
PARTIAL IDENTIFICATION APPROACH TO COUNTERFACTUAL FAIRNESS ASSESSMENT

Saeyoung Rho

Department of Computer Science
Columbia University
New York, NY
s.rho@columbia.edu

Junzhe Zhang

Department of Electrical Engineering and Computer Science
Syracuse University
Syracuse, NY
jzhan403@syr.edu

Elias Bareinboim

Department of Computer Science
Columbia University
New York, NY
eb@cs.columbia.edu

ABSTRACT

The wide adoption of AI decision-making systems in critical domains such as criminal justice, loan approval, and hiring processes has heightened concerns about algorithmic fairness. As we often only have access to the output of algorithms without insights into their internal mechanisms, it was natural to examine how decisions would alter when auxiliary sensitive attributes (such as race) change. This led the research community to come up with counterfactual fairness measures, but how to *evaluate* the measure from available data remains a challenging task. In many practical applications, the target counterfactual measure is not identifiable, i.e., it cannot be uniquely determined from the combination of quantitative data and qualitative knowledge. This paper addresses this challenge using partial identification, which derives informative bounds over counterfactual fairness measures from observational data. We introduce a Bayesian approach to bound unknown counterfactual fairness measures with high confidence. We demonstrate our algorithm on the COMPAS dataset, examining fairness in recidivism risk scores with respect to race, age, and sex. Our results reveal a positive (spurious) effect on the COMPAS score when changing race to African-American (from all others) and a negative (direct causal) effect when transitioning from young to old age.

Keywords Counterfactual Fairness · Fairness Auditing · Fairness Measure · COMPAS · Algorithmic Fairness

1 Introduction

Algorithmic decision-making systems have become an integral part of our lives, significantly influencing critical aspects of our society where fairness and equity are paramount. From determining access to healthcare resources to shaping lending practices and criminal justice outcomes, these algorithms have the ability to either uphold or compromise justice [Lee et al., 2019]. While we cannot halt the use of those algorithms, we have both the opportunity and the obligation to equip ourselves to examine their behaviors and evaluate how fair these algorithms are [Mehrabi et al., 2021, Mitchell et al., 2021].

One of the primary challenges in assessing fairness is our limited understanding of the inner workings of these algorithms, as they function as opaque “black boxes” [Adebayo et al., 2016, Saleiro et al., 2018]. Typically, we only have access to the outputs of the algorithm, such as recommendations or decisions, while the inputs and the intricate processes concealed within these black boxes remain hidden [Bandy, 2021]. Consequently, most fairness metrics discussed so far are about retrospectively evaluating the parity of fairness of outcomes—such as False Negative Rate (FNR), False Positive Rate (FPR), and Positive Predicted Value (PPV, the positive predictions that turn out to

be indeed positive)—across groups by race or gender. However, it is known that these parities cannot be achieved unless we assume that the true distribution also achieves parity [Chouldechova, 2017, Berk et al., 2021]. Moreover, the observational criteria often fail to capture indirect discrimination through proxies [Kilbertus et al., 2017].

What if we assume access to a detailed causal model that generates the data? By manipulating the model’s causal mechanisms, one could simulate thought experiments to explore hypothetical scenarios where an attribute believed by society to be irrelevant is altered. For instance, when deciding on a loan approval, one could imagine a hypothetical scenario where the applicant’s race is altered: i.e., asking a counterfactual query, “What if an individual’s race were not white but black while holding all the other attributes unchanged?” [Kusner et al., 2017]. If the algorithm’s outputs change in this alternative scenario, we may conclude that the algorithm was unfair. Based on this intuition, researchers have put forth various definitions/measurements of counterfactual fairness over the past few years [Kusner et al., 2017, Chiappa, 2019, Zhang and Bareinboim, 2018, Nabi and Shpitser, 2018, Zhang and Wu, 2017, Kilbertus et al., 2017]. These measures take into consideration the underlying disparity in true distribution across groups if it exists, and also well-represent primary anti-discrimination frameworks applied in legal systems throughout the US and the EU: disparate impact and disparate treatment [Barocas and Selbst, 2016].

Despite their transparency and intuitiveness, several challenges exist in evaluating these counterfactual fairness measures. In many applications, the detailed causal model is often not fully known; instead, the learner/investigator can access data generated by the algorithm through passive observation or auditing. First of all, it necessitates the construction of a causal diagram [Pearl, 1995] encoding qualitative causal relationships among variables in the data-generating model. Learning valid causal diagrams is an active area of research studied under the rubrics of causal discovery [Pearl, 2000, Spirtes et al., 2000, Petersen et al., 2006]. Even if we can access an accurate causal diagram, obtaining counterfactual probabilities based on observational data is still challenging (Section 4.4 of Plecko and Bareinboim [2022]). The causal knowledge encoded in the diagram is often insufficient to allow target counterfactual probabilities to be identifiable, i.e., they cannot be uniquely discernible from available data [Pearl, 2000, Ch. 3]. Some algorithms have been suggested to identify counterfactual fairness measures. Still, they are restricted to identifiable cases only and/or work under strong parametric assumptions such as linearity in the model [Nabi and Shpitser, 2018, Wu et al., 2019, Zhang et al., 2017].

Our goal in this paper is to address these challenges. We introduce a novel partial identification algorithm that could bound any counterfactual fairness measure from observational data. Our algorithm can be applied to any categorical data, without assuming strong parametric functional assumption, and derive informative bounds with theoretical guarantees. Also, we demonstrate the algorithm on the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) recidivism risk score dataset [Angwin et al., 2016]. Through this case study, we explain the causal mechanisms behind the discrimination embedded within the COMPAS algorithm on the basis of race, sex, and age. Our research contributes to the ongoing efforts to correctly evaluate the (un)fairness of the algorithms while access to the full picture is limited. More specifically, our contributions are summarized as follows:

1. We develop a novel partial identification algorithm for bounding counterfactual fairness measures from observational data, with a theoretical guarantee;
2. We evaluate our algorithm using both simulations and real-world datasets.
3. Using our algorithm, we analyze behaviors of the COMPAS algorithm and reveal new explanations for its disparate impact on defendants with minority backgrounds.

Given space constraints, additional details of the experimental setups are provided in Appendix A.

2 Preliminaries and Related Work

In this section, we introduce the necessary definitions and theorems that we use throughout the paper. The basic notations are as follows. We use capital letters to denote variables ($\mathbf{X}$), small letters for their values (x), and Ω_X for their domains. For an arbitrary set Ω , let $|\Omega|$ be its cardinality. The distribution over variables $\mathbf{X}$ is denoted by $P(\mathbf{X})$. For convenience, we consistently use $P(x)$ as a shorthand for the probability $P(\mathbf{X} = x)$. Finally, the indicator function $\mathbb{I}_{\mathbf{X}=x}$ returns 1 if an event $\mathbf{X} = x$ holds; otherwise, $\mathbb{I}_{\mathbf{X}=x}$ is equal to 0.

Structural Causal Models. The basic semantical framework of our analysis rests on *structural causal models* (SCMs) [Pearl, 2009]. An SCM $\mathcal{M}$ is a tuple $\langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$ where $\mathbf{V}$ is a set of endogenous variables and $\mathbf{U}$ is a set of exogenous variables. $\mathcal{F}$ is a set of functions where each $f_V \in \mathcal{F}$ decides values of an endogenous variable $V \in \mathbf{V}$, taking a combination of other variables in the system as an argument. That is, $v \leftarrow f_V(\text{pa}_V, u_V)$, $\text{PA}_V \subseteq \mathbf{V}$, $U_V \subseteq \mathbf{U}$. Exogenous variables $U \in \mathbf{U}$ are mutually independent, values of which are drawn from the exogenous distribution $P(\mathbf{U})$. Naturally, $\mathcal{M}$ induces a joint distribution $P(\mathbf{V})$ over endogenous variables $\mathbf{V}$, called the *observational distribution*. Each SCM $\mathcal{M}$ is also associated with a causal diagram $\mathcal{G}$ (e.g., Fig. 1), which is a directed acyclic graph

(DAG) where solid nodes represent endogenous variables V , empty nodes represent exogenous variables U , and arrows represent the arguments PA_V, U_V of each structural function f_V .

Intervention on an arbitrary subset $X \subseteq V$, denoted by $\text{do}(x)$, is an operation where values of X are set to constants x , regardless of how they are ordinarily determined. For an SCM $\mathcal{M}$, let $\mathcal{M}_x$ denote a submodel of $\mathcal{M}$ induced by intervention $\text{do}(x)$. For any subset $Y \subseteq V$, the *potential response* $Y_x(u)$ is defined as the solution of Y in the submodel $\mathcal{M}_x$ given $U = u$. Drawing values of exogenous variables U following the distribution $P(U)$ induces a *counterfactual variable* Y_x . Specifically, the event $Y_x = y$ (for short, y_x) can be read as “ Y would be y had X been x ”. For subsets $Y, \dots, Z, X, \dots, W \subseteq V$, the distribution over counterfactuals $Y_x, \dots, Z_w$ is defined as:

$$P(y_x, \dots, z_w) = \int_{\Omega_U} \mathbb{I}_{Y_x(u)=y, \dots, Z_w(u)=z} dP(u). \quad (1)$$

Distributions of the form $P(Y_x)$ are called *interventional distributions*; when $X = \emptyset$, $P(Y)$ coincides with the *observational distribution*. Throughout this paper, we assume that domains of endogenous variables V are discrete and finite; while exogenous variables U could take values in any (continuous) domains. For a more detailed survey on SCMs, see [Pearl, 2009].



