Counterfactual Risk Assessments, Evaluation, and Fairness

Amanda Coston
Heinz College & Machine Learning Department
Carnegie Mellon University

Alan Mishler & Edward H. Kennedy
Department of Statistics
Carnegie Mellon University

Alexandra Chouldechova
Heinz College
Carnegie Mellon University

ABSTRACT

Algorithmic risk assessments are increasingly used to help humans make decisions in high-stakes settings, such as medicine, criminal justice and education. In each of these cases, the purpose of the risk assessment tool is to inform actions, such as medical treatments or release conditions, often with the aim of reducing the likelihood of an adverse event such as hospital readmission or recidivism. Problematically, most tools are trained and evaluated on historical data in which the outcomes observed depend on the historical decision-making policy. These tools thus reflect risk under the historical policy, rather than under the different decision options that the tool is intended to inform. Even when tools are constructed to predict risk under a specific decision, they are often improperly evaluated as predictors of the target outcome.

Focusing on the evaluation task, in this paper we define counterfactual analogues of common predictive performance and algorithmic fairness metrics that we argue are better suited for the decision-making context. We introduce a new method for estimating the proposed metrics using doubly robust estimation. We provide theoretical results that show that only under strong conditions can fairness according to the standard metric and the counterfactual metric simultaneously hold. Consequently, fairness-promoting methods that target parity in a standard fairness metric may—and as we show empirically, do—induce greater imbalance in the counterfactual analogue. We provide empirical comparisons on both synthetic data and a real world child welfare dataset to demonstrate how the proposed method improves upon standard practice.

ACM Reference Format:
Amanda Coston, Alan Mishler & Edward H. Kennedy, and Alexandra Chouldechova. 2020. Counterfactual Risk Assessments, Evaluation, and Fairness. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3351095.3372851

1 INTRODUCTION

Much of the activity in using machine learning to help address societal problems focuses on algorithmic decision-making and algorithmic decision support systems. In settings such as health, education, child welfare and criminal justice, decision support systems commonly take the form of risk assessment instruments (RAIs), which distill rich case information into risk scores that reflect the likelihood of the case resulting in one or more adverse outcomes. [6, 8, 16, 22, 29, 46, 47]. Prior literature has raised significant concerns regarding the fairness, transparency, and effectiveness of existing RAIs [3, 4, 9, 10, 13]. Yet RAIs remain very popular in practice, and there is a large body of research on fairness and transparency promoting methods that seek to address some of these concerns [e.g 17, 20, 21, 37, 54, 55].

This paper highlights a different issue, one that has not received sufficient attention in the discussion of RAIs but that nonetheless has significant implications for fairness: RAIs are typically trained and evaluated as though the task were prediction when in reality the associated decision-making tasks are often interventions. Models trained and evaluated in this way answer the question: What is the likelihood of an adverse outcome *under the observed historical decision process*? Yet the question relevant to the decision maker is: What is the likelihood of an adverse outcome *under the proposed decision*? When decisions do not impact outcomes—when we are in what [27] call a "pure predition" setting—these are one and the same. However, many decisions take the form of interventions specifically designed to mitigate risk. RAIs for these settings must be developed and evaluated taking into account the effect of historical decisions on the observed outcomes. Failure to do so will result in RAIs that, despite appearing to perform well according to standard evaluation practices, underperform on cases such as those that have been historically receptive to intervention.

In this paper we propose an approach to **counterfactual risk modeling and evaluation** to properly account for these intervention effects. Counterfactual modeling has been proposed for medical RAIs [1, 44, 45], and prior work has used counterfactual evaluation for off-policy learning in bandit settings [14]. However, the question of adapting counterfactual evaluation for risk assessments and in particular for predictive bias assessments remains open. In this paper, we propose a new evaluation method for RAIs that uses doubly-robust estimation techniques from causal inference [40, 51]. We also argue that fairness metrics that are functions of the outcome should be defined counterfactually, and we use our evaluation method to estimate these metrics. We theoretically and empirically characterize the relationship between the standard fairness metrics and their counterfactual analogues. Our results suggest that in many cases, achieving parity in the standard metric will not achieve parity in the counterfactual metric.

Our contributions are as follows: 1) We define counterfactual versions of standard predictive performance metrics and propose doubly-robust estimators of these metrics (§ 3); 2) We provide empirical support that this evaluation outperforms existing methods using a synthetic dataset and a real-world child welfare hotline screening dataset (§ 3); 3) We propose counterfactual formulations

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02…$15.00
https://doi.org/10.1145/3351095.3372851

of standard fairness metrics that are more appropriate for decision-making settings (§ 4); 4) We provide theoretical results that only under strong conditions, which are unlikely to hold in general, does fairness according to standard metrics imply fairness according to counterfactual metrics (§ 4); 5) We demonstrate empirically that applying existing fairness-corrective methods can increase disparity in the counterfactual redefinition of the metric they target (§ 4).

2 BACKGROUND AND RELATED WORK

2.1 Counterfactual learning and evaluation

Literature on contextual bandits has considered counterfactual learning and evaluation of decision *policies*. While this literature is methodologically relevant, as we discuss below, it addresses a different problem. In the *decision support* setting we are considering, human users will ultimately decide what action to take. The goal of the learning and evaluation task is not to learn a decision policy, but rather to learn a risk model that will inform human decisions. That is, the risk assessment task is to accurately and fairly estimate the probability of an outcome under a given intervention.

While the underlying task is different, the statistical methods used in evaluation are related. [49] use propensity score weighting, a form of importance sampling, to correct for the effect of the historical treatment on the observed outcome, and they propose learning the optimal policy based on the minimization of the propensity-score weighted empirical risk. Propensity-score methods are a good candidate when one has a good model of the historical decision-making policy, but may otherwise be biased. Doubly robust (DR) methods, by contrast, are robust to parametric misspecification of the propensity score model if instead one has the correct specification of the model of the regression outcome $\mathbb{E}[Y|X]$ where Y is the outcome and X are the features/covariates [39, 41, 51]. In a non-parametric setting, DR methods have faster rates of convergence than propensity-score methods [23]. DR methods have been used for policy learning in the offline bandit setting [14]. The policy learned minimizes a DR estimate of the loss.

Prior work has considered counterfactual RAIs when the outcomes are continuous-time trajectories [44]. In this work, the trained model is evaluated on real data using the observed outcomes, and on simulated data. Evaluating against the observed outcomes can be misleading in settings in which treatment was not assigned randomly (§ 3.3.3). We propose instead to adapt DR techniques to evaluate counterfactual RAIs.

Counterfactual learning in the causal inference literature uses model selection based on DR estimation of counterfactual loss [51]. Whereas this approach evaluates counterfactual metrics implicitly, our approach does so explicitly, providing the estimators for standard classification metrics in § 3.3.3.

A line of work focuses on counterfactual learning in the presence of hidden confounders. [18] propose policy learning via minimax regret learning over uncertainty sets. Their method is not immediately applicable to decision-support settings where RAIs are more informative to decision-makers than a policy recommendation. [32] propose using deep latent variable models to model hidden confounders via proxies in the data. § 3 of our paper assumes no hidden confounders, and future work could attempt to incorporate these techniques for handling hidden confounders. Our theoretical analysis in § 4 holds even in the presence of hidden confounding.

