Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality

Aria Khademi
Pennsylvania State University
khademi@psu.edu

Sanghack Lee
Purdue University
lee2995@purdue.edu

David Foley
Pennsylvania State University
djf47@psu.edu

Vasant Honavar
Pennsylvania State University
vhonavar@psu.edu

ABSTRACT

As virtually all aspects of our lives are increasingly impacted by algorithmic decision making systems, it is incumbent upon us as a society to ensure such systems do not become instruments of unfair discrimination on the basis of gender, race, ethnicity, religion, etc. We consider the problem of determining whether the decisions made by such systems are discriminatory, through the lens of causal models. We introduce two definitions of group fairness grounded in causality: *fair on average causal effect* (FACE), and *fair on average causal effect on the treated* (FACT). We use the Rubin-Neyman *potential outcomes* framework for the analysis of cause-effect relationships to robustly estimate FACE and FACT. We demonstrate the effectiveness of our proposed approach on synthetic data. Our analyses of two real-world data sets, the Adult income data set from the UCI repository (with gender as the protected attribute), and the NYC Stop and Frisk data set (with race as the protected attribute), show that the evidence of discrimination obtained by FACE and FACT, or lack thereof, is often in agreement with the findings from other studies. We further show that FACT, being somewhat more nuanced compared to FACE, can yield findings of discrimination that differ from those obtained using FACE.

ACM Reference Format:
Aria Khademi, Sanghack Lee, David Foley, Vasant Honavar. 2019. Fairness in Algorithmic Decision Making:, An Excursion Through the Lens of Causality. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, San Francisco, USA, 7 pages. https://doi.org/10.1145/3308558.3313559

1 INTRODUCTION

With the growing adoption of algorithmic decision making systems, e.g., AI and machine learning systems, across many real-world decision making scenarios on the Web and elsewhere, there is a pressing need to make sure that such systems do not become vehicles of unfair discrimination, inequality, and social injustice [2, 3]. Of particular interest in this context is the task of detecting and preventing discrimination or unfair treatment of individuals or groups on the basis of gender, race, religion, etc. Such discrimination is traditionally addressed using one of two legal frameworks: *disparate treatment* (which aims to enforce procedural fairness, namely, the equality of treatment that prohibits the use of the protected attribute in the decision process); and *disparate impact* [3] (which aims to guarantee outcome fairness, namely, the equality of outcomes between protected groups relative to other groups). It is clear that enforcing procedural fairness within the disparate treatment framework does not guarantee non-discrimination within the disparate impact framework.

There is growing interest in algorithmic decision making systems that are demonstrably *fair* (see [4] for a review). Much of this literature relies on precise definitions that quantify fairness to avoid discrimination with respect to *protected* attributes, e.g., race, gender, on the basis of the legal notions of disparate treatment or disparate impact [3] (see [3, 4, 32, 40, 60] for reviews). Some examples include: fairness through unawareness [14], individual fairness [11], equalized odds [15, 54], calibration [8], demographic (or statistical) parity [6, 21–23], the 80% rule (disparate impact) [13, 55], representational fairness [33, 56], and fairness under composition [12].

Unfortunately, choosing the appropriate definition of fairness in a given context is extremely challenging due to a number of reasons. First, depending on the relationship between a protected attribute and data, enforcing certain definitions of fairness can actually increase discrimination [28]. Second, different definitions of fairness can be impossible to satisfy simultaneously [4, 8, 26]. Many of these difficulties can be attributed to the fact that fairness criteria are based *solely* on the joint probability distribution of the random variables of interest, namely, $\hat{Y}$ (predicted outcome), Y (actual outcome), $\tilde{X}$ (features), and A (sensitive attributes). [15] recently showed any such definition for fairness of a predictor that depends merely on the joint probability distribution is not necessarily capable of detecting discrimination. Hence, it is tempting to approach the problem of fairness through the lens of causality [1].

Answering questions of fairness through the lens of causality entails replacing the question "Is the decision discriminatory with respect to a protected attribute?" by: "Does the protected attribute have a causal effect on the decision?" A practical difficulty in using this approach is that, in general, establishing a causal relationship between a protected attribute and a decision requires the results of experimental manipulation of the protected attribute. Fortunately, however, existing frameworks for determining causal effects from observational data [20, 36] provide a rich set of theoretical results as well as practical tools for elucidating causal effects, and specifically, answering questions about *counterfactuals* or *potential outcomes*, i.e., results of *hypothetical* experimental interventions from observational data, whenever it is possible to do so. Hence, there is a growing body of work (see [32] for a recent review) focused on explicitly causal (as opposed to purely joint distribution based or *observational*) definitions for fairness (e.g., [5, 7, 25, 28–30, 35, 50, 53, 57–59]). While some, e.g., [59], have focused on testing fairness (or conversely, determining whether there is discrimination), others, e.g., [25, 28] have sought to design machine learning

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313559

algorithms that yield predictive models that are demonstrably fair. However, most of the existing work on defining fairness in causal terms has focused on variants of individual fairness. Against this background, we focus on robust methods for detecting and quantifying discrimination against protected groups, which is a necessary prerequisite for developing predictive models that are provably non-discriminatory.

