Local Causal Discovery for Structural Evidence of Direct Discrimination

Jacqueline Maasch[1], Kyra Gan[1], Violet Chen[2], Agni Orfanoudaki[3], Nil-Jana Akpinar[4*], Fei Wang[5]

[1]Cornell Tech
[2]Stevens Institute of Technology
[3]University of Oxford
[4]Amazon AWS AI/ML (*Work done outside Amazon)
[5]Weill Cornell Medicine

Abstract

Identifying the causal pathways of unfairness is a critical objective for improving policy design and algorithmic decision-making. Prior work in causal fairness analysis often requires knowledge of the causal graph, hindering practical applications in complex or low-knowledge domains. Moreover, global discovery methods that learn causal structure from data can display unstable performance on finite samples, preventing robust fairness conclusions. To mitigate these challenges, we introduce *local discovery for direct discrimination* (LD3): a method that uncovers structural evidence of direct unfairness by identifying the causal parents of an outcome variable. LD3 performs a linear number of conditional independence tests relative to variable set size, and allows for latent confounding under the sufficient condition that all parents of the outcome are observed. We show that LD3 returns a valid adjustment set (VAS) under a new graphical criterion for the *weighted controlled direct effect*, a qualitative indicator of direct discrimination. LD3 limits unnecessary adjustment, providing interpretable VAS for assessing unfairness. We use LD3 to analyze causal fairness in two complex decision systems: criminal recidivism prediction and liver transplant allocation. LD3 was more time-efficient and returned more plausible results on real-world data than baselines, which took 46× to 5870× longer to execute.

1 Introduction

Fairness holds fundamental importance in policy design and algorithmic decision-making, especially in high-stakes domains such as healthcare and policing (Starke et al. 2022; Corbett-Davies et al. 2023). Various criteria have been proposed for measuring unfairness with respect to protected or sensitive attributes, such as gender and ethnicity (Verma and Rubin 2018). Legal doctrines generally differentiate between direct discrimination and indirect or spurious forms of unfairness, codifying the notion that mechanisms of unfairness matter (Barocas and Selbst 2016; Carey and Wu 2022). However, fairness criteria based solely on statistical associations cannot disentangle these mechanisms (Kilbertus et al. 2017; Makhlouf, Zhioua, and Palamidessi 2020), limiting their informativeness and actionability for policy interventions. Consequently, there is a growing emphasis on applying causal reasoning in fairness analysis (Kilbertus et al. 2017), shifting the focus from associations to interventions and counterfactual outcomes.

Causal fairness analysis (CFA) provides a theoretical framework for disentangling the mechanisms of unfairness using the language of *structural causal models* (SCMs). Previous works in CFA (Zhang and Bareinboim 2018; Plecko and Bareinboim 2023) and the allied field of mediation analysis (Pearl 2014; VanderWeele 2016) generally assume significant prior structural knowledge in order to decompose direct, indirect, and spurious effects. In practice, structural knowledge is often incomplete, absent, or contentious for complex domains, even among experts (Petersen et al. 2023). Furthermore, the identifiability of direct and indirect effects has been a topic of extensive debate among theoreticians (Pearl 2014), raising barriers to entry for applied researchers (Vanderweele 2011). Thus, existing methodologies in CFA can be challenging to apply in complex systems.

Among the many fairness measures proposed in CFA, the *controlled direct effect* (CDE) is a relatively straightforward qualitative indicator of direct discrimination that takes nonzero values only when the exposure is a direct cause of the outcome (Zhang and Bareinboim 2018). The CDE has often been favored in policy evaluation over alternative direct effect measures (Vanderweele 2011; VanderWeele 2013) as it is more interpretable for real-world interventions and requires fewer untestable assumptions and less prior structural knowledge (Pearl 2001; Shpitser and VanderWeele 2011). To increase the practicality of CFA in complex domains, we choose to focus on the CDE as a starting point for this work.

In low-knowledge domains, we can support CFA by learning causal structure directly from observational data. However, *global causal discovery* is challenging in finite data due to high sample complexity (Spirtes, Glymour, and Scheines 2001) and exponential time complexity in unconstrained search spaces (Chickering, Heckerman, and Meek 2004; Claassen, Mooij, and Heskes 2013). Learned causal graphs often disagree with expert knowledge in complex domains (Shen et al. 2020; Petersen et al. 2023), and can yield conflicting causal fairness conclusions in CFA (Binkytė et al. 2023 and Section 6 of this paper).