Figure 1: Causal diagrams representing a standard fairness model containing a protected attribute A (e.g., race), an outcome Y (recidivism score), a confounder Z (birthplace) and a mediator W (prior criminal records).

Causal Fairness Measures. The language of structural causality allows one to formalize and articulate concepts that are not easily defined in classic statistical theory. For instance, we can measure the impact of unfair and discriminatory practices with mathematical precision by simulating thought experiments through hypothetical interventions on structural equations. We will consistently use A to stand for a protected attribute; Y for the primary outcome; Z for all the observed confounders affecting A and Y , i.e., their common causes; and W for the descendants of A that also affect Y , which we call mediators.

For example, Fig. 1a shows a causal diagram of a standard fairness model [Zhang and Bareinboim, 2018] representing the data-generating process of some recidivism score [Angwin et al., 2016]. Here, A stands for the race of the defendant; Y for the predicted recidivism score; Z for their birthplace; and W for their prior criminal records. The judge may exhibit increased strictness towards the same action due to racial bias ($A \rightarrow W$). Also, birthplace can influence an individual’s race ($Z \rightarrow A$), potentially due to historical racial segregation. Finally, all observed variables may impact the algorithmic prediction Y .

The counterfactual probability inspired researchers to define *counterfactual fairness* measures [Kusner et al., 2017]. Kusner et al. say that an algorithm is *counterfactually fair* if one’s prediction outcome Y in the real world is similar to the Y in the counterfactual world where A has a different value. Let $X = Z \cup W$ be other predictors affecting outcome Y . Formally,

Definition 2.1 (Counterfactual Effect). Under any context $X = x$ and $A = a$, we define the counterfactual effect of an intervention $A = a_0$ on Y , with baseline $A = a_0$, as

$$\text{CE}_{a_0, a_1}(y|x, a) = P(y_{a_1}|x, a) - P(y_{a_0}|x, a). \quad (2)$$

Kusner et al. [2017] declares that an algorithm is fair if $\text{CE}_{a_0, a_1}(y|x, a) = 0$ for all y and for any value $a_0 \neq a_1$. In the example presented in Fig. 1, this means that the distribution of the algorithm’s recidivism score prediction would not change had the race(A) been altered, given the conditions where all other observed variables remain the same.

Zhang and Bareinboim [2018] suggest a set of more granular counterfactual measurements to explain observed statistical disparity in the outcome Y over the protected attribute A over the underlying causal pathways between them, including direct, indirect, and spurious paths.

Definition 2.2 (Direct, Indirect, Spurious Effects). Given a SCM $\mathcal{M}$, a counterfactual direct effect (DE), indirect effect (IE), and spurious effect (SE) of an intervention $A = a_1$ on Y , with baseline $A = a_0$, conditioned on $A = a$ are

$$DE_{a_0, a_1}(y|a) = P(y_{a_1, W_{a_0}}|a) - P(y_{a_0}|a) \quad (3)$$

$$IE_{a_0, a_1}(y|a) = P(y_{a_0, W_{a_1}}|a) - P(y_{a_0}|a) \quad (4)$$

$$SE_{a_0, a_1}(y) = P(y_{a_0}|a_1) - P(y_{a_0}|a_0) \quad (5)$$

Among quantities in the above equations, $Y_{a, W_{a'}}$ is a *nested counterfactual variable* such that given a unit $U = u$, the potential outcome $Y_{a, w}(u)$ is equal to the solution when the input w is set as the values of mediator $W_{a'}(u)$ under intervention $do(a')$. Hereinafter, we will consistently refer to distributions over counterfactual variables of the form $Y_{a, W_{a'}}$ as nested counterfactual measures. Using nested counterfactuals allows us to decompose an intervention's effect, i.e., race change, on the recidivism score into three distinct components [Zhang and Bareinboim, 2018]. First, IE measures the effect caused solely by changes in prior records (W) induced by the intervention, achieved by altering race as an input for the function f_W . Next, DE represents the impact of an intervention directly influencing the recidivism score, removing the effect from any changes in race that may have affected prior counts. Lastly, SE captures the effect mediated by changes in confounding variables, such as birthplace (Z). Similarly, Chiappa [2019] utilized Path-Specific Effects (PSEs) [Pearl, 2001] to measure the impact of potentially discriminatory mechanisms along causal paths (i.e., one-directional paths) from the protected attribute A to outcome Y .

The intuitive nature of counterfactual notion has inspired many other fairness definitions, counterfactual predictive parity [Coston et al., 2020] and counterfactual equalized odds [Coston et al., 2020, Mishler et al., 2021], to name a few. This paper will primarily focus on the counterfactual measures described in Defs. 2.1 and 2.2, but our proposed algorithm applies to any counterfactual probabilities.

3 Estimating Counterfactual Fairness Measures

As research progresses in defining more sophisticated counterfactual fairness measures, methods for identifying these measures have also been proposed. Nabi and Shpitser [2018] and Wu et al. [2019] have introduced algorithms for measuring path-specific counterfactual effects, albeit under the assumption of a linear model. On the other hand, Zhang et al. [2017] avoids assuming linearity but is still limited to identifiable cases only. In this section, we present an algorithm that improves the current state-of-the-art approaches by eliminating the linearity assumption and expanding the scope to include unidentifiable cases by adopting a Bayesian sampling approach.

The general procedure of our proposed approach is provided in Algorithm 1, which is designed to estimate a given counterfactual fairness measure μ based on observational data $\mathcal{D}$. It utilizes two sub-algorithms: (1) a causal discovery algorithm, called Fast Causal Inference (FCI) [Spirtes, 2001], which learns causal relationships from the data; and (2) a partial identification algorithm, SampleCTF, which samples the posterior counterfactual measure conditioning on the observational data and inferred causal relationships.

More specifically, Step 1 applies FCI to identify an equivalence class $\mathcal{E}$ of candidate causal diagrams compatible with the observational data. This is a standard approach required in analyses based on structural causal models. Any causal discovery algorithm can replace FCI in this step, depending on the user's assumptions about the data and structure, such as noise distributions or parametric assumptions. In Step 2, the algorithm refines the learned equivalence class by filtering out candidate causal diagrams violating the domain knowledge. In Steps 3-5, the algorithm enumerates through each candidate causal diagram $\mathcal{G}_i$ and obtains posterior samples of the target counterfactual fairness measure conditioning on the data $\mathcal{D}$ and causal knowledge $\mathcal{G}_i$. We will discuss details of the sampling algorithm, SampleCTF, later in this section. Finally, the algorithm sorts posterior samples and returns an interval at the $(1 - \delta)\%$ confidence level for a fixed error rate $\delta \in [0, 1)$. The learner could decide the error rate δ based on the goal of the analysis. When $\delta = 0$, the return interval converges to the optimal bound guaranteed to contain the target counterfactual measure when the number of observed samples increases [Manski, 1990, Chickering and Pearl, 1996, Zhang et al., 2022].

3.1 Partial Counterfactual Identification

In this section, we introduce a subroutine to draw posterior samples of unknown counterfactual probabilities conditioning on the observed data and causal knowledge of the environment. Details of this subroutine, SampleCTF, are described in Algorithm 2. It is a Bayesian sampling algorithm that takes three elements as inputs: dataset containing T samples of $\mathbf{V}$, $\mathcal{D} = \{(\mathbf{V}^t) : t = 1, \dots, T\}$, where $\mathbf{V}$ is a set of observed variables, a causal graphical model $\mathcal{G}$, and a target fairness measure μ . The causal graphical model $\mathcal{G}$ provides quantitative information about a set of unobserved variables $\mathbf{U}$ and the structure of f_V , a function determining an observed variable $v \in V$. Although the specifics of f_V are not specified, the graph $\mathcal{G}$ specifies which variables are eligible to be input for that function. The fairness measure μ is defined by the

Algorithm 1: IDENTIFYING FAIRNESS MEASURE: IDFair($\mathbf{D}, \mu, \delta$)

Input: $\mathbf{D} = \{(V^t) : t = 1, \dots, T\}$, fairness measure μ , and error rate δ
Output: A bound containing the fairness measure $\mu \in [a, b]$

- 1: Learn an equivalence class $\mathcal{E} = \text{FCI}(\mathbf{D})$ which is a finite set of causal diagrams $\mathcal{E} = \{\mathcal{G}_1, \dots, \mathcal{G}_n\}$ compatible with the observational data $\mathbf{D}$
- 2: Construct a subset of causal diagram $\mathcal{E}^* = \{\mathcal{G}_1, \dots, \mathcal{G}_m\}$ from the equivalence class $\mathcal{E}$ using the domain knowledge.
- 3: **for** $\mathcal{G}_i \in \mathcal{E}^*$ **do**
- 4: $\mu_i = \text{SampleCTF}(\mathbf{D}, \mathcal{G}_i, \mu, N)$ # N samples of fairness measure
- 5: **end for**
- 6: $\Gamma = [\mu_{1,1}, \dots, \mu_{1,N}, \dots, \mu_{n,1}, \dots, \mu_{n,N}]$ #concatenated $n \times N$ samples
- 7: $\Gamma = \text{sort}(\Gamma)$ #sorted in an increasing order
- 8: **return** mean $\frac{\sum_{j=1}^{n \times N} \Gamma_j}{n \times N}$, $(1 - \delta)\%$ confidence interval $(\Gamma_{\lfloor \delta/2 \times n \times N \rfloor}, \Gamma_{\lceil (1-\delta/2) \times n \times N \rceil})$