Contributions. We reduce the problem of quantifying discrimination against protected groups to the well-studied problem of estimating the causal effect of some variable(s) on a target (outcome) variable. We introduce two explicitly causal definition of fairness in a population, *fair on average causal effect* (FACE), and in a protected group, *fair on average causal effect on the treated* (FACT), both with respect to a protected attribute (e.g., gender, race). We use the Rubin-Neyman *potential outcomes* framework [20, 45, 49] for robust estimation of FACE and FACT. We demonstrate the effectiveness of the proposed approach in detecting and quantifying group fairness using synthetic data, as well as two real-world data sets: the Adult income data from the UCI repository [10] (with gender being the protected attribute), and the NYC Stop and Frisk data (with race being the protected attribute). We show that the evidence of discrimination, or lack thereof, obtained by FACE and FACT is often in agreement with other studies. We further show that FACT, being somewhat more nuanced compared to FACE, can yield findings of discrimination that differ from those obtained using FACE.

2 FAIRNESS: A CAUSAL PERSPECTIVE

Assume we have observational data on a population of individuals. Let $\tilde{X} \in \mathcal{X}$ be the vector of non-protected attributes, $A \in \mathcal{A} = \{a, a'\}$ be a binary protected attribute, and $Y \in \mathcal{Y}$ an outcome of interest. The question we want to answer is: Are individuals being discriminated against, on average, with respect to outcomes or decisions Y on the basis of a protected attribute A? From a causal perspective, such a question is *equivalent* to the following question: Does A have a causal effect on Y? In other words, how much would Y change, on average, were the value of A to change? Both Structural Causal Models [36] and the Rubin-Neyman Causal Model (RCM) [20] (also called the *potential outcomes* model) offer methods for estimating such causal effects from observational data.

We introduce two explicitly causal definitions for fairness "on average" in a population or a protected group (as opposed to causal definitions of individual fairness, e.g., counterfactual fairness [28]) with respect to a protected attribute (e.g., gender, race). Let $Y_i^{(a)}$ and $Y_i^{(a')}$ be the potential outcomes of a data point i had their value of A been a and a', respectively. Let $h : \mathcal{X} \times \mathcal{A} \rightarrow \mathcal{Y}$ be a decision function (or a predictive model trained using machine learning) that is used to support decision making. $\mathbb{E}[\cdot]$ is the expectation of a random variable. We define the following.

Definition 2.1. (FACE: Fair on Average Causal Effect). A decision function h is said to be fair, on average over all individuals in the population, with respect to A, if $\mathbb{E}[Y_i^{(a)} - Y_i^{(a')}] = 0$.

Definition 2.2. (FACT: Fair on Average Causal Effect on the Treated). A decision function h is said to be fair with respect to A, on average over individuals with the same value of A, if $\mathbb{E}[Y_i^{(a)} - Y_i^{(a')} \,|\, A_i = a] = 0$.

Example. Imagine we are given the hiring data of a company containing demographic information about applicants, as well as A = {male, female} as their gender, and Y = {hired, rejected} as whether they were hired by the company. Our task is to determine whether the company's hiring decisions are fair on average with respect to gender. FACE contrasts the expected outcomes (i.e., hiring) between men vs. women with the expectation taken over the entire population. FACT contrasts the expected outcomes observed for a specific protected group (e.g., women) and the hypothetical (counterfactually inferred) outcomes for the group had they not been members of the protected group (with the expectation taken only over the members of the protected group), e.g., hiring outcomes for women contrasted with outcomes for the same individuals had their gender been different with all other attributes remaining unchanged. Obviously, such counterfactual outcomes cannot be obtained from observational data[1] and ought to be estimated.

3 ESTIMATING FACE AND FACT

We use tools offered by the potential outcomes framework [20] to estimate FACE and FACT. These tools rely on the following key assumptions: i) *Consistency* which requires that for a data point i, the potential outcome of i under any level of treatment a, i.e., Y_i^a, equals the *actual* outcome observed for that data point, Y_i^{obs}, had they been exposed to treatment a. Formally, under consistency, $Y_i^{obs} = a\, Y_i^{(a)} + a'\, Y_i^{(a')}$ would hold for all i. This assumption, used in existing literature [7, 34, 35, 38], is a rather natural one to make in our setting. ii) *Positivity* which asserts that the probability $Pr(A = a \,|\, X = x) > 0$ for all values of A. In our setting, this means each value of the protected attribute has a non-zero probability. iii) *Stable Unit Treatment Value Assumption (SUTVA)* [47] which consists of two sub-assumptions: 1) Absence of *interference* between individuals [9], which means that an individual's potential outcome is unaffected by the treatment assigned to any other individual. While this assumption is plausible in our setting, it may be violated in some settings, in which case, such violations should be accounted for [16]. 2) Presence of only one form of treatment (and control). For example, if a treatment involves administering a drug, then all individuals who take the drug, take it in the same form (e.g., injection). This assumption is trivially satisfied in our setting because treatment is simulated by the protected attribute. iv) *Unconfoundedness of the treatment mechanism* which implies that given a set of observables, the potential outcomes of each individual are jointly independent of the corresponding treatment [46]. Unconfoundedness cannot be verified or contradicted entirely on the basis of observational data. However, sensitivity analysis [31, 42] can be a useful tool for analyzing the estimated causal effects under violations of the unconfoundedness assumption. Strong ignorability refers to the combination of unconfoundedness and positivity [43]. Strong ignorability is a sufficient condition for the causal effect to be identifiable [16] and is equivalent to the *back-door criterion* [37], which is required for identifiability of the causal effects in Pearl's