2.2 Fairness and causality

A growing literature on counterfactual fairness has offered notions of fairness based on the counterfactual of the protected attribute (or its proxy) [26, 31, 53]. In this work, a policy is considered fair if it would have made the same decision had the individual had a different value of the protected attribute (and hence, potentially different values of features affected by the attribute). In this setting, the treatment decision is the outcome, and the protected attribute is the 'treatment'. By contrast, we consider counterfactual treatment decisions and consider a future observation to be the outcome.[1]

Another line of work considers unfair causal pathways between the protected attribute (or its proxy) and the outcome variable or target of prediction [35, 57]. These papers characterize discrimination via path-specific effects, which are defined by interventions on the protected attribute. We do not consider interventions on (i.e. counterfactuals of) the protected attribute; rather, we propose methods that account for interventions on treatment decisions.

Fairness definitions based on the counterfactual of the protected attribute are not widely used in RAI settings for two reasons. First, the assumptions required to estimate these counterfactual metrics prohibit the use of important features, such as prior history, or require full specification of the structural causal model (SCM) [30, 31, 56] These requirements are too restrictive for our settings where we have insufficient domain knowledge to construct the SCM and where we are unable to disregard important predictors like prior history. Second, and more significantly, these definitions are ill-suited for RAI settings like child welfare screening. As we discuss in § 4, decisions made based on the counterfactual protected attribute may cause further harm to the protected groups.

Our work bears conceptual similarity to the analysis of residual unfairness when there is selection bias in the training data that induces covariate shift at test time as discussed in [19]. In settings where cases are systematically screened out from the training set, such as loan approvals, they find that applying fairness-corrective methods is insufficient to achieve parity. We consider a different but related setting in which we observe outcomes for all cases, but these outcomes are under different treatments. We propose fairness definitions that account for the effect of these treatments on the observed outcomes, and analyze the conditions under which existing methods can achieve this notion of counterfactual fairness.

3 COUNTERFACTUAL MODELING AND EVALUATION

Before introducing the learning approaches and evaluation methods considered in this work, we pause to clarify the types of risk-based decision policies to which our evaluation strategy as presented is tailored. RAIs typically inform human decisions either by identifying cases that are the most (or least) *risky*, or by identifying cases that are the most (or least) *responsive*. The evaluation metrics we consider are relevant in the paradigm where human decision-makers wish to intervene on the *riskiest* cases. Our method can be adapted (as discussed in § 3.3) for paradigms based on responsiveness.

[1] This distinction is also made in a survey of fairness literature [34].

The motivating application for our work is child welfare screening. Child welfare service agencies across the nation field over 4 million child abuse and neglect calls each year [50]. Call workers must decide whether to "screen in" a call, which refers to opening an investigation into the family. The child welfare system is responsible for responding to all cases where there is significant suspicion that the child is in present or impending danger. The standard of practice is therefore to identify the *riskiest* cases. Jurisdictions in California, Colorado, Oregon and Pennsylvania are all in various stages of developing and integrating RAIs into their call screening processes. The RAIs are trained on historical data to predict adverse child welfare outcomes, such as re-referral to the hotline or out-of-home foster care placement [8]. The decision to investigate a call can affect the likelihood of the target outcomes.

3.1 Notation

We use $Y \in \{0, 1\}$ to denote the observed binary outcome, and for exposition we assume $Y = 1$ is the unfavorable outcome. $T \in \{0, 1\}$ denotes the decision which for simplicity we take to be binary. We note that DR estimation methods can be used in any treatment setting, including for continuous treatments such as dosing [24, 52]. Throughout the remainder of the paper we will use the term 'decision' and 'treatment' interchangeably. In describing counterfactual learning and evaluation, we rely on the potential outcomes framework common in causal inference [25, 36, 42]. In this framework, Y^t denotes the outcome under treatment t. For any given case we only get to observe Y^0 or Y^1. We will take $T = 0$ to be the baseline treatment, the decision under which it is relevant to assess risk. Most risk assessment settings have a natural baseline, which is often the decision to not intervene. For instance, in education one might assess the likelihood of poor outcomes if a student is not offered support; in child welfare it is natural to assess risk if the call is not investigated. We refer to the baseline treatment as *control* and the non-baseline treatment as *treatment*. $X \in \mathcal{X} \subseteq \mathbb{R}^d$ denotes the covariates (or features) which may include a protected or sensitive attribute $A \in \{0, 1\}$. $\pi(X) = \mathbb{P}(T = 1 \mid X)$ denotes the propensity score, whose estimate we denote by $\hat{\pi}(X)$. In the child welfare setting, X contains call details and historical information on all associated parties, T is whether the case is screened-in for investigation, and Y is whether the case is re-referred to the hotline in a six-month period. We use subscripts i to index our data; e.g., X_i are the features for case i. We use $\hat{Y} : \mathcal{X} \mapsto \{0, 1\}$ to denote our predicted label and $\hat{s} : \mathcal{X} \mapsto [0, 1]$ to denote the predicted score which is the model's estimate of the target outcome (our RAI).[2]

3.2 Learning models of risk

In this section we introduce "observational" (standard practice) and "counterfactual" forms of model training.

3.2.1 Observational. The *observational* RAI produces risk estimates by regressing Y on X for the entire observed dataset. i.e., this RAI estimates $\mathbb{E}[Y \mid X]$. This model answers the question: What is the likelihood of an adverse outcome under the *observed historical decision process*? The observational RAI is ill-suited for guiding future decisions; it will, for instance, underestimate (baseline) risk for cases that were historically responsive to treatment.

3.2.2 Counterfactual. The counterfactual model of risk estimates the outcome under the baseline treatment. Our counterfactual model of risk targets $\mathbb{E}[Y^0 \mid X]$. Even though we only observe Y^0 or Y^1 for any given observation, we may nevertheless draw valid inference about both potential outcomes under a set of standard identifying assumptions[3]. These assumptions hold by design in our synthetic dataset, and we discuss why they may be reasonable in the child welfare setting under each point.

(1) Consistency: $Y = TY^1 + (1 - T)Y^0$.
This assumes there is no interference between treated and control units. This is a reasonable assumption in the child welfare setting since opening an investigation into one case will not likely affect another case's observed outcome.[4]
(2) Exchangeability: $Y^0 \perp T \mid X$. This assumes that we measured all variables that jointly influence the intervention decision T and the potential outcome Y^0. This is an untestable assumption; it may be reasonable in the child welfare setting where the measured variables capture most of the information the call screeners use to make their decision (see § 3.4.2).
(3) Weak positivity requirement: $\mathbb{P}(\pi(X) < 1) = 1$ requires that each example have some non-zero chance of the baseline treatment. This can hold by construction in decision support settings. We can filter out cases that violate this assumption since the decision for these cases is nearly certain.[5]

Our assumptions identify the target: $\mathbb{E}[Y^0|X] = \mathbb{E}[Y|X, T = 0]$.

The counterfactual model estimates $\mathbb{E}[Y^0 \mid X]$ by computing an estimate of $\mathbb{E}[Y \mid X, T = 0]$. We can train such a model by applying any probabilistic classifier to the control population. Since the control population may have a different covariate distribution than the full population, reweighing can be used to correct this covariate shift [38]. This may be useful in a setting with limited data or where model misspecification is a concern [48].

3.3 Evaluation

To evaluate how well our models of risk might inform decision-making in the paradigm targeting the riskiest cases, we assess precision, true positive rate (TPR), false positive rate (FPR), and calibration.[6] Since the task is to evaluate how well the model predicts risk under a baseline intervention, we specify the performance metrics in terms of Y^0. The target counterfactual TPR is

$$\mathbb{E}[\hat{Y} \mid Y^0 = 1] \tag{1}$$

The target counterfactual precision is

$$\mathbb{E}[Y^0 \mid \hat{Y} = 1] \tag{2}$$

[2] $\hat{Y}(X)$ is typically obtained by thresholding $\hat{s}(X)$.

[3] Identification is the process of using a set of assumptions to write a counterfactual quantity in terms of observable quantities

[4] We set the treatment to be the same value for all children in a family.