Algorithm 2: SAMPLING COUNTERFACTUAL PROBABILITIES: SampleCTF($\mathbf{D}, \mathcal{G}, \mu$)

Input: $\mathbf{D} = \{(V^t) : t = 1, \dots, T\}$, $\mathcal{G}, \mu, N; \alpha, M, K$

- 1: Initialize $\mathbf{q}^0 = \frac{1}{K} \cdot (1, \dots, 1)$ to be a uniform distribution with K cases
- 2: Initialize f_V as a random function for all $V \in \mathbf{V}$.
- 3: **for** $t = 1, \dots, T$ **do**
- 4: Randomly assign U^t
- 5: Update f_V to match V^t and U^t for all $V \in \mathbf{V}$
- 6: **end for** { # Initialization done, sampling starts }
- 7: **for** $i = 1, \dots, M + N$ **do**
- 8: **for** $t = 1, \dots, T$ **do**
- 9: Compute $P(U|V = V^t; \mathbf{q}^{i-1}) \propto P(V^t|U)\mathbf{q}^{i-1}$.
- 10: Draw $U^t \sim P(U|V = V^t; \mathbf{q}^{i-1})$.
- 11: **end for**
- 12: Initialize Dirichlet Prior $\theta = \alpha \cdot (1, \dots, 1)$ for $\mathbf{q}$
- 13: Initialize f_V as a random function for all $V \in \mathbf{V}$
- 14: **for** $t = 1, \dots, T$ **do**
- 15: Update f_V to match V^t and U^t for all $V \in \mathbf{V}$
- 16: Update θ as the sum of the occurrence of U
- 17: **end for**
- 18: Draw $\mathbf{q}^i \sim \text{Dirichlet}(\theta)$
- 19: **if** $i > M$ **then**
- 20: $\mu_i = \mu(\mathbf{q}^i; f_V)$
- 21: **end if**
- 22: **end for**
- 23: **return** Sampled counterfactual measure $[\mu_{M+1}, \dots, \mu_{M+N}]$

probability over unobserved U , and Algorithm 2 is to estimate the exogenous distribution $P(U)$ and how it is pushed down to each element through functions f_V . These components serve as the foundational input for the algorithm.

In addition to the aforementioned core inputs, several parameters need to be specified. For simplicity and without loss of generality, we assume the existence of only one unobserved variable U with a finite cardinality of K in this section. In practical scenarios, there may be multiple unobserved variables $U_1, U_2, \dots$, each with corresponding cardinalities $K_1, K_2, \dots$ (or one can think of U as a vector). Exogenous probabilities $P(u)$ are a probability vector drawn from a Dirichlet distribution, i.e., $P(u) \sim \text{Dirichlet}(\alpha, \dots, \alpha)$, where $\alpha > 0$ is a small constant chosen for the Dirichlet prior.

A natural question arising at this point is how to determine the cardinality K for the domain of any exogenous variable $U \in \mathbf{U}$. To answer this question, we first introduce some necessary concepts in causal inference. We will utilize a special type of clustering of endogenous variables in the causal diagram, which is called *confounded components* [Tian and Pearl, 2002]. For convenience, let a *bi-directed arrow* $V_i \leftrightarrow V_j$ between endogenous nodes $V_i, V_j \in \mathbf{V}$ be defined as a sequence $V_i \leftarrow U_k \rightarrow V_j$ where $U_k \in \mathbf{U}$ is an exogenous parent shared by V_i, V_j . A *bi-directed path* is a consecutive sequence of bi-directed arrows, —i.e., a path composed entirely of bi-directed arrows. Formally,

Definition 3.1 (C-Component [Tian and Pearl, 2002]). For a causal diagram $\mathcal{G}$, a subset of variables $C \subseteq V$ is a c-component if any pair of nodes $V_i, V_j \in C$ is connected by a bi-directed path.

For an arbitrary exogenous variable $U \in \mathcal{U}$, we denote by $C(U)$ the c-component covering U in $\mathcal{G}$, i.e., $U \in \bigcup_{V \in C(U)} U_V$. For instance, every node in Fig. 1a is a c-component due to the lack of bidirected arrows. On the other hand, exogenous variables U_1, U_2 in Fig. 1b are covered by a single c-component $C(U_1) = C(U_2) = \{A, Z, W, Y\}$ due to the bi-directed path $A \leftrightarrow Z \leftrightarrow Y \leftrightarrow W$.

To compute the cardinality K for unobserved variables U , the C-component of U is first identified, and the number of states required to represent all states of the observed variables is counted. More specifically, we will model every unobserved exogenous variable $U_i \in \mathcal{U}$ as a discrete variable taking values in a finite domain $\{1, \dots, K_i\}$. For every $U_i \in \mathcal{U}$, the cardinality K_i of the exogenous domain of U_i is bounded by

$$K_i = d_i + 1, \quad \text{where } d_i = \Pi_{V \in \text{Pa}(C(U_i))} |V|. \quad (6)$$

In the above equation, $C(U_i)$ is the c-component covering U_i in the causal diagram of the model $\mathcal{M}$; $\text{Pa}(C(U_i))$ are observed direct parents of nodes in the c-component $C(U_i)$ (including $C(U_i)$). For example, suppose our goal is to infer the counterfactual direct effect $\text{DE}_{a_0, a_1}(y|a)$ from the observational distribution $P(A, Z, W, Y)$ in Fig. 1b; A, Z, W, Y are binary variables taking values in $\{0, 1\}$. Since U_1, U_2 share the same c-component $\{A, Z, W, Y\}$, SampleCTF will set their cardinality K_1, K_2 are set as $K_1 = K_2 = 2^4 + 1 = 17$.

Since we do not assume any specific forms of the exogenous domains and the exogenous variables U could take any values, one may wonder whether bounding the exogenous cardinality following Eq. 6 could impose additional restrictions, rendering the inferred posterior counterfactual probabilities invalid. We will next show this is not the case. First, Proposition 2.6 in Zhang et al. [2022] for any structural causal model $\mathcal{M}$ with discrete observed variables V , one could generate its observational distribution $P(V)$ using discrete latent variables U , where for every $U_i \in \mathcal{U}$, its cardinality is bounded by d_i defined in Eq. 6. We could extend this discretization result to represent the observational data $P(V)$ and the target counterfactual measure μ simultaneously. The additional +1 state in Eq. 6 allows us to represent the probability mass associated with the target counterfactual measurement μ . The following proposition ensures that this exogenous cardinality bound is sufficient in representing the observed data $P(V)$ and the target fairness measure μ , without violating qualitative knowledge encoded in graph $\mathcal{G}$.

Theorem 3.2. For an SCM $\mathcal{M}$, let $\mathcal{G}$ be its associated causal graph, $P(V)$ be its observational distribution, and μ be a nested counterfactual measure. Then there exists an alternative SCM $\mathcal{N}$ with exogenous cardinalities bounded in Eq. 6 such that $\mathcal{M}$ and $\mathcal{N}$ induce the same $\mathcal{G}$, $P(V)$ and μ .

Proof. Proposition 2.6 in Zhang et al. [2022] for any structural causal model $\mathcal{M}$ with discrete observed variables V , one could generate its observational distribution $P(V)$ using discrete latent variables U , where for every $U_i \in \mathcal{U}$, its cardinality is bounded by d_i defined in Eq. 6. The additional +1 state in Eq. 6 allows one to represent the probability mass associated with the target counterfactual measure μ . The proof follows a similar procedure for [Zhang et al., 2022, Lemma A.6]. $\square$

After initializing exogenous cardinalities, Algorithm 2 begins by initializing the distribution of unobserved variables $P_0(U)$ and the functions f_V (see Algorithm 2, Steps 1 to 6). It chooses a uniform distribution over U , i.e., $P_0(U) = \mathbf{q}^0 = \frac{1}{K} \cdot (1, \dots, 1)$ (Step 1). The initialization of f_V involves randomly defining the function (Step 2) and drawing U^t based on the uniform distribution $P_0(U)$ (Step 4). The structural function f_V is then updated to match V^t and U^t by iterating over the data one by one. The order of the samples can be randomized if desired, as the goal is to have random functions that align well with the data.

In each round i of the sampling procedure (Algorithm 2, Steps 8 to 23), new U^t is drawn from $P(U|V = V^t; \mathbf{q}^{i-1})$, and f_V is re-initialized accordingly. While we update the f_V , a Dirichlet prior θ is initialized to $\theta = \alpha \cdot (1, \dots, 1)$ and updated to obtain a Dirichlet posterior based on the occurrence of U . Finally, an updated distribution $\mathbf{q}^i$ over U is drawn from the Dirichlet posterior: $\mathbf{q}^i \sim \text{Dirichlet}(\theta)$. This process is iterated for M rounds to converge to a stable period, and the last N rounds are used for generating counterfactual probabilities in Step 20, computed using $\mathbf{q}^i$ and f_V . In summary, Algorithm 2 samples counterfactual probabilities by exploring the entire function space while updating the conditional probability of unobserved variables given observations.

Theorem 3.3. Given priors ρ over the exogenous probabilities $P(U)$ and structural functions $\mathcal{F}$, SampleCTF($\mathcal{D}, \mathcal{G}, \mu$) draws a posterior sample of the target counterfactual measure μ conditioning on the observed data $\mathcal{D}$, i.e., $\mu \sim P(\mu | \mathcal{D}; \mathcal{G}, \rho)$.