[1] This is called the Fundamental Problem of Causal Inference (FPCI) from observational data [18].

model of causality [37]. In our work, as in the case of existing work on causal definitions of fairness [35], we assume strong ignorability.

3.1 Estimating and Interpreting FACE

We use Inverse Probability Weighting (IPW), also known as Inverse Probability of Treatment Weighting (IPTW) in Marginal Structural Models (MSM) [39] to estimate FACE. Specifically, for each individual i, we calculate a *stabilized weight*: $sw_i = \frac{Pr(A_i=a)}{Pr(A_i=a \mid \tilde{X}_i=\tilde{x}_i)}$ (call it the weight model). We obtained stabilized weights using the R package *ipw* (version 1.0-11) [52]. Assigning such a weight to every data point, we generate a "pseudo-population" in which there are sw_i copies of each data point i. Subsequently, the associative parameter β in the weighted regression (call it the outcome model) of the (continuous) outcome Y on the protected attribute A: $\mathbb{E}[Y^{(A)}] = \delta + \beta A + \tilde{\theta}^\top \tilde{X}$, would be the causal effect of A on Y. For a binary output Y, we use the weighted logistic regression model: $\text{logit}\,(\mathbb{E}[Y^{(A)}]) = \delta + \beta A + \tilde{\theta}^\top \tilde{X}$. In the absence of unmeasured confounders, if *either* the weight model *or* the outcome model are correctly specified, then $\hat{\beta}$ is an unbiased estimator of the average causal effect [39]. For example, suppose Y is salary and A is gender. At the chosen level of statistical significance α, $\hat{\beta} = 0$ implies that salary is fair with respect to gender on average over the entire population of individuals; $\hat{\beta} \neq 0$ implies that, on average, women's salary differs from that of men by a factor of $\hat{\beta}$ (across the entire population). For a continuous outcome Y, $\hat{\beta}$ is simply the average causal effect of A on Y. For a binary outcome Y, $\hat{\beta}$ corresponds to the causal odds ratio of salary for women versus men.

3.2 Estimating and Interpreting FACT

We use *matching* to estimate FACT. Consider the example of salary discrimination based on gender. For a woman, we can never observe what the salary would have been, *had she been a man* (i.e., her counterfactual salary). Hence, we estimate the counterfactual salary as follows [20]: 1) Using a suitable matching technique (see Section Matching Methods), we match the woman i, to a man j who is closest to i with respect to a distance measure $d(i, j)$. 2) The matching process is repeated as needed until matches are of acceptable quality (see Section Quality of Matches). 3) After matching, we use the salary of the matched man j (i.e., Y_j), as the counterfactual salary of the woman i.

Matching Methods. The results of matching depend on the choice of distance measure $d(\cdot, \cdot)$ as well as the matching process. Several matching methods exist (see [51] for a survey). In what follows, for simplicity and brevity, we refer to individuals with protected attribute set to $A = a$ as the *treated* individuals and those with the protected attribute set to $A = a'$ as the *controlled* individuals. We used the matching methods implemented within the R package *MatchIt* (version 3.0.2) [17] with all parameters set to their default values unless otherwise noted: (i) Exact Matching (EM); (ii) Nearest Neighbor Matching (NNM) with propensity score [43]. Following [48], we estimated the propensity scores using the logit link and transformed them to the linear scale. Then, we ran NNM with replacement, based on the linear propensity scores, and discarded the data points (both from treated and controlled) that fall outside the support of the distance measure; (iii) Nearest Neighbor Matching with a Propensity Caliper (NNMPC). NNMPC includes only matches within a certain number of standard deviations of the distance measure and discards the rest. In NNMPC, we use the same procedure as in NNM, augmented with a caliper = 0.25 [44], resulting in the matches outside 0.25 times the standard deviation of the (transformed) linear propensity score, being discarded; (iv) Mahalanobis Metric Matching within the Propensity Caliper [48] (MMMPC). MMMPC determines for each data point, a "donor pool" of available matches within the propensity caliper. Mahalanobis metric matching is then performed among the data points chosen in the previous step mimicking blocking in randomized experiments [48]. We ran MMMPC with caliper, replacement, and discarding strategy as described above in NNM; and (v) Full Matching (FM) [41]. We used the same distance measure and discarding strategy as described above in NNM.