[5] Risk assessments are unnecessary for these cases since the decision-maker already knows what to do.

[6] In the paradigm where interventions are to be targeted at the *most responsive* cases, performance metrics such as discounted cumulative gain (DCG) or Spearman's rank correlation coefficients are more natural choices for evaluation. DR estimates can be constructed for these metrics as well.

The target counterfactual FPR is

$$\mathbb{E}[\hat{Y} \mid Y^0 = 0] \tag{3}$$

A model is well-calibrated in the counterfactual sense when

$$\mathbb{E}\Big[Y^0 \mid r_1 \leq \hat{s}(X) \leq r_2\Big] \approx \frac{r_1 + r_2}{2} \tag{4}$$

where r_1, r_2 define a bin of predictions.

3.3.1 Observational Evaluation. A standard approach evaluates the model against the observed outcomes. An observational Precision-Recall (PR) curve plots observational precision, $\mathbb{E}[Y \mid \hat{Y} = 1]$, against observational TPR[7], $\mathbb{E}[\hat{Y} \mid Y = 1]$. An observational ROC curve plots observational TPR against observational FPR $\mathbb{E}[\hat{Y} \mid Y = 0]$. An observational calibration curve plots $\mathbb{E}[Y \mid r_1 < \hat{s}(X) < r_2]$, the observational outcome rate for scores in the interval $[r_1, r_2]$. The observational evaluation answers the question: Does the RAI accurately predict the likelihood of an adverse outcome under the *observed historical decisions*? This evaluation approach can be misleading since $Y \not\equiv Y^0$. For instance, it will conclude that a valid counterfactual model of risk under baseline performs poorly because its predictions will be systematically inaccurate for cases that are responsive to treatment.

3.3.2 Evaluation on the Control Population. The standard counterfactual approach to evaluation computes error metrics on the control population [44]. The PR curve evaluated on the control population plots $\mathbb{E}[Y \mid \hat{Y} = 1, T = 0]$ against $\mathbb{E}[\hat{Y} \mid Y = 1, T = 0]$, and the ROC and calibration curves are similarly defined by conditioning on $T = 0$. When the control population is not representative of the full population (i.e. $T \not\perp X$), as is the case in non-experimental settings, this evaluation may be misleading since $\mathbb{E}[Y \mid T = 0] = \mathbb{E}[Y^0 \mid T = 0] \neq \mathbb{E}[Y^0]$. A method that performs well on the control population may perform poorly on the treated population (or vice-versa). In child welfare, cases where the perpetrator has a history of abuse are more likely to be screened in. Since there is more information associated with these cases, a model may be able to discriminate risk better for these cases than on cases in the control population with little history.

3.3.3 Doubly-robust (DR) Counterfactual Evaluation. We propose to improve upon the control population evaluation procedure by using DR estimation to perform counterfactual evaluation using both treated and control cases. This ensures that performance is assessed on a representative sample of the population. Our method estimates the counterfactual outcome for all cases and evaluates metrics on this estimate. Other approaches such as inverse-probability weighing (IPW) or plug-in estimates could be used for a counterfacutal evaluation, but DR techniques are preferable because they have faster rates of convergence for non-parametric methods, and for parametric methods they are robust to misspecification in one of the nuisance functions, i.e. the treatment propensity $\pi(X)$ and the outcome regression $\mathbb{E}[Y^0 \mid X]$ [23, 39, 41]. Under sample splitting and $n^{1/4}$ convergence rates in the nuisance function error terms, these estimates are $\sqrt{n}$-consistent and asymptotically normal. This enables us to compute confidence intervals (see Appendix B).

We first consider estimates of the average outcome under control $\mathbb{E}[Y^0]$. Under our causal assumptions in Section 3.2.2, $\mathbb{E}[Y^0] = \mathbb{E}[\mathbb{E}[Y \mid X, T = 0]]$. The plug-in estimate is: $\frac{1}{n}\sum_{i=1}^{n} \hat{s}_0(X_i)$ where $\hat{s}_0(X) = \hat{E}[Y^0|X]$ denotes the score of our counterfactual model. The IPW estimate uses the observed outcome on the control population and reweighs the control population to resemble the full population: $\frac{1}{n}\sum_{i=1}^{n} \frac{1-T_i}{1-\hat{\pi}(X_i)} Y_i$. DR estimators[8] combine the plug-in estimate with an IPW-residual bias-correction term for the control cases:

$$DR_{Y^0} = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1-T_i}{1-\hat{\pi}(X_i)}(Y_i - \hat{s}_0(X_i)) + \hat{s}_0(X_i)\Big] \tag{5}$$

Next we consider estimators for the counterfactual targets in Equations 1-4. We emphasize that $\hat{s}$ in (4) is the score of any model we wish to evaluate, while $\hat{s}_0$ refers specifically to our counterfactual model in § 3.2.2.

TPR (Recall): Counterfactual TPR is identified as

$$\mathbb{E}[\hat{Y} \mid Y^0 = 1] = \frac{\mathbb{E}\Big[\hat{Y}\mathbb{E}[Y \mid X, T = 0]\Big]}{\mathbb{E}\Big[\mathbb{E}[Y \mid X, T = 0]\Big]} \tag{6}$$

.

The DR estimate for the numerator is

$$\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i\Big[\frac{1-T_i}{1-\hat{\pi}(X_i)}(Y_i - \hat{s}_0(X_i)) + \hat{s}_0(X_i)\Big] \tag{7}$$

The DR estimate for the denominator is DR_{Y^0} in Equation 5.

Precision: The target counterfactual precision is identified as

$$\mathbb{E}[Y^0 \mid \hat{Y} = 1] = \mathbb{E}[\mathbb{E}[Y \mid X, T = 0] \mid \hat{Y} = 1] \tag{8}$$

The DR estimator for precision is

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1-T_i}{1-\hat{\pi}(X_i)}(Y_i - \hat{s}_0(X_i)) + \hat{s}_0(X_i)\Big]\mathbb{I}\{\hat{Y}_i = 1\}}{\mathbb{P}(\hat{Y}_i = 1)} \tag{9}$$

where $\mathbb{I}$ denotes the indicator function.

FPR:. The target counterfactual FPR is identified as

$$\mathbb{E}[\hat{Y} \mid Y^0 = 0] = \frac{\mathbb{E}\Big[\hat{Y}\mathbb{E}[1 - Y \mid X, T = 0]\Big]}{\mathbb{E}\Big[\mathbb{E}[1 - Y \mid X, T = 0]\Big]} \tag{10}$$

The DR estimator for the numerator is

$$\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i\Big[\frac{1-T_i}{1-\hat{\pi}(X_i)}(\hat{s}_0(X_i) - Y_i) + (1 - \hat{s}_0(X_i))\Big] \tag{11}$$

For the denominator we use $1 - DR_{Y^0}$ where DR_{Y^0} is in Eq 5.

Calibration: The target in Equation 4 is identified as

$$\mathbb{E}\Big[\mathbb{E}[Y \mid X, T = 0] \mid r_1 \leq \hat{s}(X) \leq r_2\Big]$$

The DR estimate for calibration is

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1-T_i}{1-\hat{\pi}(X_i)}(Y_i - \hat{s}_0(X_i)) + \hat{s}_0(X_i)\Big]\mathbb{I}\{r_1 \leq \hat{s}(X_i) \leq r_2\}}{\mathbb{P}(r_1 \leq \hat{s}(X_i) \leq r_2)} \tag{12}$$

We show how to compute confidence intervals in Appendix B.

[7] TPR and recall are equivalent.

[8] In survey inference, this is known as the generalized regression estimator [43].