Proof. It follows from Theorem 3.2 that discrete exogenous domains with cardinality bounds in Eq. 6 are sufficient in representing the observational distribution $P(V)$ and an arbitrary nested counterfactual measure μ in a casual graphical

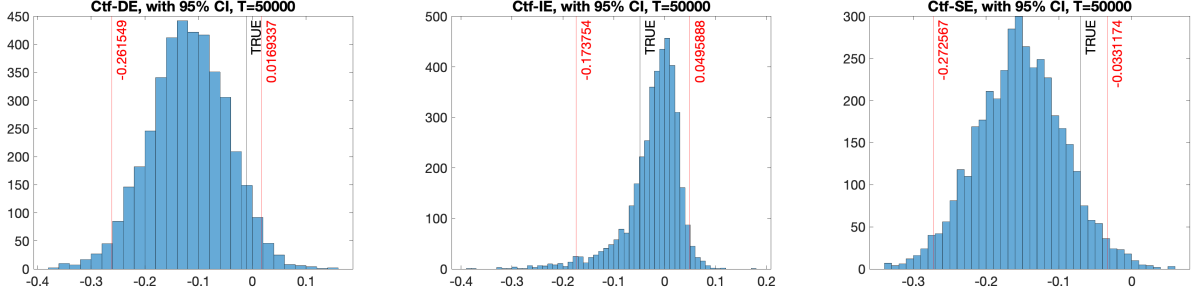


Figure 2: Histograms for DE, IE, and SE, obtained from the simulation dataset. The black vertical line is the ground-truth value (labeled as TRUE) and the two red lines show 95% confidence interval (2.5% top, 2.5% bottom).

model $\mathcal{G}$. Given priors ρ of the exogenous probabilities $P(u)$ and structural functions $\mathcal{F}$, the statement follows from the validity of Gibbs sampling [Geman and Geman, 1984]. $\square$

Choosing the appropriate burn-in period M involves recognizing that Algorithm 2 operates as a special type of Gibbs sampling method. Consequently, we anticipate the sampled $P(U)$ to converge to a certain distribution. As the true $P(U)$ is inaccessible, we suggest computing a probability obtainable from the observed distribution and verifying that the sampled quantity indeed converges to the observed quantity. For instance, in Sec. 5 that follows, we calculate $P(Y = 1 \mid A = 0)$ from the sampled $P(U)$ and the functions, comparing it against the same quantity computed from the observed data (assumed to be the ground truth).

4 Simulation study

In this section, we evaluate the accuracy of our method using a simulation dataset. The data was generated based on the graphical model in Fig. 1a, with a random distribution for each unobserved variable. The final dataset consists of 50,000 observations for the binary variables Z , A , W , and Y .

By applying the FCI algorithm [Spirtes, 2001], one could infer that the covariate Z and the mediator W are not adjacent. That is, there is no direct $Z \rightarrow W$, $Z \leftarrow W$, or bidirected arrow $Z \leftrightarrow W$ in the underlying causal diagram. We also assume access to the domain knowledge that variables A , Z , W could be potential direct causes for the outcome Y , but not vice versa. Synthesizing the domain knowledge with the causal relationships inferred from the observational data leads to a causal diagram $\mathcal{G}$ described in Fig. 1b. Compared with the ground-truth graph in Fig. 1a, our learned model $\mathcal{G}$ has additional latent variables U_1, U_2 affecting variable clusters $\{A, Z, Y\}$ and $\{A, W, Y\}$ respectively, since the combination of the observational data and domain knowledge is unable to rule out the presence of unobserved confounding.

With this $\mathcal{G}$, the cardinality of unobserved variables U_1 and U_2 should be greater than or equal to 17. We set $K = 22$ and apply Algorithm 1 to obtain bounds for the direct effect (DE), indirect effect (IE), and total effect (TE) of changing A . Fig. 2 shows the histograms of the DE, IE, and SE drawn by Algorithm 2. The results consistently contained the ground truth within the $(1 - \delta)\%$ confidence interval for all three quantities with $\delta = 0.05$, corroborating Theorems 3.2 and 3.3. More detailed results are shown in Appendix A.1.

5 COMPAS Case study

In this section, we demonstrate Algorithm 1, IDFair, on a real-world dataset. We describe the COMPAS dataset in Section 5.1, identify its causal graphical structure in Section 5.2, and share the results of the sampling DE, IE, and SE in Section 5.3. Then, we instantiate the causal explanation formula from Zhang and Bareinboim [2018] and show how our results aligns with it in Section 5.3.4. Finally, we use our algorithm to estimate other counterfactual fairness measures and compare them against DE, IE, and SE in Section 5.4.

5.1 COMPAS Dataset

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is an algorithm developed by Northpointe (now Equivant) used in the judicial process to predict a criminal defendant’s recidivism score. In 2016,

ProPublica raised a concern that black defendants were often predicted to be at a higher risk of recidivism than they actually were (45% vs. 23%), whereas white defendants were often predicted to be less risky than they were (48% vs. 28%) based on a 2-year follow-up data [Angwin et al., 2016]. ProPublica publicized their dataset on Github¹, which contains COMPAS scores for *Risk of Recidivism* ranging from 1 to 10, as well as each defendant’s race, sex, age, criminal history (prior counts), and charge degree.

In our analysis, we assume $A = \{\text{Race, Sex, Age}\}$ be protected attributes that shall not affect the recidivism score directly, $X = \{\text{Charge degree, Prior counts}\}$ be measurements that could be related to an individual’s actual recidivism (X could be either confounder Z or mediator W), and $Y = \{\text{Risk of Recidivism}\}$ be the outcome of the algorithm to be assessed if it is fair or not. Table 1 shows the definition of our variables.

Table 1: Definition of variables

Variable	0	1
Race(A)	Others	African-American
Age(A)	Less than 30	Over 30
Sex(A)	Female	Male
Charge Degree (W_1)	Misdemeanor	Felony
Prior Counts (W_2)	≤ 2	> 2
Score (Y , 1 ~ 10)	≤ 5	> 5

5.2 Causal Graphical Structure of COMPAS

This section illustrates Steps 1-2 of Algorithm 1, where the graphical structure $\mathcal{G}$ is learned using the FCI algorithm, implemented in an R package `pcalg` [Kalisch et al., 2012]. First, we come up with an equivalent class of candidate causal diagrams using FCI algorithm, then use qualitative domain knowledge to fine-tune the result and filter out incompatible diagrams. This approach allows non-Markovian cases where unobserved confounders generally exist. We assume that there exist no direct edges from X (Charge degree, Prior counts) to the protected attributes A (Age, Sex, or Race).

We show in Fig. 8 in Appendix A.2, the equivalent class inferred by the FCI algorithm. Based on this, we leverage our background knowledge and make informed decisions about edge types. First, we assume that there cannot be a direct edge between the protected attributes, i.e., age cannot be influenced by sex. Next, considering that the determination of protected attributes precedes the establishment of charge degree or prior counts, we preclude any directed edges from charge degree or prior counts to the protected attributes. Similarly, since prior counts are already decided before the charge degree, we prohibit directed edges from charge degree to prior counts. Finally, we permit bi-directed edges where such a possibility exists. The finalized graphical model assumed throughout the analysis is shown in Fig. 3.

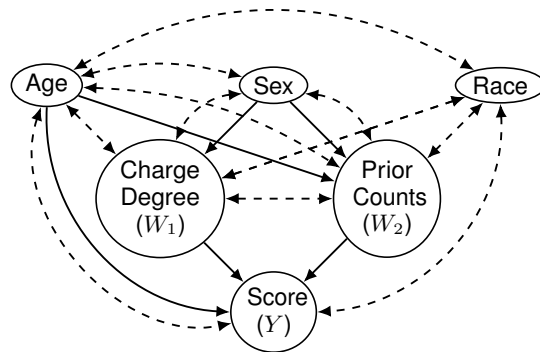


Figure 3: Inferred causal diagram for the COMPAS dataset using the FCI algorithm and domain knowledge.

From this composite graphical model, we can obtain a simplified version for each protected attribute. Fig. 4 shows graphical models for (a) $A = \{\text{Race}\}$, (b) $A = \{\text{Age}\}$, and (c) $A = \{\text{Sex}\}$. In this case, we are only considering one graph for each case, i.e., the equivalence class $\mathcal{E}^*$ in Step 2 of Algorithm 1 has only one candidate.

¹<https://github.com/propublica/compas-analysis/>

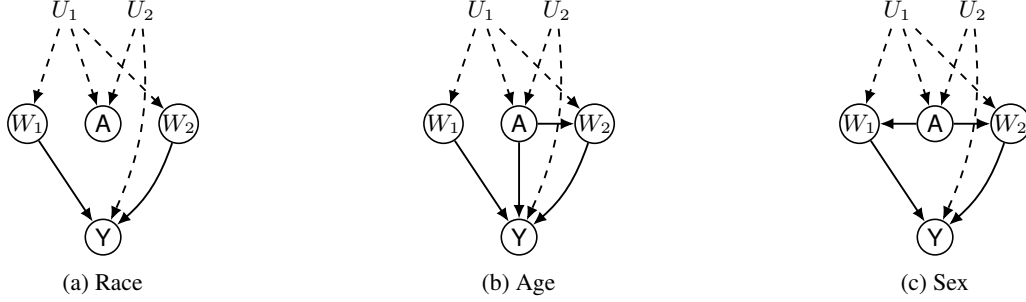


Figure 4: Graphical model for each protected variable with two exogenous variables. Each exogenous variables have 17 states. A denotes (a) race, (b) age, (c) sex, W_1 denotes charge degree, W_2 denotes prior counts, and Y denotes the predicted COMPAS score.