Quality of Matches. To ensure accurate estimation of FACT, it is crucial to measure the "goodness-of-match." If the data points are well matched, then one can proceed to estimate FACT. Common diagnostics for examining the quality of match include both numerical and graphical criteria. Among the numerical criteria, following [48], we compare the standardized difference in the means of the treated and the controlled data points in terms of the distance measure. We denote the absolute value of this difference in means on the original, and matched data, by $\overline{D}_{a,a'}$, and $\overline{D}^m_{a,a'}$, respectively. For the match to be of good quality, $\overline{D}^m_{a,a'}$ has to be close to 0. Among the graphical criteria, we use quantile-quantile (QQ), and jitter plots recommended by [17, 51].[2]

Outcome Analysis After Matching. With good quality matched pairs identified, we can proceed to conduct outcome analysis for FACT estimation. Matching methods often assign appropriate weights to the matched data points to balance the treated and controlled data distributions. After obtaining the weights via matching, we run the following weighted regression models: $\mathbb{E}[Y^{(A)}] = \delta + \gamma A + \tilde{\theta}^\top \tilde{X}$, for continuous, and $\text{logit}\,(\mathbb{E}[Y^{(A)}]) = \delta + \gamma A + \tilde{\theta}^\top \tilde{X}$, for binary outcomes, both *on the matched data set*, to estimate FACT. The estimated coefficient for A in the equations above, i.e., $\hat{\gamma}$, estimates FACT. The resulting estimate is "doubly robust" in that if *either* the matching model, *or* the outcome model, are correctly specified, $\hat{\gamma}$ would be statistically consistent [17].

Interpreting $\hat{\gamma}$ as a Measure of FACT. Suppose Y is salary and A is gender. At the chosen level of statistical significance α, $\hat{\gamma} = 0$ implies that there is no significant difference in expected salary for women compared to what their salary would have been had they been men (with all non-protected attributes $\tilde{X}$ remaining unchanged, a condition that is approximated by counterfactual inference using matching), thus implying no gender-based discrimination in salary for women; $\hat{\gamma} \neq 0$ implies that, on average, women's salary is statistically significantly different from what it would have been, had they been men, thus implying gender-based discrimination in salary. For a continuous outcome Y, e.g., the salary in US dollars, if statistically significant, $\hat{\gamma} \neq 0$ means that on average, *considering men and women that are matched based on their feature*

[2] We avoid the commonly used hypothesis tests for *assessing feature balance* in diagnosing the quality of matches because such tests have been shown to be misleading in general [19].

vector $\tilde{X}$, the difference between women's salary and that of men is $\hat{\gamma}$. For a binary outcome, e.g., salaries binarized with an arbitrary threshold τ, $\hat{\gamma}$ is the causal odds ratio of women's salary compared to that of men, for those women and men who are similar.

Impact of Unmeasured Confounders on $\hat{\gamma}$. What if the strong ignorability assumption (i.e., no hidden confounders) is violated? In the absence of unmeasured confounding, matching estimators are unbiased if the matching model is specified correctly, i.e., if balance is achieved over the *observed attributes*. However, it is conceivable that the results of matching could change in the presence of *unobserved confounders* (i.e., hidden bias). We perform sensitivity analysis [31, 42] to investigate the degree to which the unmeasured confounders impact $\hat{\gamma}$. Let Γ be the odds ratio of matched (using any matching method) data points i and j receiving a treatment. Sensitivity analysis proceeds by first assuming $\Gamma = 1$ (i.e., no hidden bias). Then, it increases Γ (e.g., $1, \ldots, 5$), thus mimicking the presence of hidden bias, and examines the resulting changes to statistical significance of $\hat{\gamma}$. The Γ at which the significance of the upper bound for the p-value would change (e.g., from < 0.05 to > 0.05) is the point at which $\hat{\gamma}$ is no longer robust to hidden bias. We ran sensitivity analysis using the R package *rbounds* (version 2.1) [24].

4 EXPERIMENTS AND RESULTS

We tested our approach on a synthetic data set (where the discrimination based on a protected attribute can be varied in a controlled fashion), and two real-world data sets that have been previously used in studies of fairness. In each case, we designated a protected attribute and estimated FACE and FACT as measures of discrimination based on that attribute. We run all of our statistical significance tests with $\alpha = 0.05$. We proceed to describe the data sets, experiments, as well as our FACE and FACT analyses in detail.

4.1 Data sets

Synthetic data set. We generated 1000 data points, each with a feature vector $\tilde{X} = (X_1, \ldots, X_5)$, a protected attribute $A \in \{0, 1\}$, and an outcome variable Y according to the following: $X_1, \ldots, X_5 \overset{iid}{\sim} \mathcal{N}(0, 1)$; $A \mid X \sim \text{Bernoulli}\left(\text{logit}^{-1}(\sum_{i=1}^{5} X_i)\right)$; $Y \mid X, A = \sum_{i=1}^{5} X_i W_i$, where $\tilde{W} = (W_1, \ldots, W_5)$ is a weight vector (fixed for all data points) with each element drawn randomly in [0,1]. The resulting generative model ensures there are no hidden confounders and there is no discrimination, as measured by FACE and FACT, with respect to the outcome variable Y on the basis of the protected attribute A.