3.4 Results

3.4.1 Synthetic example. We begin with a synthetic dataset so that we can compare methods in a setting where we observe both potential outcomes. We specify two groups with different treatment propensities, where the treatment is constructed to be equally effective at reducing the likelihood of adverse outcome ($Y = 1$) for both groups. We generate 100,000 data points (X_i, Y_i^0, Y_i^1, T_i) where $X_i = (Z_i, A_i)$ and $Z_i \sim \mathcal{N}(0, 1)$, a normal distribution with mean 0 and variance 1. $A_i \sim Bern(0.5)$, a Bernoulli with mean 0.5. $Y_i^0 \sim Bern(\sigma(Z_i - 0.5))$ where $\sigma(z) = \frac{1}{1+e^{-z}}$. $Y_i^1 \sim Bern(c\sigma(Z_i - 0.5))$ where $c = 0.1$ controls the treatment effect. $T_i \sim Bern(\sigma(Z_i - 0.5 + kA_i))$ where $k = 1.6$ describes the bias in treatment assignment toward group $A = 1$.[9] We set $Y = TY^1 + (1 - T)Y^0$.

We use logistic regression to train both the observational $\mathbb{E}[Y \mid X]$ and counterfactual models $\mathbb{E}[Y^0 \mid X]$ as well as the propensity model $\mathbb{E}[T \mid X]$.[10] Under this choice of model, the propensity model and counterfactual model are both correctly specified, and accordingly, the plug-in and IPW estimates are both consistent in this setting. However, in practice, there is no way to know whether the models are correctly specified, so DR estimates are preferable for real-world settings.

Figure 1 displays PR, ROC, and calibration curves.[11] DR evaluation most closely aligns with the true counterfactual evaluation. Notably, the observational evaluation suggests the observational model outperforms the counterfactual model. The true counterfactual evaluation shows the counterfactual model performs better.

3.4.2 Child Welfare. We also apply counterfactual learning and evaluation to the problem of child welfare screening. The baseline intervention is screen-out (which means no investigation occurs). The data consists of over 30,000 calls to the hotline in Allegheny County, Pennsylvania, each containing more than 1000 features describing the call information as well as county records for all individuals associated with the call. The call features are categorical variables describing the allegation types and worker-assessed risk and danger ratings. The county records include demographic information such as age, race and gender as well as criminal justice, child welfare, and behavioral health history. The outcome is re-referral within a six month period. Our approach contrasts to prior work which used placement out-of-home as the outcome [8, 11]. This outcome is only observed for cases under investigation; therefore it cannot be used to identify Y^0, the risk under no investigation.

We use random forests to train the observational and counterfactual risk assessments as well as the propensity score model. We used reweighing to correct for covariate shift but did not observe a boost in performance, likely because we have sufficient data and we used a non-parametric model.

We present the PR, ROC and calibration curves in Figure 2. The observational evaluation suggests that the observational model performs better. The control evaluation suggests that the counterfactual and observational models of risk perform equally well. Our DR evaluation suggests the counterfactual model has both better discrimination and calibration in estimating the probability of re-referral under screen-out. The observational evaluation suggests that the observational model is well-calibrated whereas the counterfactual model is overestimating risk; this is expected because the counterfactual model assesses risk under no investigation whereas the observed outcomes include cases whose risk was mitigated by child welfare services. The control evaluation suggests that the two models are similarly calibrated. The DR evaluation shows that the counterfactual model is well-calibrated and the observational model underestimates risk. This makes intuitive sense because the observational model is not accounting for that fact that treatment reduced risk for the screened-in cases.

We see further evidence that the observational model performs poorly on the treated population in the drop in ROC curves between the control evaluation and DR evaluation in Figure 2. Deploying such a model would mean failing to identify the people who need and would benefit from treatment. The observational and control evaluations do not show this significant limitation; DR evaluation is the only evaluation that illustrates the poor performance of the observational model on the treated population.

We also evaluate the different models according to whether they are equally predictive, in the sense of being equally well calibrated, across racial groups. Research suggests child welfare processes may disproportionately involve black families [12]. Here we ask whether the observational or counterfactual model is more equitable. We compare calibration rates by race in Figure 3. The observational evaluation suggests that the counterfactual model of risk is poorly calibrated by race. The DR evaluation shows that the counterfactual model is well-calibrated by race and indicates that the observational model underestimates risk on both black and white cases.

Overall the observational evaluation suggests that the observational model performs better whereas the DR evaluation suggests the counterfactual model performs better. Since we do not have access to the true counterfactual to validate these results, we further consider how well the models align with expert assessment of risk.

3.4.3 Expert Evaluation. At various stages in the child welfare process, social workers assign treatment based on their assessment of risk. Social workers decide 1) whether to screen in a case for investigation; 2) whether to offer services for a case under investigation; 3) whether to place a child out-of-home after an investigation. Assuming that social workers are competent at assessing risk, we expect the group placed out-of-home to have the highest risk distribution, followed by those offered services, and we expect those screened out to have the lowest risk. Figure 4 shows that the counterfactual model exhibits this expected behavior whereas the observational model does not. The observational model assesses the screened out population to have more high risk cases than any other treatment group. The observational model underestimates risk on the treated groups. These cases *should* be assigned treatment, but the observational model would suggest they should be screened out.

Such a mistake can have cascading effects. We are particularly concerned about screening out cases that, had they been screened in, would have been accepted for services or placed out-of-home. Figure 5 shows the recall for placed cases and serviced cases as we vary the proportion of cases classified as high-risk. The counterfactual model has much higher recall for both services and placement.

[9] We present results for alternative values of c and k in Appendix E. The offset -0.5 is to roughly balance the number of treated/control units

[10] In Appendix E.1 we include T as a feature in the observational model to see if this can appropriately control for treatment effects, but we find that it does not.

[11] The code for this experiment is given in https://github.com/mandycoston/counterfactual

