 25), https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html.

Molinari F (2019) Microeconometrics with partial identification. Heckman JJ, Leamer EE, eds. *Handbook of Econometrics* (North Holland). Vol 7A:1–145.

Monahan J, Skeem JL (2016) Risk assessment in criminal sentencing. *Annual Rev. Clin. Psych.* 12:489–513.

Narayanan A (2018) Translation tutorial: 21 fairness definitions and their politics. Tutorial presentation at the Fairness, Accountability and Transparency Conference 2018, February 23.

Obermeyer Z, Mullainathan S (2019) Dissecting racial bias in an algorithm that guides health decisions for 70 million people. *Proc. Conf. Fairness Accountability Transparency* (ACM, New York), 89.

Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH (2018) Ensuring fairness in machine learning to advance health equity. *Ann. Internal Medicine* 169(12):866–872.

Ridder G, Moffitt R (2007) The econometrics of data combination. Heckman JJ, Leamer EE, eds. *Handbook of Econometrics*, Vol. 6B (North-Holland, Amsterdam), 5469–5547.

Robinson SM (1975) Stability theory for systems of inequalities. Part I: Linear systems. *SIAM J. Numer. Anal.* 12(5):754–769.

Rockafellar RT (2015) *Convex Analysis* (Princeton University Press, Princeton, NJ).

Rüschendorf L (2013) *Mathematical Risk Analysis: Dependence, Risk Bounds, Optimal Allocations and Portfolios*, Springer Series in Operations Research and Financial Engineering (Springer, Heidelberg, Germany).

Rutherglen G (1987) Disparate impact under Title VII: An objective theory of discrimination. *Virginia Law Rev.* 73(7):1297–1345.

Schuessler AA (1999) Ecological inference. *Proc. Natl. Acad. Sci. USA* 96(19):10578–10581.

Sweeney L (2013) Discrimination in online ad delivery. Preprint, submitted January 29, https://arxiv.org/abs/1301.6822.

Ulmer C, McFadden B, Nerenz DR, eds. (2009) *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement* (National Academies Press, Washington, DC).

U.S. Census Bureau (2010) 2010 Census Summary File 1. Data set, Accessed May 31, 2019. https://www.census.gov/data/datasets/2010/dec/summary-file-1.html.

U.S. Census Bureau (2018) Current population survey (CPS). Accessed May 31, 2019. https://www.census.gov/topics/income-poverty/income/data/tables/cps.html.

Verma S, Rubin J (2018) Fairness definitions explained. *2018 IEEE/ACM Internat. Workshop Software Fairness (FairWare)* (IEEE, Piscataway, NJ), 1–7.

Villani C (2008) *Optimal Transport: Old and New*, Vol. 338 (Springer-Verlag, Berlin).

Wakefield J (2004) Ecological inference for 2 × 2 tables (with discussion). *J. Roy. Statist. Soc. Ser. A* 167(3):385–445.

Weissman JS, Hasnain-Wynia R (2011) Advancing healthcare equity through improved data collection. *New England J. Medicine* 364(24):2276–2277.

Zhang Y (2016) Assessing fair lending risks using race/ethnicity proxies. *Management Sci.* 64(1):178–197.

Zimmer MJ (1996) The emerging uniform structure of disparate treatment discrimination litigation. *Georgia Law Rev.* 30:563–626.