The Adult data set. The Adult income data set [27][3], contains information about individuals as well as their salaries. The data set includes 48842 individuals each with 14 attributes, 6 continuous and 8 categorical, including demographic and work-related information such as age, gender, hours of work per week, etc. We examined whether there is gender-based discrimination in salaries by designating *gender* as the sensitive attribute. We encoded categorical variables using one-hot-encoding and removed data records with missing values, yielding a data set with 46033 individuals and 45 features (excluding gender, the protected attribute). We designated the outcome Y to be a binary variable denoting whether the person's annual salary is $> \$50K$ (Y=1), or $\leq \$50K$ (Y= 0).

The NYC Stop and Frisk (NYCSF) data set. We retrieved the publicly available stop, search, and frisk data from The New York Police Department (NYPD)[4] website which serves demographic and other information about drivers stopped by the NYC police force. Our question is whether the arrests made after stops have been discriminatory with respect to race. Following [28], we restricted our experiment to the year 2014 yielding a total of 45787 records. We selected the subset of records corresponding to only Black-Hispanic and White men. We designated *race* as the protected attribute with $A = 1$ denoting Black-Hispanic and $A = 0$ denoting White. We dropped the data records with missing values and encoded categorical variables with one-hot-encoding. The resulting data consist of 7593 records each with 73 features (excluding race, the sensitive attribute). The outcome Y denotes whether an arrest was made ($Y = 1$), or not ($Y = 0$).

4.2 FACE Check: Fairness Analysis Using FACE

We report our analysis of fairness using FACE, for the synthetic, Adult, and NYCSF data sets. The estimated FACE ($\hat{\beta}$) are shown in Table 1. In all cases, the null hypothesis is $H_0 : \beta = 0$. In the case of synthetic data, we find insufficient evidence to reject H_0, suggesting the outcome is fair with respect to the protected attribute (an expected conclusion given the design of the generative model in Section 4.1, which ensures that the outcome is fair with respect to the protected attribute). In the case of Adult data, we reject H_0 and find that $\hat{\beta}$, the average causal effect of gender on salaries, is -1.069. This means that on average, over the entire population, the odds of women having a salary $> \$50K$ a year is $\exp(-1.069) \approx 0.34$ times that of men, suggesting gender-based discrimination against women as measured by FACE. This finding is in agreement with the conclusions reported in [30, 35]. In the case of NYCSF data, we reject H_0 and find that $\hat{\beta}$ is 0.273 which means that on average, the odds of Black-Hispanics being arrested after a stop by the police, is $\exp(0.273) \approx 1.31$ times that of Whites, suggesting possible racial bias against non-Whites.

Table 1: Estimates of FACE ($\hat{\beta}$) obtained on the synthetic, Adult, and NYCSF data sets.

Data set	$\hat{\beta}$	Standard Error	P-value
Synthetic	-1.130×10^{-16}	6.991×10^{-17}	**0.106**
Adult	**−1.069**	3.614×10^{-1}	0.003
NYCSF	**0.273**	1.259×10^{-1}	0.030

4.3 FACT Check: Fairness Analysis Using FACT

We report our analysis of fairness using FACT, for the synthetic, Adult, and NYCSF data sets.

[3] https://archive.ics.uci.edu/ml/datasets/adult

[4] https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page

Matching Quality Analyses. Because the quality of matched pairs used to estimate FACT impacts the conclusions that can be drawn using it, we compare the FACT estimates obtained using several widely-used matching methods described in Section 3.2. We present some analyses to verify that the generated matches are of sufficiently high quality for estimating FACT.

We observe that before matching, $\overline{D}_{a,a'}$ is 1.6400, 3.3508, and 1.1616, on the synthetic, Adult, and NYCSF data sets, respectively. The matching methods dramatically reduced $\overline{D}_{a,a'}$ on all of the data sets (see Table 2). Overall, NNM and FM achieved the lowest $\overline{D}^m_{a,a'}$ as compared to other matching methods on all data sets. The greater the number of pairs that are matched, the harder it is to achieve balance, and the trade-off between the two can be application dependent. We observed that considering the trade-off between the number of matches and $\overline{D}^m_{a,a'}$, FM yields higher quality matches on all data sets as compared to other methods.

The QQ plots are generated for each feature in each data set. In Figure 1 we show the QQ plots before and after FM for the first three features of the synthetic data set. The features lie far away from the 45 degree line before FM. After FM, the features are much better aligned to the diagonal line showing a more desirable feature balance. We also show the jitter plots of FM on all data sets in Figure 2. It is clear that the distribution of propensity scores of the treated and controlled data points are very similar to each other after matching. Having verified that the results of matching are of adequate quality, we proceed to use them for estimating FACT.