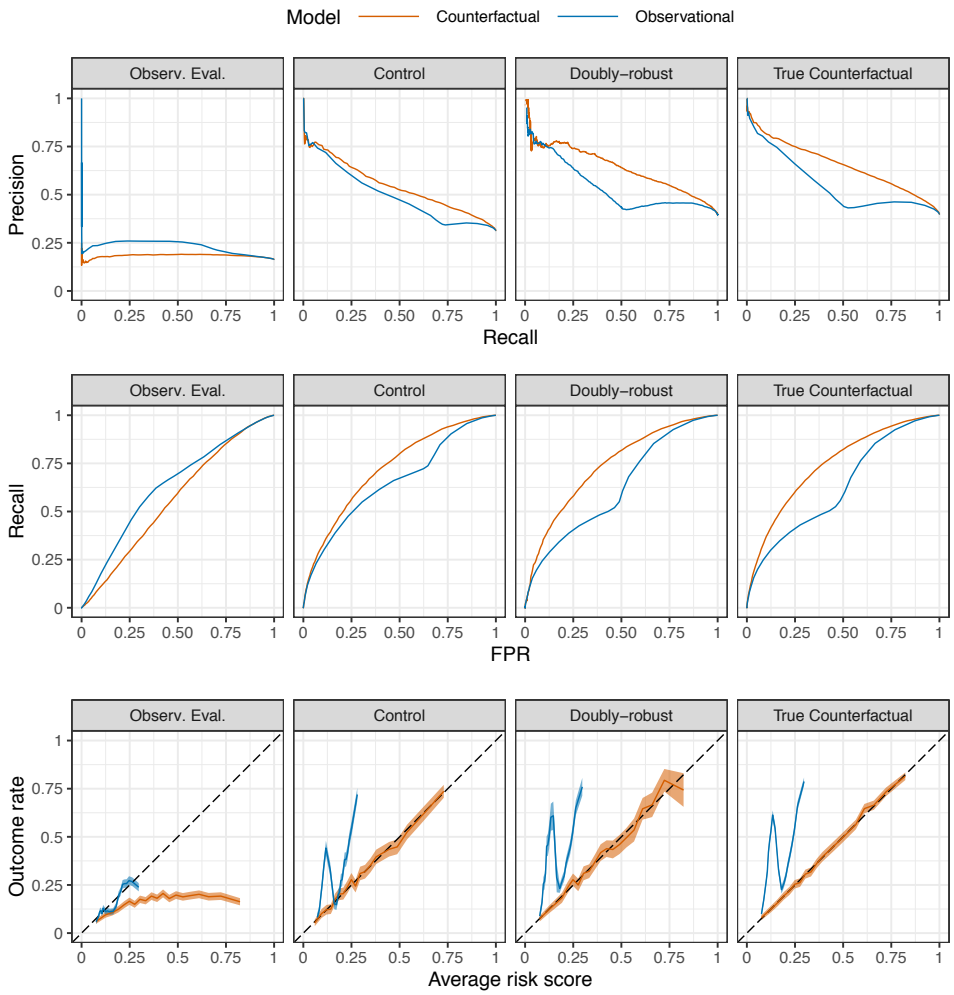

Figure 1: Synthetic data results. From top to bottom the rows give PR curves, ROC curves, and calibration curves with 95% pointwise confidence bounds. Each column is an evaluation method (§ 3.3). Colors denote the learning method (§ 3.2). DR evaluation most accurately represents the true counterfactual evaluation. Observational evaluation erroneously suggests the observational model performs better than the counterfactual model, because it includes units whose risk was mitigated by treatment. Control evaluation produces inaccurate curves because it does not assess how well the models perform on the treated population. (See § 3.4.1.)

3.4.4 Task adaptation. Another way to evaluate the models is to assess performance on related risk tasks. While the counterfactual risk models $\mathbb{E}[Y^0|X]$, we can assess how well it estimates $\mathbb{E}[Y^1|X]$, the risk under investigation. If we have reason to believe there will be common risk factors for risk under no investigation and risk under investigation, then we expect our model to perform well on this task. We use placement out-of-home, an adverse child welfare outcome that is observed for cases under investigation.

Table 1 shows that the observational model performs worse than a random classifier on the placement task whereas the counterfactual model shows some degree of discrimination. This suggests that the counterfactual model is learning a risk model that is useful in related risk tasks whereas the observational model is not.

	Observ. model	Counterfact. model	Random
AUROC	0.48 (0.46,0.49)	0.62 (0.61,0.63)	0.50
AUPR	0.13 (0.11,0.14)	0.18 (0.16,0.19)	0.14

Table 1: Area under ROC and PR curves using our re-referral models to predict a related risk task, out-of-home placement (95% confidence intervals given in parentheses). The observational model performs worse than a random classifier. The counterfactual learns a model of risk that transfers to related risk tasks. (See § 3.4.4.)

The comparison to expert assessment of risk and the performance on a downstream risk task support the conclusions of our

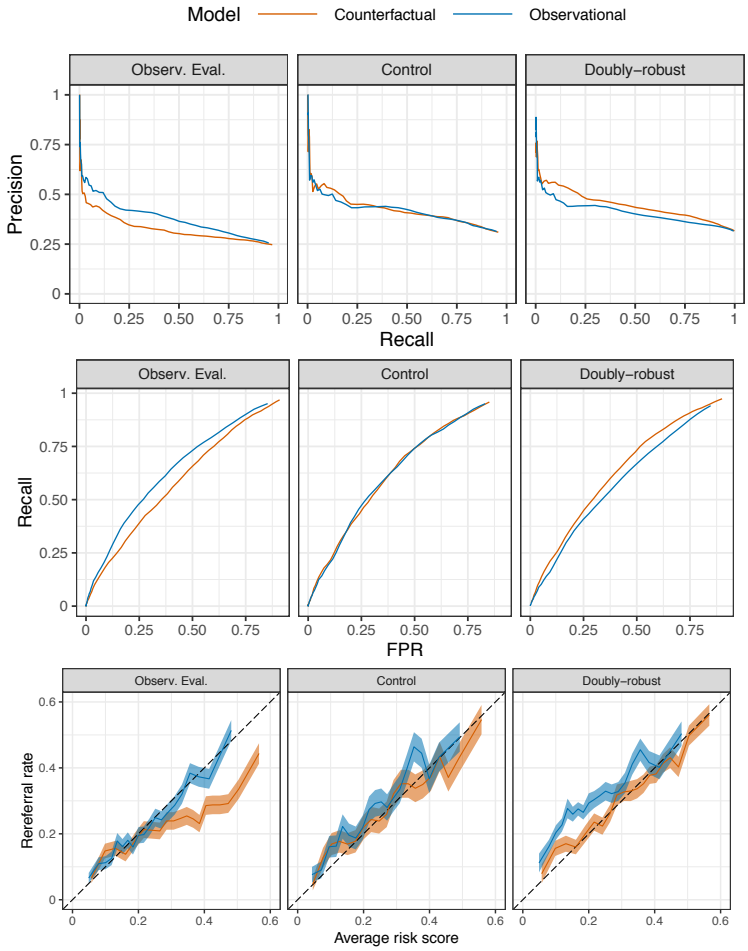

Figure 2: Child welfare results. From top to bottom the rows give PR curves, ROC curves, and calibration curves with 95% pointwise confidence bounds. Each column is an evaluation method § 3.3. Colors denote the learning method § 3.2. Observational evaluation suggests the observational model has better discrimination and calibration than the counterfactual model because it evaluates against the observed outcomes which include cases whose risk was mitigated by child welfare services. Control evaluation suggests the two models perform similarly on cases that did not receive treatment. DR evaluation shows that the observational model does not perform well on treated cases. (See § 3.4.2 and 3.4.3.)

DR evaluation. In decision-making contexts, failure to account for treatment effects can lead one to the wrong conclusions about model performance, leading to the deployment of a model that underestimates risk for those who stand to gain most from treatment.

4 COUNTERFACTUAL FAIRNESS

Standard observational notions of algorithmic fairness are subject to the same pitfalls as observational model evaluation. In this section we propose counterfactual formulations of several fairness metrics and analyze the conditions under which the standard (observational) metric implies the counterfactual one.

We motivate the importance of defining these metrics counterfactually with an example. Suppose teachers are assessing a model that predicts who is likely to fail an exam which they intend to use to assign tutoring resources. Suppose anyone tutored will pass. The tutoring session conflicts with girls' sports practice so only male students are tutored. A model that perfectly predicts who will fail without the help of a tutor will have a higher observational FPR for men than women because some male students were tutored, which enabled them to pass. It would be wrong to conclude that this model is unfair with regards to FPR. Someone who would have been high-risk had they not been treated but whose risk was mitigated under treatment should not be considered a false positive. Failure to make this distinction could lead to unfairness, not only in settings where the treatment assignment varies according to the protected attribute but also in settings where the risk under treatment varies according to the protected attribute. For instance suppose now both girls and boys are tutored but the tutor is only effective in preparing male students to pass. The model that perfectly predicts who will

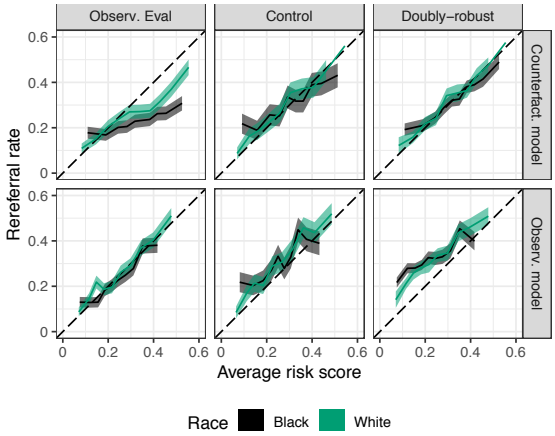

Figure 3: Calibration curves by race for child welfare. Counterfactual model (top row) is well-calibrated by race according to the control and DR evaluations but shows inequities according to the observational evaluation because black cases were more likely to get treatment which mitigates risk (see § 3.4.2 for more details). The observational model (bottom row) is poorly calibrated for both black and white cases according to the DR evaluation.