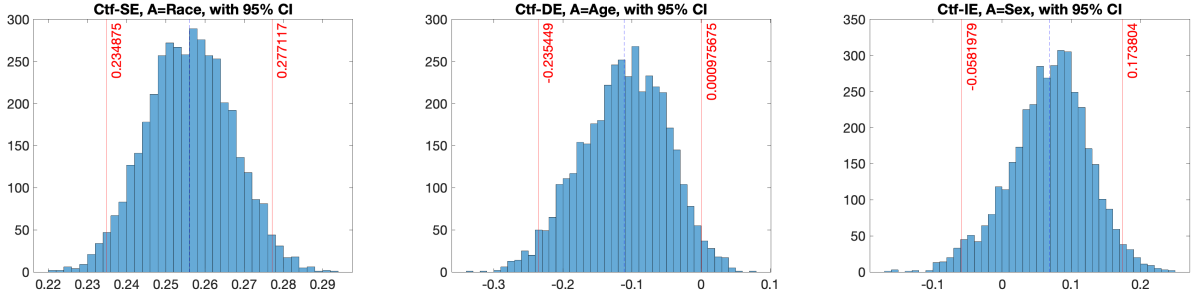


Figure 5: Histogram of SE when $A = \text{Race}$ (left), DE when $A = \text{Age}$ (middle), IE when $A = \text{Sex}$ (right). The two red lines show 95% confidence interval (2.5% top, 2.5% bottom).

Finally, we obtain the minimum number of states K for each exogenous variable U as described in Section 3.1. Given that all endogenous variables are binary, it follows that for exogenous variables U_1 and U_2 , we require a minimum of $K_1 = K_2 = 2^4 + 1 = 17$, $K_1 = K_2 = 2^5 + 1 = 33$, or $K_1 = K_2 = 2^6 + 1 = 65$ states when $A = \text{Race}$, $A = \text{Age}$, or $A = \text{Sex}$, respectively. For the experiments, both K_1 and K_2 were set to have the same value of 20, 40, and 70 for $A = \text{Race}$, $A = \text{Age}$, and $A = \text{Sex}$, respectively.

5.3 Estimated Counterfactual DE, IE, and SE

In this section, we show the results of running Algorithm 2 on COMPAS dataset to obtain bounds for counterfactual fairness measures. We start with a uniform initialization for $P(U_1)$ and $P(U_2)$, using respective K_1 and K_2 . The burn-in period M was set to 2000, and $N = 4000$ samples were collected after the burn-in for all experiments. Throughout all experiments, we consistently set the error rate $\delta = 0.05$. However, as previously mentioned, one could specify δ to be any reasonable real value in $[0, 1)$ based on the analysis. We highlight main findings here; more visualizations and discussion are deferred to Appendix A.3.

5.3.1 (a) $A = \text{Race}$

When we analyze the counterfactual effects of Race, the direct and indirect effects were zero by the graphical structure—there is no direct or indirect path from A to Y . The first histogram in Fig. 5 shows the distribution of 4000 samples' SE we obtained from Gibbs sampling. The width of the 95% confidence interval is 0.0423 (about 4%p). In words, this means that if a non-African-American individual had been African-American, the probability that the individual is assigned to a high score (greater than 5) would have become at least about 23% higher and at most about 27% higher, solely through the spurious paths (backdoor paths).

5.3.2 (b) $A = \text{Age}$

The second histogram in Fig. 5 shows the confidence interval for DE, $(-0.2354, 0.0009)$, when $A = \text{Age}$. The CI is wide but still indicates some portion of negative effect, i.e., if you were older, you are less likely to be assigned high risk. The confidence interval for IE is centered around zero and ranges from -10% to 10% , hence it is hard to draw any

Table 2: 95% Confidence interval for counterfactual family of effects and Total variation of each protected attribute (A).

A	$DE_{A=1,A=0} A=1$	$IE_{A=0,A=1} A=1$	$SE_{A=0,A=1}$	TV Bound	TV
Race	(0, 0)	(0, 0)	(0.2348, 0.2771)	(0.2348, 0.2771)	0.2544
Age	(0.0333, 0.4511)	(-0.1113, -0.0352)	(-0.2609, 0.3164)	(-0.3399, -0.1413)	-0.2018
Sex	(0, 0)	(-0.0196, 0.1656)	(-0.1284, 0.1521)	(0.0066, 0.0872)	0.0629

conclusion about the indirect effect. The confidence interval of SE is 41.68% p and this bound is too wide to draw any causal conclusion.

5.3.3 (c) Sex

When $A = \text{Sex}$, similar to the Race case, the DE is zero by definition (no direct path from A to Y). The last histogram in Fig. 5 shows a wide confidence interval for IE (23.1% p , CI=(-0.0581, 0.1738)). This could stem from the inherent complexity of the graphical model, as evidenced by the wider spectrum shown in the convergence graph (Fig. 12 in Appendix A.3).

5.3.4 Comparison to Total Variation (TV)

Lastly, we demonstrate Theorem 1 of Zhang and Bareinboim [2018] with the results we obtained, which shows total variation as a linear combination of DE, IE, and SE. We take the first version of the theorem with $a_0 = 0$ and $a_1 = 1$:

$$TV_{A=0,A=1}(Y=1) = SE_{A=0,A=1}(Y=1) + IE_{A=0,A=1}(Y=1|A=1) - DE_{A=1,A=0}(Y=1|A=1).$$

Table 2 shows the values for each term in the above equation. The second last column (TV Bound) indicates the bound for the $TV_{A=0,A=1}(Y=1)$, and the total variation (the last column) falls within the TV bound for all three cases.

5.4 Comparison to Other Fairness Measures

This section compares the quantitative and qualitative findings using various fairness measures with our results. First, we use our algorithms to estimate other *counterfactual* measures, such as the Counterfactual Effect (CE) based on the counterfactual fairness definition by Kusner et al. [2017] and the Path-specific Counterfactual Effect (PSE) by Chiappa [2019], and compare with our findings. Then, we qualitatively compare our results to previous findings with different statistical fairness measures.

5.4.1 Counterfactual Effect (CE)

When $A = \text{Race}$, as in Fig. 4a, the CE is zero by the graphical structure—there is no direct or indirect path from A to Y . This means that the COMPAS algorithm abides by the counterfactual fairness by Kusner et al. [2017] with respect to race.

With $A = \text{Age}$, we calculate the effect of changing $A = 0$ (baseline, age less than 30) to $A = 1$ (intervention, age over 30) conditioned on $A = 0$, W_1 , and W_2 . Since the sampling did not converge when $W_2 = 0$, we present CE conditioned on $W_2 = 1$ only. The 95% confidence interval for CE conditioned on $A = 0$, W_1 and $W_2 = 1$ were estimated to be (0.0290, 0.3665) when $W_1 = 0$ and (0.0120, 0.2544) when $W_1 = 1$ (Fig. 6, left and middle). The confidence interval lies on the positive side, regardless of the conditioning value for W_1 . This means that changing age from less than 30 to over 30 has a positive counterfactual effect on the COMPAS score, although the precise magnitude is challenging to ascertain due to the wide confidence interval. Although these results do not necessarily contradict the results from Section 5.3.2, this definition fails to capture the negative DE we obtained.

With $A = \text{Sex}$, we compute the effect of changing $A = 0$ (baseline, female) to $A = 1$ (intervention, male) conditioned on $A = 0$, W_1 , and W_2 . The counterfactual effect (CE) bound exhibits a relatively positive trend when $W_2 = 0$ (fewer prior charges; bounds are (0.0794, 0.4737) when $W_1 = 0$, (-0.0626, 0.3098) when $W_1 = 1$) and a negative trend when $W_2 = 1$ (more prior charges; bounds are (-0.3556, 0.0658) when $W_1 = 0$, (-0.4081, -0.0200) when $W_1 = 1$)—see Appendix A.4, Fig. 15. This indicates that a female is more likely to receive a higher COMPAS score if she were male, particularly when she has fewer prior charges. However, this is reversed when she has more than two prior charges; she is likely to receive a lower COMPAS score had she been male if $W_2 = 1$.