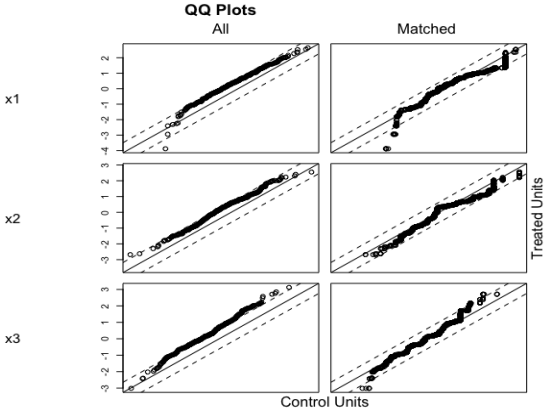


Figure 1: QQ plots of the first three features from the synthetic data set before (left) and after (right) FM.

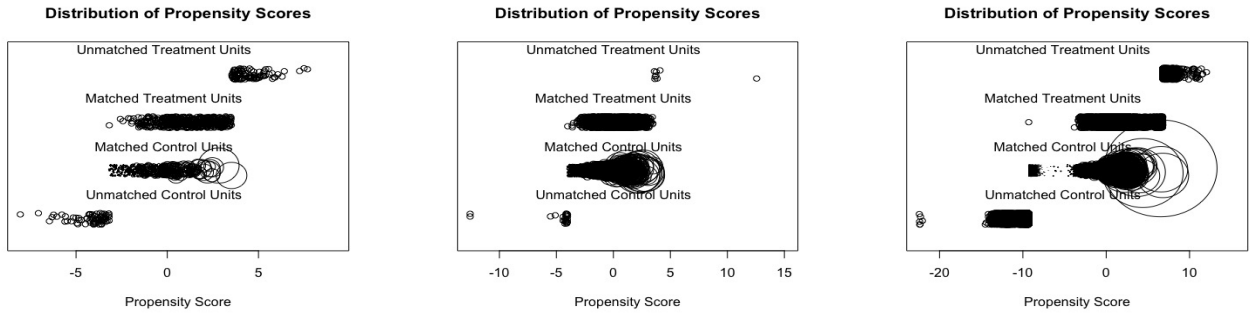
FACT Estimates. The results of FACT analyses on the synthetic, Adult, and NYCSF data sets are summarized in Table 2 (Note that EM did not yield any matches and hence is omitted from Table 2). In all cases, the null hypothesis is $H_0 : \gamma = 0$. In the case of synthetic data, FACT analyses show that for NNM and MMMPC, there is not enough evidence to reject H_0. The p-values in the case of NNMPC and FM are < 0.05, but the magnitude of the estimated $\hat{\gamma}$ is close to zero. We conclude that the synthetic data set is fair on average with respect to FACT. On the Adult data, we can reject H_0, suggesting that salaries of women are significantly lower than those of men who match them on the non-protected attributes. For example, using FM, we find that $\hat{\gamma} = -0.573$, thus the odds of women earning $> \$50K$ a year, is $\exp(-0.573) \approx 0.56$ times that of men. We conclude that in the Adult data, there is evidence of gender-based discrimination in salary, on average, against women. On the NYCSF data, interestingly, FACT analyses show that H_0 cannot be rejected, suggesting a lack of evidence for racial bias, on average, in arrests after stops (when Black-Hispanics are compared with Whites who match them on non-protected attributes). This conclusion contradicts the finding of racial bias based on counterfactual fairness analysis (Supplementary Material S6 in [28]) which suggests discrimination against *individuals*, as well as FACE analysis (see Section 4.2). We conjecture that the apparent discrepancy can be explained by noting that (i) fairness (or discrimination) on average does not necessarily imply individual-level fairness (or individual-level discrimination), and (ii) FACT compares the observed outcomes of members of a protected group with the hypothetical (counterfactual) outcomes they would have experienced had they not been members of the protected group (with all non-protected attributes remaining unchanged), whereas FACE compares such counterfactual outcomes on the entire population.

Impact of Unmeasured Confounders. We ran sensitivity analysis of our estimates of FACT for $\Gamma = 1, \ldots, 10$ (where larger values of Γ correspond to greater bias introduced by hidden confounders) on the Adult and NYCSF data sets. We find that all of our estimates obtained with various matching methods are quite robust to hidden confounder bias. Specifically, on the Adult data set, for all matching methods except FM, the estimates are robust to such bias, and for FM, they are robust up to $\Gamma = 4.5$, which corresponds to a fairly large amount of bias. On the NYCSF data set, estimates obtained via NNM and MMMPC are robust to hidden confounder bias, and NNMPC and FM are robust up to Γ equals 8.5, and 3, respectively. These results mean that our FACT estimates (and hence our findings of discrimination on the basis of protected attributes, or lack thereof) are fairly robust to hidden confounder bias.