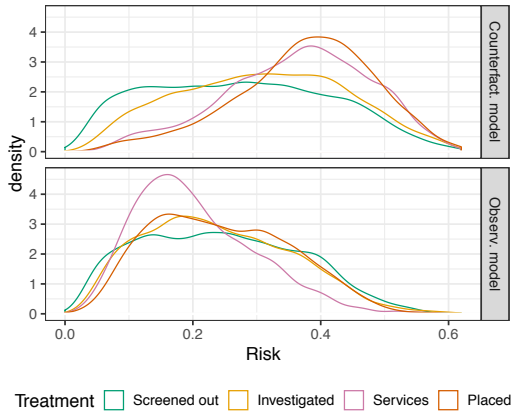

Figure 4: Child welfare risk distributions by treatment type for counterfactual and observational models. We expect risk to increase with the severity of treatment assigned, with 'Placed' out-of-home having the highest risk distribution and 'Screened out' of investigation having the lowest (see § 3.4.3). The counterfactual model displays this trend. The observational model does not, underestimating risk on cases where child welfare effectively mitigated the risk

fail without a tutor has a higher observational FPR for men, but as before, it is wrong to conclude that the model is unfair.

We distinguish our notion of counterfactual fairness from prior work which considered counterfactuals of the protected attribute [26, 31, 53], an approach which is counterproductive in our settings of interest. Consider a female student who is at high risk of failing because of gender discrimination at home or in the classroom, e.g.

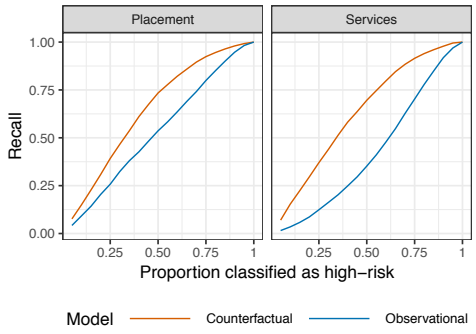

Figure 5: Recall for downstream child welfare decisions. At current screen-in rates (0.5), the observational model would screen out nearly 50% of very high risk cases that were placed out-of-home. The counterfactual model has higher recall at 73%. The gap is even larger for cases that were accepted for services. (See § 3.4.3.)

parents or previous teachers have not given her the support they would have had she been male. Treating this student "counterfactually as if she had been male all along" may suggest that we should not assign this student a tutor. In fact we *must* assign her a tutor in order to correct historical discrimination. Similar arguments can be made in settings like child welfare screening and loan approvals.

4.1 Theoretical results

For three definitions of fairness (parity), we show that observational parity implies counterfactual parity if and only if a balance condition holds. We show that an independence condition is sufficient for observational parity to imply counterfactual parity. We discuss why it is generally unlikely that the independence condition holds and even more unlikely that the finer balance condition holds when the independence condition fails. Appendix C contains the proofs.

4.1.1 Base Rate Parity. Base rate plays a core role in statistical definitions of fairness (also known as group fairness). Base rate parity is similar to the fairness notion of demographic parity, which requires $\hat{Y} \perp A$ [5, 15, 54]. In § 4.2, we perform experiments on a fairness corrective method that targets base rate parity in order to encourage demographic parity [20]. A related fairness notion, prediction-prevalence parity, requires $\mathbb{E}[Y \mid a] = \mathbb{E}[\hat{Y} \mid a]$. Satisfying both prediction-prevalence parity and demographic parity requires parity in the base rates. We distinguish observational base rate parity (oBP) $Y \perp A$ from counterfactual base rate parity (cBP), which requires $Y^0 \perp A$, where Y^0 is the potential outcome under the baseline treatment.

Theorem 1 (Base Rate Parity). *Assume* $\mathbb{P}(T = 0 \mid y^0, a) \neq 0$. *If oBP holds, then cBP holds if and only if the following balance condition holds.*

Condition (balBP).

$$\mathbb{P}(Y^1 = y)\mathbb{P}(T = 1 \mid Y^1 = y) - \mathbb{P}(Y^1 = y \mid a)\mathbb{P}(T = 1 \mid Y^1 = y, a)$$
$$= \mathbb{P}(Y^0 = y)\Big(\mathbb{P}(T = 1 \mid Y^0 = y) - \mathbb{P}(T = 1 \mid Y^0 = y, a)\Big) \quad (13)$$

BalBP holds under the following independence conditions, which provide sufficient conditions for oBP to imply cBP.

Condition (indBP).

$$\begin{aligned} T &\perp A \mid Y^0 \\ (Y^1, T) &\perp A \end{aligned} \tag{14}$$

It is unlikely that indBP holds in many contexts. In child welfare and criminal justice, research suggests that even when controlling for true risk, certain races are more likely to receive treatment [2, 12, 33]. indBP cannot hold since $T \not\perp A \mid Y^0$. Even in settings where there is no such bias, indBP will not hold if the risk distributions under treatment vary by protected attribute since indBP requires $Y^1 \perp A$. indBP also requires $T \perp A \mid Y^1$, which forbids discrimination in treatment assignment when controlling for risk under treatment. If indBP does not hold, it is possible that balBP still holds if the conditional and marginal probabilities are such that all terms exactly cancel; however there is no semantic reason why this should hold. Theorem 1 assumes $\mathbb{P}(T = 0 \mid y^0, a) \neq 0$, a positivity-like assumption that holds in all settings that are suitable for algorithmic RAIs. Violations of this assumption indicate perfect or imperfect treatment assignment historically for a group.

4.1.2 Predictive parity. Base parity and demographic parity may be ill-suited for settings where base rates differ by protected attribute due to disparate needs. Here we may instead desire parity in an error metric, such as precision. Positive predictive parity requires the precision (also known as positive predictive value) to be independent of the protected attribute, and negative predictive parity requires the negative predictive value to be independent of the protected attribute [7, 28]. We define observational Predictive Parity (oPP) as $Y \perp A \mid \hat{Y} = \hat{y}$ and counterfactual Predictive Parity (cPP) as $Y^0 \perp A \mid \hat{Y} = \hat{y}$ where $\hat{y} = 0$ corresponds to negative predictive parity and $\hat{y} = 1$ corresponds to positive predictive parity.

Theorem 2 (Predictive Parity). *Assume* $\mathbb{P}(T = 0 \mid y^0, a, \hat{y}) \neq 0$. *If oPP holds, then cPP holds if and only if the following balance condition holds.*

Condition (balPP).

$$\begin{aligned} &\mathbb{P}(Y^1 = y \mid \hat{y})\mathbb{P}(T = 1 \mid Y^1 = y, \hat{y}) \\ &- \mathbb{P}(Y^1 = y \mid a, \hat{y})\mathbb{P}(T = 1 \mid Y^1 = y, a, \hat{y}) \\ &= \mathbb{P}(Y^0 = y \mid \hat{y})\Big(\mathbb{P}(T = 1 \mid Y^0 = y, \hat{y}) - \mathbb{P}(T = 1 \mid Y^0 = y, a, \hat{y})\Big) \end{aligned} \tag{15}$$

BalPP is satisfied under the following independence conditions, which provide sufficient conditions for oPP to imply cPP.