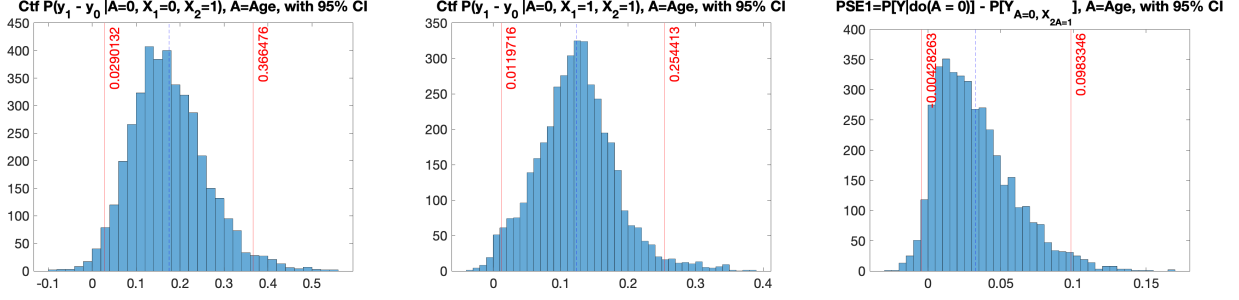


Figure 6: Histograms of other counterfactual fairness measures for $A = \text{Age}$. Each histogram shows CE conditioned on $A = 0, W_1 = 0, W_2 = 1$ (left), CE conditioned on $A = 0, W_1 = 1, W_2 = 1$ (middle), PSE for the path $A \rightarrow Y$ (right). The two red lines show 95% confidence interval (2.5% top, 2.5% bottom). More graphs from Section 5.4 are deferred to Appendix A.4

5.4.2 Path-Specific Effect (PSE) with respect to Age

For $A = \text{Age}$, one can define the direct edge from A to Y to be unfair and the indirect path from A to W_2 to Y to be fair. In this case, The PSE is identical to the unconditioned DE , and the obtained bound $(-0.0043, 0.0983)$ mostly lies on the positive side (Fig. 6, right).

Another way to define an *unfair* path is to choose the path from A to W_2 . We can estimate this PSE by setting $A = 0$ as the input for f_Y and $A = 1$ as the input for f_{W_2} . The obtained 95% confidence interval $(0.0513, 0.2426)$ indicates that the direction of the PSE is positive when we change from $A = 0$ to $A = 1$ (Appendix A.4, right side of Fig. 16). This aligns with Kusner’s CE estimation, although the magnitude of the effect is different.

5.4.3 Other (Non-counterfactual) Fairness Measures

The first accusation against COMPAS scores was that the false positive rate was higher for black defendants than for white defendants, and the false negative rate was higher for white defendants than for black defendants [Angwin et al., 2016]. While we cannot directly compare our findings to analyses of false positive or false negative rates—since the notion of counterfactual fairness does not necessarily account for the actual outcome—we do find evidence of racial bias in COMPAS scores through a spurious path. This spurious effect may explain the differences in false positive and false negative rates between races, favoring white defendants.

It was also suggested that the COMPAS score assigns higher score to women than men “[b]ecause women reoffend at lower rates than men with similar criminal histories” [Corbett-Davies et al., 2023]. Based on our analyses, the direct path from $A = \text{Sex}$ to the COMPAS score (Y) was absent. If the aforementioned claim were to be true, having a direct path connecting A (Sex) and Y (COMPAS Score) could help mitigate the currently observed unfair behavior against women.

6 Discussion

Interpretation of Estimated Bounds The confidence interval output by Algorithm 1 is obtained by testing any possible distribution $P(U)$ over hidden variables and the function f_V determining observed variables against the observation data. For example, 95% confidence interval means that of all $P(U)$ and f_V that can generate the observation data, 95% of the cases will result in the (counterfactual) fairness measure that falls within the calculated confidence interval. Hence, the confidence level accounts for the randomness from the ignorance region (non-identifiable parts in the structure), as well as having a finite number of samples. If desired, our approach can be utilized to approximate the worst-case bound by taking the minimum and the maximum of the samples probabilities. In such cases, increasing the number of samples N may also help.

Choice of Fairness Definition In the realm of counterfactual fairness, a variety of proposed definitions often lack alignment with one another, leading to instances where certain definitions are deemed inadmissible with respect to others. A notable example is the CE, which is inadmissible with respect to Direct Effect (DE), Indirect Effect (IE), and Spurious Effect (SE) [Plecko and Bareinboim, 2022]. This implies that observing $CE = 0$ does not guarantee the non-zero DE, IE, or SE. This phenomenon was evident in the case of $A = \text{Race}$, where CE (or PSE) was inherently

zero, but a significant SE was estimated. Thus, the choice of fairness metric becomes crucial, as it can lead to divergent conclusions.

Our analysis of fairness concerning Race serves as an illustrative example. Among all the tested definitions, only SE was capable of capturing racial bias, exhibiting a relatively narrow confidence interval around 25%. Some may assert that the spurious effect is not *causal*, operating solely through a backdoor path. However, we propose that it should still be regarded as an unfair path, particularly when the testing variable Y represents an algorithmic output, and the variables in the spurious paths are not exhaustively known.

Consider the spurious path in the Race graph involving, for instance, $U_2 =$ the zip code of residence. If the outcome variable of interest Y were the actual observed recidivism in the real world, we might contend that the spurious path is not unfair due to its lack of causality. However, when Y represents the recidivism score generated by an algorithm, it introduces tools for the algorithm to *infer* zip code either through input or estimation, potentially leading to actively discriminating against African-American individuals.

Counterfactual fairness stands as a logically sound concept, yet the consensus on its precise definition remains elusive. In our paper, we opted for the counterfactual family of effects (DE, IE, and SE) proposed by Zhang and Bareinboim [2018], deeming it the most appropriate among available options. However, alternative ways to define counterfactual fairness may exist, and as we gain a deeper understanding of counterfactual probabilities and formulate refined definitions, our Algorithm 1 can be applied to accommodate the new definitions.

Analysis of COMPAS scores The qualitative analysis of COMPAS scores requires several caveats. First, the validity of our findings depends on the assumption that the underlying causal structural model is correct. The causal discovery process is influenced by various factors, including the choice of algorithm and the definition of variables; for example, adopting a different causal discovery algorithm could result in slightly different causal diagrams. While this does not invalidate our results, it emphasizes the need to interpret the findings with the understanding that the underlying causal model may change. To address this, Algorithm 1 is designed to account for multiple causal models, providing a more robust framework for analysis.

Another key caveat is that the interpretation of results is closely tied to the definitions of the variables used. In our analysis, race is defined as a binary variable: Black versus Others. As a result, the analysis specifically addresses counterfactual scenarios involving changes in race between Black and non-Black groups. This framing does not extend to other populations, such as the Hispanic demographic. For instance, if we had defined the Race variable as Hispanic versus Others, the results would have been entirely different. These limitations must be carefully considered when interpreting and generalizing the findings.

7 Conclusion

This paper addressed the non-identifiability issue associated with counterfactual probabilities by introducing a sampling algorithm (Algorithm 1) for their estimation. The simulation study in Section 4 affirmed the validity of our algorithm—all results encompass the ground-truth value within the 95% confidence interval. Then, we demonstrated our algorithm in Section 5 by evaluating algorithmic fairness of COMPAS recidivism scores with respect to race, age, and sex. The findings revealed a significant spurious effect (SE) of $25.5 \pm 2\%$ when the race changed to African-American from others. Additionally, a negative direct effect (DE) was identified when age transitioned from less than 30 to over 30. In the case of sex, the confidence interval was too wide and included zero, not yielding decisive conclusions. This lends credibility to our COMPAS analysis, affirming the reliability of our approach in efficiently assessing algorithmic fairness.

References

- Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.
- Shira Mitchell, Eric Potash, Solon Barocas, Alexander D’Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.
- Julius A Adebayo et al. *FairML: ToolBox for diagnosing bias in predictive modeling*. PhD thesis, Massachusetts Institute of Technology, 2016.

- Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.
- Jack Bandy. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction*, 5(CSCW1):1–34, 2021.
- Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.
- Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.
- Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.
- Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.
- Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.
- Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- Lu Zhang and Xintao Wu. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics*, 4:1–16, 2017.
- Solon Barocas and Andrew D Selbst. Big data’s disparate impact. *Calif. L. Rev.*, 104:671, 2016.
- Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.
- Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.
- Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.
- M.L. Petersen, S.E. Sinisi, and M.J. van der Laan. Estimation of direct causal effects. *Epidemiology*, 17(3):276–284, 2006.
- Drago Plecko and Elias Bareinboim. Causal fairness analysis, 2022.
- Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the twenty-eighth international joint conference on Artificial Intelligence*, 2019.
- Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3929–3935, 2017. doi: 10.24963/ijcai.2017/549. URL <https://doi.org/10.24963/ijcai.2017/549>.
- Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. ProPublica, 2016. URL <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Judea Pearl. *Causality*. Cambridge university press, 2009.
- J. Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI 2001)*, pages 411–420. Morgan Kaufmann, San Francisco, CA, 2001.
- Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 582–593, 2020.
- Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 386–400, 2021.
- Peter Spirtes. An anytime algorithm for causal inference. In *International Workshop on Artificial Intelligence and Statistics*, pages 278–285. PMLR, 2001.
- Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.
- David Maxwell Chickering and Judea Pearl. A clinician’s tool for analyzing non-compliance. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1269–1276, 1996.

- Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. In *International Conference on Machine Learning*, pages 26548–26558. PMLR, 2022.
- J. Tian and J. Pearl. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 567–573. AAAI Press/The MIT Press, Menlo Park, CA, 2002.
- Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.
- Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H Maathuis, and Peter Bühlmann. Causal inference using graphical models with the r package pcalg. *Journal of statistical software*, 47:1–26, 2012.
- Sam Corbett-Davies, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *The Journal of Machine Learning Research*, 24(1):14730–14846, 2023.
- Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

A Additional Results

In this appendix, we provide more visualizations to better understand our paper. Some of the representative results are already included in the main part, but we show them again in this section for completeness.

A.1 From Section 4. Simulation Study

In this section, we show more visualizations of the results from the simulation study.

In Fig. 7, the top left graph shows the convergence of $P[Y_{A=0} = 1|A = 0]$ computed from the samples. The X axis is time (1^{st} to 4000^{th} sample) and the horizontal red line indicates the ground truth value for $P[Y = 1|A = 0]$. For the three histograms, the black vertical line is the groundtruth value (labeled as TRUE) and the two red lines show 95% confidence interval (2.5% top, 2.5% bottom).