5 SUMMARY AND DISCUSSION

We have approached the problem of detecting whether a group of individuals that share a sensitive attribute, e.g., race, gender, have been subjected to discrimination in an algorithmic decision-making system, through the lens of causality. We have introduced two explicitly causal definitions of group fairness: *fair on average causal effect* (FACE), and *fair on average causal effect on the treated* (FACT). We have shown how to robustly estimate FACE and FACT, and use the resulting estimates to detect and quantify discrimination based on specific attributes (e.g., gender, race). The results of our experiments on synthetic data show that our proposed methods are effective at detecting and quantifying group fairness. Our analyses of the Adult data set for evidence of gender-based discrimination in salary, and of the NYCSF data set for evidence of racial bias in arrests after traffic stops, yield evidence of discrimination, or lack thereof, that is often in agreement with other studies.[5] We show

[5] The regression and matching-based methods we employed to estimate FACE and FACT adjust for covariates that might be potential confounders of the protected attribute, which although necessary in general, may be unnecessary in the case of gender and race, because they are unlikely to be caused by any other covariate. Consequently, the



(a) Synthetic data set. **(b) NYCSF data set.** **(c) Adult data set.**

Figure 2: Jitter plots of distribution of the propensity scores on the linear logit scale after FM on the synthetic (left), NYCSF (middle), and Adult (right) data sets. Each circle represents a data point. Area of the circle is proportional to the weight given to the data point. Female, Black-Hispanic = treated, and male, White = controlled.

Table 2: Estimates of FACT ($\hat{\gamma}$) obtained via various matching methods on the synthetic, NYCSF, and Adult data sets.

Synthetic data set						
Matching Method	# of Treated Matches	# of Control Matches	$\overline{D}^m_{a,a'}$	$\hat{\gamma}$	Standard Error	P-value
NNM	454	155	0.0032	-6.972×10^{-17}	4.947×10^{-17}	0.159
NNMPC	454	168	0.0234	-9.196×10^{-17}	3.020×10^{-17}	0.002
MMMPC	454	174	0.0308	7.424×10^{-18}	3.192×10^{-17}	0.816
FM	454	386	0.0031	-8.263×10^{-17}	3.702×10^{-17}	0.026
Adult data set						
Matching Method	# of Treated Matches	# of Control Matches	$\overline{D}^m_{a,a'}$	$\hat{\gamma}$	Standard Error	P-value
NNM	13330	4922	0.0009	−0.637	0.128	6.560×10^{-7}
NNMPC	13301	5258	0.0714	−0.650	0.113	1.050×10^{-8}
MMMPC	13301	5838	0.0584	−0.586	0.131	7.650×10^{-6}
FM	13330	15320	0.0009	−0.573	0.115	5.700×10^{-7}
NYCSF data set						
Matching Method	# of Treated Matches	# of Control Matches	$\overline{D}^m_{a,a'}$	$\hat{\gamma}$	Standard Error	P-value
NNM	2605	1305	0.0001	0.049	0.186	0.788
NNMPC	2605	1414	0.0266	0.246	0.160	0.124
MMMPC	2605	1264	0.0231	0.324	0.183	0.078
FM	2605	4958	0.0000	0.155	0.171	0.364

on the real-world data that our estimates of FACE and FACT are robust to unmeasured confounding. Our results further show on the real-world data that FACE and FACT based findings do not always agree. Our FACT analyses also demonstrate that group-fairness (or discrimination) does not necessarily imply individual-level fairness (or individual-level discrimination).

Some directions for further research include: relaxing the assumption that the data are independent and identically distributed (i.i.d.) in settings where individuals are related to each other through family ties or other relationships; examining the relationships between different causal notions of fairness; and designing automated decision support systems that are demonstrably non-discriminatory with respect to given outcome(s) and protected attribute(s).

Acknowledgements. This work was funded in part by grants from the NIH NCATS through the grant UL1 TR000127 and TR002014 and by the NSF through the grants 1518732, 1640834, and 1636795, the Edward Frymoyer Endowed Professorship in Information Sciences and Technology at Pennsylvania State University and the Sudha Murty Distinguished Visiting Chair in Neurocomputing and Data Science funded by the Pratiksha Trust at the Indian Institute of Science (both held by Vasant Honavar). The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors.

reported estimates of FACE and FACT are likely to represent *direct* causal effects as opposed to *total* causal effects.

REFERENCES

[1] C. Barabas, M. Virza, K. Dinakar, J. Ito, and J. Zittrain. 2018. Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. In *Conference on Fairness, Accountability and Transparency*. 62–76.

[2] S. Barocas, E. Bradley, V. Honavar, and F. Provost. 2017. Big Data, Data Science, and Civil Rights. *arXiv preprint arXiv:1706.03102* (2017).

[3] S. Barocas and A. D. Selbst. 2016. Big data's disparate impact. *Cal. L. Rev.* 104 (2016), 671.

[4] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. 2017. Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207* (2017).

[5] F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti. 2017. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics* 3, 1 (2017), 1–21.

[6] T. Calders, F. Kamiran, and M. Pechenizkiy. 2009. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*. IEEE, 13–18.