Condition (indPP).

$$\begin{aligned} T &\perp A \mid Y^0, \hat{Y} \\ (Y^1, T) &\perp A \mid \hat{Y} \end{aligned} \tag{16}$$

IndPP will not hold in many settings. Note that $(Y^1, T) \perp A \mid \hat{Y} \iff T \perp A \mid Y^1, \hat{Y}$ and $Y^1 \perp A \mid \hat{Y}$. Conditions $T \perp A \mid Y^t, \hat{Y}$ require $\hat{Y}$ to contain all the information that A tells us about treatment assignment that is not contained in Y^t. Since $\hat{Y}$ is typically trained to predict Y and not T, it is quite unlikely that these conditions will hold in settings where there is bias in treatment assignment even when controlling for true risk. Condition $Y^1 \perp A \mid \hat{Y}$ allows differences in the risk distribution under treatment if we can fully explain these differences with $\hat{Y}$. In the best case $\hat{Y} \approx Y$, but it is unlikely that the observed outcome, which is not causally well-defined, would explain differences in the risk distribution under treatment. As above, even if indPP does not hold, balPP may hold, but it is difficult to reason why this should hold in any setting. Like Theorem 1, Theorem 2 assumes a mild positivity-like assumption.

4.1.3 Equalized odds. In settings where TPR and FPR are more important than predictive value, we may desire parity in TPR and FPR, a fairness notion known as Equalized Odds [17]. Observational Equalized Odds (oEO) requires that $\hat{Y} \perp A \mid Y$, and counterfactual Equalized Odds (cEO) requires that $\hat{Y} \perp A \mid Y^0$.

Theorem 3 (Equalized Odds). *Assume* $\mathbb{P}(Y = y \mid a) \neq 0$ *and* $\mathbb{P}(T = 0 \mid y^0, a, \hat{y}) \neq 0$. *If oEO holds, then cEO holds if and only if the following balance condition holds. .*

Condition (balEO).

$$\begin{aligned} &\mathbb{P}(\hat{Y} = 1 \mid Y^1 = y)\frac{\mathbb{P}(T = 1 \mid \hat{Y} = 1, Y^1 = y)\mathbb{P}(Y^1 = y)}{\mathbb{P}(Y = y)} \\ &- \mathbb{P}(\hat{Y} = 1 \mid Y^1 = y, a)\frac{\mathbb{P}(T = 1 \mid \hat{Y} = 1, Y^1 = y, a)\mathbb{P}(Y^1 = y \mid a)}{\mathbb{P}(Y = y \mid a)} \\ &= \mathbb{P}(\hat{Y} = 1 \mid Y^0 = y)\Bigg(\frac{\mathbb{P}(T = 0 \mid \hat{Y} = 1, Y^0 = y, a)\mathbb{P}(Y^0 = y \mid a)}{\mathbb{P}(Y = y \mid a)} \\ &- \frac{\mathbb{P}(T = 0 \mid \hat{Y} = 1, Y^0 = y)\mathbb{P}(Y^0 = y)}{\mathbb{P}(Y = y)}\Bigg) \end{aligned} \tag{17}$$

The balance condition is satisfied under the following independence conditions, which comprise sufficient conditions for oEO to imply cEO.

Condition (indEO).

$$\begin{aligned} Y &\perp A \\ Y^0 &\perp A \\ T &\perp A \mid \hat{Y}, Y^0 \\ (Y^1, \hat{Y}, T) &\perp A \end{aligned} \tag{18}$$

The first two conditions of indEO require oBP and cBP, so indEO requires balBP to hold. In settings where there is discrimination in treatment assignment even when controlling for true risk, indEO is unlikely to hold. Even if there is no such discrimination, indEO will not hold if there are differences in the risk distributions under treatment since the last condition of 18 requires $Y^1 \perp A$. indEO requires further conditions such as parity in the TPR/FPR against the outcome under treatment. If these conditions are not met, oEO could imply cEO if balEO holds, but it is difficult to reason about why this would hold when the independence conditions do not.

Our theoretical analysis suggests that in many settings, equalizing the observational fairness metric will not equalize the counterfactual fairness metric. We conclude by noting that the theorems hold when conditioning on any feature(s) $\subseteq X$, and in this context, these theorems are relevant to individual notions of fairness.

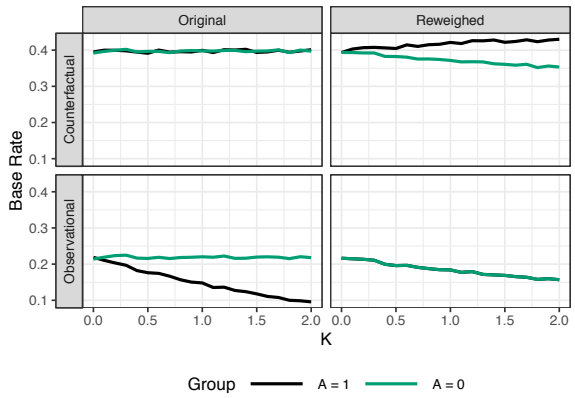

Figure 6: Counterfactual and observational base rates before and after applying a fairness-corrective method that reweighs training data (§ 4.2.1). X-axis controls the bias of treatment assignment toward group $A = 1$. Before reweighing ("Original"), counterfactual base rates are equal (cBP holds), but observational base rates are different (oBP doesn't hold) for $k > 0$ since group $A = 1$ is more likely to get treated. Reweighing achieves oBP but cBP no longer holds.

4.2 Experiments on synthetic data

We empirically demonstrate that equalizing the observational metric can increase disparity in the counterfactual metric.[12]

4.2.1 Reweighing. One approach to encourage demographic parity reweighs the training data to achieve base rate parity [20]. Figure 6 shows that without any processing ("Original"), the counterfactual base rates are equal, while the observational base rates show increasing disparity with k. Reweighing the observational outcome achieves oBP but induces disparity in the counterfactual base rate. Theorem 1 suggested this result: For $k > 0$, $A \not\perp T \mid Y^0$; then it is unlikely that oBP implies cBP.

4.2.2 Post-processing. We evaluate a method that modifies scores to achieve a generalized version of equalized odds [17, 37].[13] This method targets parity in the generalized FNR/FPR, where GFPR is $\mathbb{E}[\hat{s}(X) \mid Y = 0]$ and GFNR is $\mathbb{E}[1 - \hat{s}(X) \mid Y = 1]$. We refer to these observational rates as oGFPR/oGFNR and define their counterfactual counterpart: cGFPR = $\mathbb{E}[\hat{s}(X) \mid Y^0 = 0]$ and cGFNR = $\mathbb{E}[1 - \hat{s}(X) \mid Y^0 = 1]$. We use the scores of the counterfactual model as inputs. We compute the cGFNR and cGFPR using our DR method from § 3.3.3.[14]

Table 2 shows that post-processing to equalize oGFPR and oGFNR induces imbalance in cGFPR and cGFNR.[15] In Figure 7 we see that the original model achieved cEO, but post-processing induced disparity to the detriment of the group that was less likely to be treated.

[12]We perform these experiments on the synthetic data (§ 3.4.1). We do not use the child welfare data since it is balanced in terms of base rates and FPR/TPR with respect to race.
[13]We use the Pleiss implementation on https://github.com/gpleiss/equalized_odds_and_calibration that extends the method in [17] to probabilistic classifiers.
[14]The estimator is nearly identical to the estimators for FPR/FNR if we use $\hat{s}(X)$ in place of the predicted label $\hat{Y}(X)$
[15]We use $c = 0.1$ and $k = 1.6$. We report results for other values in Appendix F.