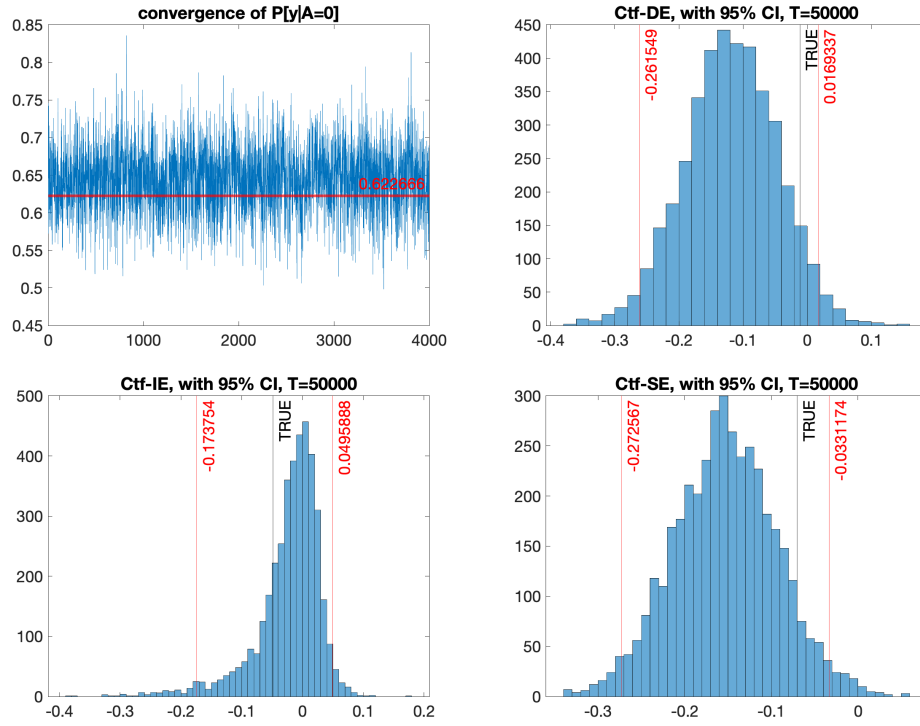
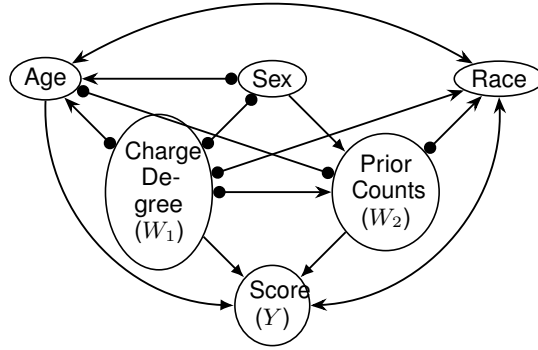


Figure 7: The convergence graph (top left), and histograms for DE (top right), ID (bottom left) and SE (bottom right), obtained from the simulation dataset.

A.2 From Section 5.2. Causal Graphical Structure of COMPAS

Figure 8 show the equivalence class identified from the COMPAS data using the FCI algorithm. FCI algorithm identifies a equivalence class $\mathcal{E}$ of causal diagrams in a compact form of a partial ancestral graph (PAG), where the circle edges represent undetermined edge marks Zhang [2008]. To translate the PAG into a DAG, we need to decide the mark for circle edges using qualitative knowledge such as time order among the variables. More details about how to interpret the equivalence class can be found in Section 3 of Zhang [2008].


 Figure 8: Identified equivalent class $\mathcal{E}$ for the COMPAS dataset

Note that categorical variables must be binary for causal discovery; thus, Race is defined as 1 for African-American and 0 otherwise. For Age, Prior Counts, and Score, we binarize the variables for computational efficiency, following the rules in Table 1. However, our sampling algorithm (Algorithm 2) can accommodate any finite discrete variables.

A.3 From Section 5.3. Estimated Counterfactual DE, IE, and SE

In this section, we show more visualizations of the results from the COMPAS Case Study (Section 5.3).

A.3.1 $A = \text{Race}$

The graph on the left side of Fig. 9 shows the time series of a counterfactual probability $P[Y_{A=0} = 1 | A = 0]$ computed from the samples as a blue line. The red line is the same value obtained from the observation (because $P[Y_{A=0} = 1 | A = 0] = P[Y = 1 | A = 0]$.) The values in blue line should converge to the red line, if the sampling method worked out. In this case, the counterfactual probability computed by the Algorithm 2 lies between around 0.22 and 0.25, and the observational distribution (around 0.2350) also falls within the obtained bound.

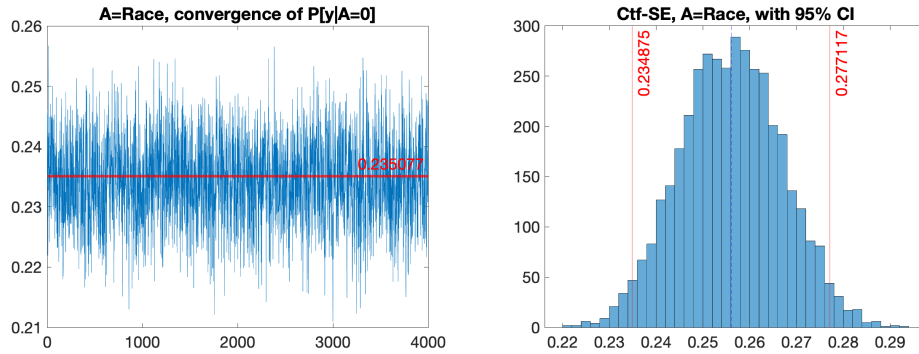


Figure 9: *Left*: the convergence of $P[Y_{A=0} = 1 | A = 0]$ computed from the samples. The X axis is time (from first to 4000-th sample) and the horizontal red line indicates $P[Y = 1 | A = 0]$ computed from the data. *Right*: SE of $A = \text{Race}$. The two red lines show 95% confidence interval (2.5% top, 2.5% bottom).

A.3.2 $A = \text{Age}$

Fig. 10 shows the calculated $P(y|A = 0)$, $A = \text{Age}$, over each round of draw (similar to the graph on the right side of Fig. 9). This is to test the convergence, where we expect the blue lines to converge to the red horizontal line (ground truth from observation). Indeed, the blue line and the red line overlaps well: blue lines falling mostly between 0.46 and 0.49, and the red line indicates 0.47. The variance in the blue line is slightly bigger than the Race case, but still within a reasonable range.

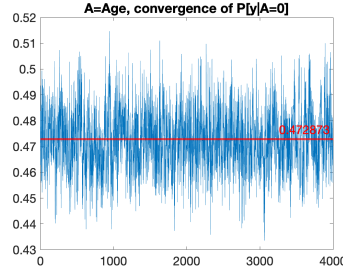


Figure 10: The convergence of $P[Y_{A=0} = 1|A = 0]$, when $A = \text{Age}$, computed from the samples. The X axis is time (1^{st} to 4000^{th} sample) and the horizontal red line indicates $P[Y = 1|A = 0]$ computed from the data.

Fig. 11 shows the histogram of all three counterfactual fairness quantities, when A is set to be Age. As discussed earlier, the width of the confidence interval of SE is 41.68% and this bound is much wider than the bound for SE of $A = \text{Race}$. From this confidence interval, it is hard to argue if there exists any discrimination, let alone the magnitude of direction of it.

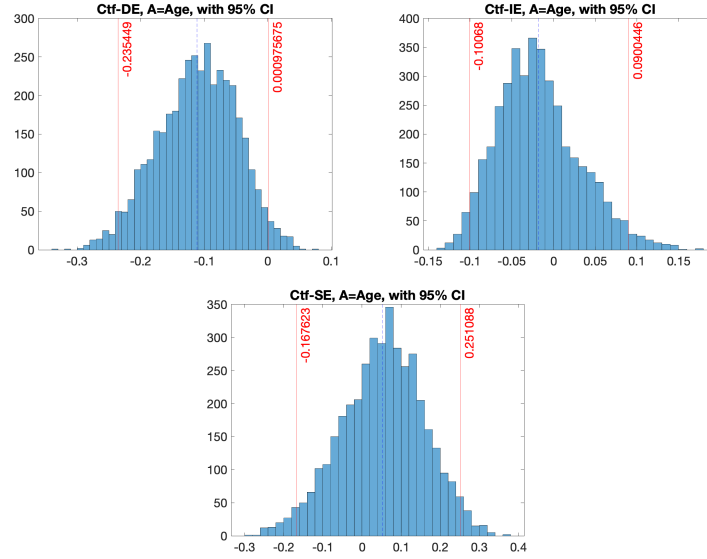


Figure 11: DE (top left), IE (top right), and SE (bottom) of $A = \text{Age}$. The blue line is the mean and the two red lines show 95% confidence interval (2.5% top, 2.5% bottom).

A.3.3 $A = \text{Sex}$

Fig. 12 shows the largest variance among all three A 's we tested. The blue line ranges mostly between 0.3 and 0.36, showing a higher variance than the two previous cases. The red line (0.3146) lies within that bound, but it is much closer to 0.3 than 0.36. This may imply that our Algorithm 2 found it harder to converge in the case of $A = \text{Sex}$.