[7] S. Chiappa and T. PS. Gillam. 2018. Path-specific counterfactual fairness. *arXiv preprint arXiv:1802.08139* (2018).

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

[9] D. R. Cox. 1958. Planning of experiments. (1958).

[10] D. Dheeru and E. Karra Taniskidou. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 214–226.

[12] C. Dwork and C. Ilvento. 2018. Fairness Under Composition. *arXiv preprint arXiv:1806.06122* (2018).

[13] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.

[14] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, Vol. 1. 2.

[15] M. Hardt, E. Price, and N. Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*. 3315–3323.

[16] M. A. Hernan and J. M. Robins. 2018. *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming.

[17] D. E. Ho, K. Imai, G. King, and E. A. Stuart. 2011. MatchIt: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software* 42, 8 (2011), 1–28.

[18] P. W. Holland. 1986. Statistics and Causal Inference. *J. Amer. Statist. Assoc.* 81, 396 (1986), 945–960.

[19] K. Imai, G. King, and E. Stuart. 2008. Misunderstandings Among Experimentalists and Observationalists about Causal Inference. *Journal of the Royal Statistical Society, Series A* 171, part 2 (2008), 481–502.

[20] G. W. Imbens and D. B. Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[21] J. E. Johndrow and K. Lum. 2017. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv preprint arXiv:1703.04957* (2017).

[22] F. Kamiran and T. Calders. 2009. Classifying without discriminating. In *Computer, Control and Communication. IC4 2009. 2nd International Conference on*. IEEE, 1–6.

[23] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[24] L. Keele. 2010. An overview of rbounds: An R package for Rosenbaum bounds sensitivity analysis with matched data. *White Paper. Columbus, OH* (2010), 1–15.

[25] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

[26] J. Kleinberg, S. Mullainathan, and M. Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[27] R. Kohavi. 1996. Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Vol. 96. 202–207.

[28] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4069–4079.

[29] M. J. Kusner, C. Russell, J. R. Loftus, and R. Silva. 2018. Causal Interventions for Fairness. *arXiv preprint arXiv:1806.02380* (2018).

[30] J. Li, J. Liu, L. Liu, T. D. Le, S. Ma, and Y. Han. 2017. Discrimination detection by causal effect estimation. In *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE, 1087–1094.

[31] W. Liu, S. J. Kuramoto, and E. A. Stuart. 2013. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science* 14, 6 (2013), 570–580.

[32] J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva. 2018. Causal Reasoning for Algorithmic Fairness. *arXiv preprint arXiv:1805.05859* (2018).

[33] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830* (2015).

[34] D. Madras, E. Creager, T. Pitassi, and R. Zemel. 2019. Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 349–358.

[35] R. Nabi and I. Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 2018. NIH Public Access, 1931.

[36] J. Pearl. 2009. *Causality*. Cambridge university press.

[37] J. Pearl. 2010. The foundations of causal inference. *Sociological Methodology* 40, 1 (2010), 75–149.

[38] J. Pearl. 2019. On the Interpretation of do(x). *Journal of Causal Inference, forthcoming* (2019).

[39] J. M. Robins, M. A. Hernán, and B. Brumback. 2000. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* 11, 5 (2000), 550–560.

[40] A. Romei and S. Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 5 (2014), 582–638.

[41] P. R. Rosenbaum. 1991. A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society. Series B (Methodological)* (1991), 597–610.

[42] P. R. Rosenbaum. 2005. Sensitivity analysis in observational studies. *Encyclopedia of Statistics in Behavioral Science* 4 (2005), 1809–1814.

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

[44] P. R. Rosenbaum and D. B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39, 1 (1985), 33–38.

[45] D. B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 5 (1974), 688.

[46] D. B. Rubin. 1978. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics* (1978), 34–58.

[47] D. B. Rubin. 1980. Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* 75, 371 (1980), 591–593.

[48] D. B. Rubin. 2001. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology* 2, 3-4 (2001), 169–188.

[49] D. B. Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

[50] C. Russell, M. J. Kusner, J. Loftus, and R. Silva. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*. 6417–6426.

[51] E. A. Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science: a review journal of the Institute of Mathematical Statistics* 25, 1 (2010), 1.

[52] W. M. van der Wal, R. B. Geskus, et al. 2011. Ipw: an R package for inverse probability weighting. *J Stat Softw* 43, 13 (2011), 1–23.

[53] T. J. VanderWeele and W. R. Robinson. 2014. On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology (Cambridge, Mass.)* 25, 4 (2014), 473.

[54] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1171–1180.

[55] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Artificial Intelligence and Statistics*. 962–970.

[56] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

[57] J. Zhang and E. Bareinboim. 2018. Fairness in Decision-Making–The Causal Explanation Formula. In *32nd AAAI Conference on Artificial Intelligence*.

[58] L. Zhang, Y. Wu, and X. Wu. 2016. Situation testing-based discrimination discovery: a causal inference approach. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 2718–2724.

[59] L. Zhang, Y. Wu, and X. Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.

[60] I. Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148* (2015).