Group	Method	cGFNR	cGFPR	oGFNR	oGFPR
A=1	Original	0.50	0.33	0.58	0.39
A=0	Original	0.50	0.33	0.56	0.39
A=1	Post-Proc.	0.58	0.30	0.63	0.35
A=0	Post-Proc.	0.64	0.34	0.63	0.35

Table 2: Counterfactual and observational generalized FNR/FPR before and after post-processing to equalize odds (§ 4.2.2) using threshold = 0.5. Before post-processing ("Original"), the counterfactual generalized rates (cGFNR and cGFPR) are the same for both groups. Post-processing equalizes the observational rates (oGFNR and oGFPR) but induces noticeable disparity in both cGFNR and cGFPR.

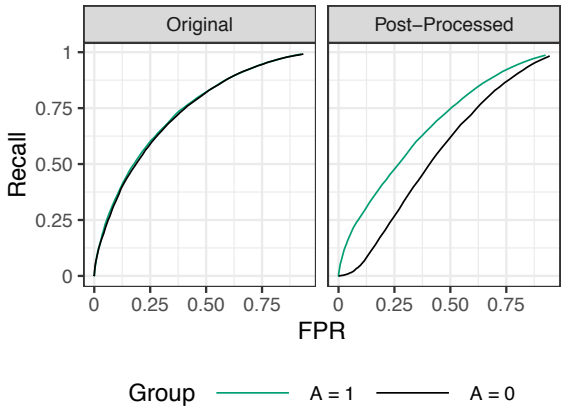

Figure 7: Counterfactual ROC curves before and after post-processing to equalize odds (§ 4.2.2). Before post-processing, ROC curves are identical for both groups, indicating that counterfactual equalized odds (cEO) holds. Post-processing induces imbalance, harming group $A = 0$ and compounding initial unfairness in treatment assignment.

Since treatment is beneficial, this "fairness" adjustment actually compounded the discrimination in the treatment assignment.

5 CONCLUSION

This paper demonstrates that training and evaluating models using observed outcomes can lead to the misallocation of resources due to the misestimation of risk for those most receptive to treatment. Furthermore, fairness-correcting methods that seek to achieve observational parity can lead to disparities on the relevant counterfactual metrics, and may further compound inequities in intial treatment assignment. The counterfactual approaches to learning, evaluation and predictive fairness assessment introduced in this paper provide more accurate and relevant indications of model performance.

ACKNOWLEDGMENTS

We are grateful to the Block Center for Technology and Society for funding this research, and to our collaborators at the Allegheny County Department of Human Services. Thanks also to our reviewers for helpful suggestions for improving the manuscript.

REFERENCES

[1] Ahmed M Alaa and Mihaela van der Schaar. 2017. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*. 3424–3432.

[2] Michelle Alexander. 2011. The new jim crow. *Ohio St. J. Crim. L.* 9 (2011), 7.

[3] Chelsea Barabas, Karthik Dinakar, Joichi Ito, Madars Virza, and Jonathan Zittrain. 2017. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. *arXiv preprint arXiv:1712.08238* (2017).

[4] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[5] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18.

[6] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1721–1730.

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

[8] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. 134–148.

[9] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[10] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 797–806.

[11] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. 2018. Learning under selective labels in the presence of expert consistency. *arXiv preprint arXiv:1807.00905* (2018).

[12] Alan J Dettlaff, Stephanie L Rivaux, Donald J Baumann, John D Fluke, Joan R Rycraft, and Joyce James. 2011. Disentangling substantiation: The influence of race, income, and risk on the substantiation decision in child welfare. *Children and Youth Services Review* 33, 9 (2011), 1630–1637.

[13] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[14] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601* (2011).

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

[16] Andrew Guthrie Ferguson. 2016. Policing predictive policing. *Wash. UL Rev.* 94 (2016), 1109.

[17] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[18] Nathan Kallus and Angela Zhou. 2018. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*. 9269–9279.

[19] Nathan Kallus and Angela Zhou. 2018. Residual Unfairness in Fair Machine Learning from Prejudiced Data. In *Proc. International Conference on Machine Learning*. Stockholm, Sweden, 2439–2448.

[20] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[21] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 643–650.

[22] Danielle Leah Kehl and Samuel Ari Kessler. 2017. Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. (2017).

[23] Edward H Kennedy. 2016. Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*. Springer, 141–167.

[24] Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. 2017. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79, 4 (2017), 1229–1245.

[25] Edward H Kennedy, Wyndy L Wiitala, Rodney A Hayward, and Jeremy B Sussman. 2013. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. *Medical care* 51, 3 (2013), 251.

[26] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

[27] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review* 105, 5 (2015), 491–95.

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

[29] Amanda Kube, Sanmay Das, and Patrick J Fowler. 2019. Allocating interventions based on predicted outcomes: A case study on homelessness services. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[30] Matt Kusner, Chris Russell, Joshua Loftus, and Ricardo Silva. 2019. Making Decisions that Reduce Discriminatory Impacts. In *International Conference on Machine Learning*. 3591–3600.

[31] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[32] David Madras, Elliot Creager, Toniann Pitassi, and Richard 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.

[33] Marc Mauer. 2010. Justice for all-challenging racial disparities in the criminal justice system. *Hum. Rts.* 37 (2010), 14.

[34] Shira Mitchell, Eric Potash, and Solon Barocas. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867* (2018).

[35] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[36] J Neyman. 1923. Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes (Masters Thesis); Justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. Excerpts English translation (Reprinted). *Stat Sci* 5 (1923), 463–472.

[37] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.

[38] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2009. *Dataset shift in machine learning*. The MIT Press.

[39] James M Robins and Andrea Rotnitzky. 1995. Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* 90, 429 (1995), 122–129.

[40] James M Robins and Andrea Rotnitzky. 2001. Inference for semiparametric models: Some questions and an answer-Comments.

[41] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89, 427 (1994), 846–866.

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

[43] Carl-Erik Särndal, Bengt Swensson, and Jan H Wretman. 1989. The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika* 76, 3 (1989), 527–537.

[44] Peter Schulam and Suchi Saria. 2017. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*. 1697–1708.

[45] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3076–3085.

[46] Vernon C Smith, Adam Lange, and Daniel R Huston. 2012. Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses. *Journal of Asynchronous Learning Networks* 16, 3 (2012), 51–61.

[47] Megan Stevenson. 2018. Assessing risk assessment in action. *Minn. L. Rev.* 103 (2018), 303.

[48] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. 2007. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research* 8, May (2007), 985–1005.

[49] Adith Swaminathan and Thorsten Joachims. 2015. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research* 16, 1 (2015), 1731–1755.

[50] Administration U.S. Department of Health & Human Services. 2019. Child Maltreatment 2017. https://www.acf.hhs.gov/cb/research-data-technology/statistics-research/child-maltreatment

[51] Mark J Van der Laan, MJ Laan, and James M Robins. 2003. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

[52] Tyler J VanderWeele and Miguel A Hernan. 2013. Causal inference under multiple versions of treatment. *Journal of causal inference* 1, 1 (2013), 1–20.

[53] Yixin Wang, Dhanya Sridhar, and David M Blei. 2019. Equal Opportunity and Affirmative Action via Counterfactual Predictions. *arXiv preprint arXiv:1905.10870* (2019).

[54] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2015. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259* (2015).

[55] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

[56] Junzhe Zhang and Elias Bareinboim. 2018. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*. 3671–3681.
[57] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making – the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*.