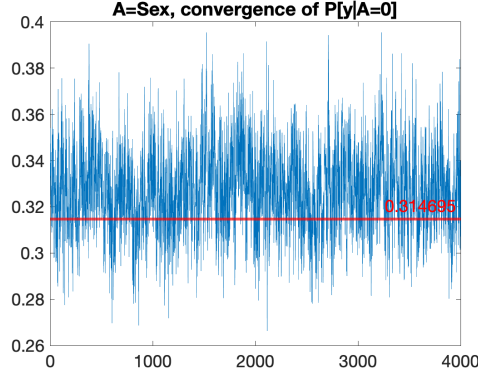


Figure 12: The convergence of $P[Y_{A=0} = 1 | A = 0]$ computed from the samples. The X axis is time (from first to 4000-th sample) and the horizontal red line indicates $P[Y = 1 | A = 0]$ computed from the data.

Fig. 13 shows the distribution of IE (left) and SE (right). Again, the confidence interval of IE is 23.1%p ($CI = (-0.0581, 0.1738)$), while that of SE is 28.05%p ($CI = (-0.1284, 0.1521)$). Although it is difficult to decide which direction the spurious effect is heading to, IE seems to be marginally negative or positive. These confidence intervals are much wider than the previous cases, which is already alluded by the convergence in Fig. 12.

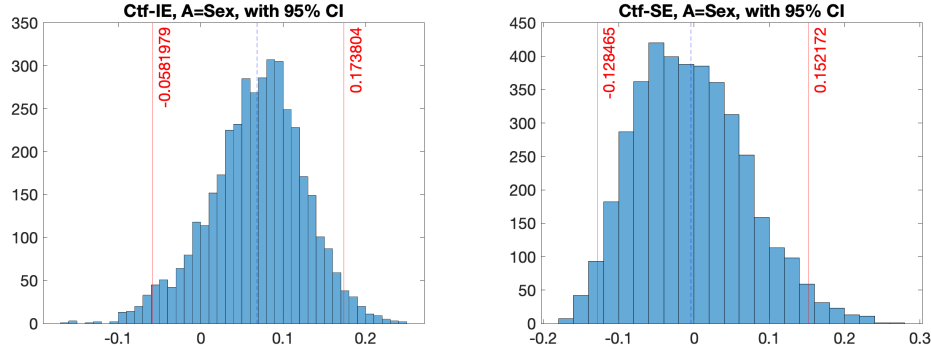


Figure 13: IE (left) and SE (right) of $A = \text{Sex}$. The blue line is the mean and the two red lines show 95% confidence interval (2.5% top, 2.5% bottom).

A.4 From Section 5.4. Comparison to Other Fairness Measures

In this section, we show more visualizations of the results from bounding other fairness measures. All results are already presented and discussed in Section 5.4.

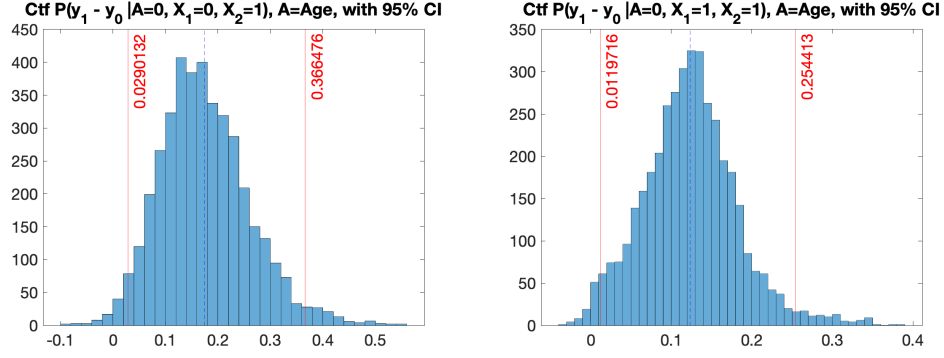


Figure 14: CE with $A = \text{Age}$ conditioned on $A = 0, W_1 = 0, W_2 = 1$ (left) and $A = 0, W_1 = 1, W_2 = 1$ (right). The two vertical red lines show 95% confidence interval (2.5% top, 2.5% bottom).

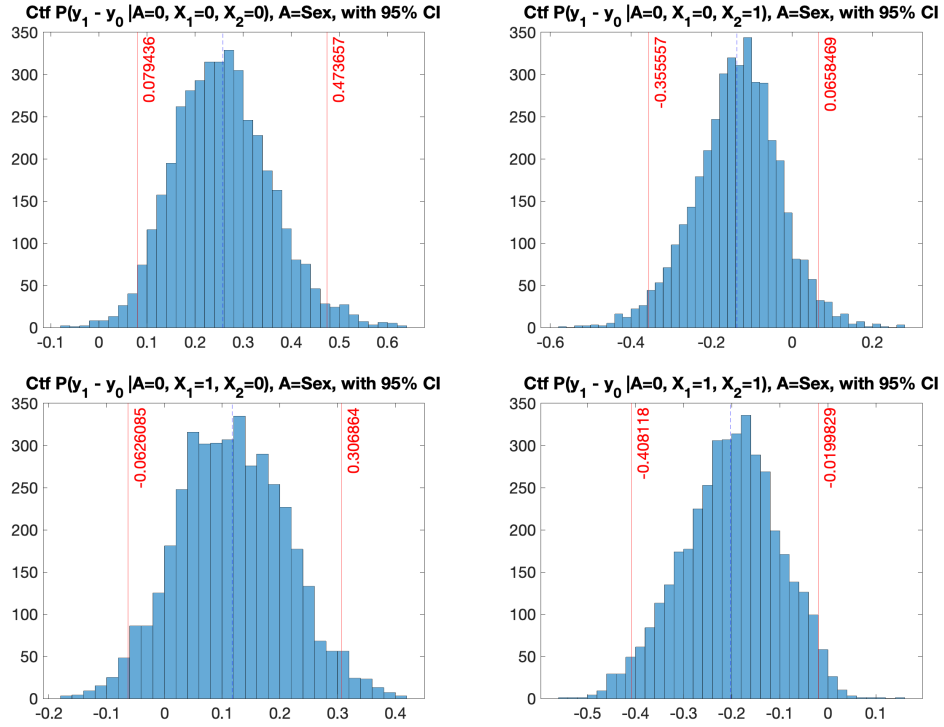


Figure 15: CE with $A = \text{Sex}$ conditioned on $A = 0$ and different values of W_1 and W_2

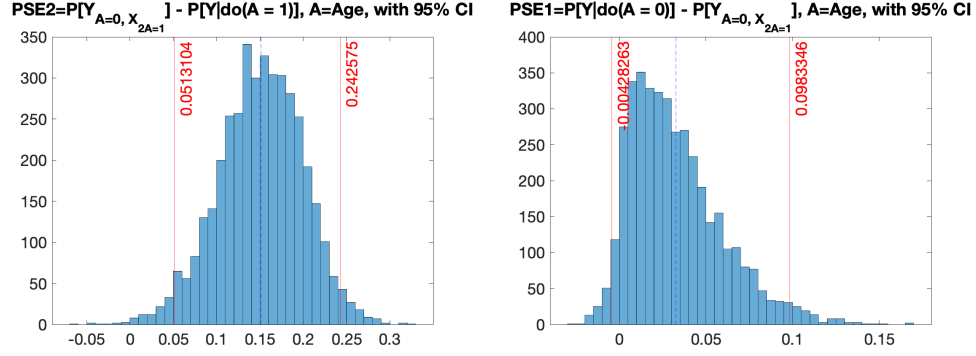


Figure 16: PSE of $A = \text{Age}$ for the path $A \rightarrow Y$ (left) and the path $A \rightarrow W_2 \rightarrow Y$ (right).

B Frequently Asked Questions

In this section, we provide answers to basic questions for the readers not too familiar with causal inference.

Q. What does it mean for counterfactual probabilities to be unidentifiable, and how prevalent is the case? A. An unidentifiable problem means that the target quantity is underdefined from the available observed data and assumptions on the causal structural model. Usually, counterfactual probabilities cannot be specified even if we know the true causal diagram with the data Pearl [2000]. More detailed explanations can be found in Chapters 1 and 7 of Pearl [2000].

Q. Can we use this algorithm without knowledge of the true data-generating process? A. Yes. Our Algorithms only require that the observed variables are discrete and finite, and the equivalence class is obtained by FCI. Section 4 uses a simulated dataset generated from a known data-generating process to test the accuracy of our method. However, in real-world cases, as demonstrated in Section 5, inference will be based solely on the observed data and detailed knowledge of the true data-generating process is not required.

Q. Then, are we assuming that the cardinality of the variables are bounded? A. No. The bounded cardinality is not an assumption, but a sufficient statistic that allows us to represent the observational distribution and the target counterfactual measurement in the ground-truth causal model. This result was first introduced in Zhang et al. [2022], and Theorem 3.2 extends it to nested counterfactual fairness measures, while improving the latent cardinality. In essence, Theorem 3.2 shows that there exists a natural parametrization for the latent space as a function of the cardinality of the observation distribution, which helps a user to make a data-driven decision.

Q. What if there are multiple candidates for $\mathcal{G}$ in $\mathcal{E}^*$? A. Then, the probability mass is equally distributed across all candidates. Its effect on the final bound will depend on the causal diagrams and the data, but the width of the bound is expected to increase when more options are considered (models with fewer assumptions).