While global discovery learns the relations among all observed variables, *local causal discovery* only learns the substructures relevant for downstream tasks, such as causal effect estimation (Gupta, Childers, and Lipton 2023; Maasch

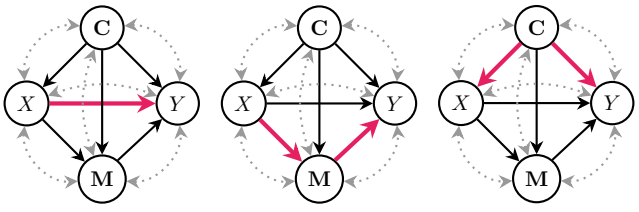

Figure 1: The *standard fairness model* (SFM) is compactly represented as a local subgraph around protected attribute X and outcome Y (Plečko and Bareinboim 2024). Variables that are irrelevant to CFA are abstracted away, leaving confounders ($\mathbf{C}$) and mediators ($\mathbf{M}$). Directed edges represent active paths and bidirected edges denote potential latent confounding. This work aims to identify *direct* mechanisms of unfairness in a data-driven way.

et al. 2024; Shah, Shanmugam, and Kocaoglu 2024) or feature selection (Yu et al. 2021; Yu, Liu, and Li 2021). As the *standard fairness model* (SFM; Figure 1) is represented as a local subgraph (Plečko and Bareinboim 2024), local discovery offers a natural framework for CFA. However, since task-specific local discovery algorithms are definitionally one-size-*does-not*-fit-all, existing methods may not be optimal for fairness tasks.

Contributions This work aims to increase the practicality of CFA for direct discrimination in complex domains with unknown causal graphs. Our contributions are three-fold.

1. **Local discovery for direct discrimination (LD3)**. This local causal discovery method leverages the problem structure in CFA to efficiently detect graphical signatures of direct discrimination.[1] LD3 discovers the parents of an outcome variable in a linear number of conditional independence tests with respect to variable set size.
2. **A graphical criterion for the weighted controlled direct effect (WCDE).** This criterion is sufficient to identify a valid adjustment set (VAS) for the WCDE, a qualitative indicator of direct discrimination. This criterion is satisfied by the knowledge returned by LD3.
3. **Real-world fairness analysis.** We deploy LD3 for two fairness problems: (1) racial discrimination in recidivism prediction and (2) sex-based discrimination in liver transplant allocation. LD3 recovered more plausible causal relations than local and global baselines, which performed 11–1021× more tests and took 46–5870× longer to run.

2 Preliminaries

Let capital letters denote univariate random variables (e.g., X), with their values in lowercase (e.g., $X = x$). Multivariate random variables or sets are denoted by boldface capital letters (e.g., $\mathbf{X}$), with vector values in bold lowercase (e.g., $\mathbf{X} = \mathbf{x}$). Graphs or function sets are denoted by calligraphic script (e.g., $\mathcal{F}$). Let $pa(\cdot)$ and $de(\cdot)$ denote the parent and descendant sets for a variable in causal graph $\mathcal{G}$, respectively.

[1]Code on GitHub: `https://github.com/jmaasch/LD3`

2.1 Causal Fairness Analysis

CFA can be framed in the language of SCMs and their graphical representations (Plečko and Bareinboim 2024).

Definition 1 (Structural causal model, Bareinboim et al. 2022)**.** An SCM is a 4-tuple $\langle \mathbf{V}, \mathbf{U}, \mathcal{F}, p(\mathbf{u}) \rangle$ where $\mathbf{U} = \{U_i\}_{i=1}^n$ denotes a set of exogenous variables determined by factors external to the model, $\mathbf{V} = \{V_i\}_{i=1}^n$ denotes a set of observed endogenous variables determined by $\mathbf{U} \cup \mathbf{V}$, $\mathcal{F} = \{f_i\}_{i=1}^n$ denotes a set of structural functions such that $V_i = f_i(pa(V_i), U_i)$, and $p(\mathbf{u})$ is the distribution over $\mathbf{U}$.

An SCM can be visually represented by a graphical model. To facilitate CFA, the true causal graph for an SCM can be compactly represented using the SFM (Figure 1).

Definition 2 (Standard fairness model, Plečko and Bareinboim 2024)**.** Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a causal graph with vertices $\mathbf{V}$ and edges $\mathbf{E}$. Let $\mathcal{G}_{\text{SFM}}$ be the *projection* of $\mathcal{G}$ onto the SFM, which is obtained by (1) selecting a protected attribute-outcome pair $\{X, Y\} \subset \mathbf{V}$ and (2) identifying sets $\mathbf{M}, \mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$ that meet the following conditions:

- $\mathbf{M}$ is the set of mediators with respect to X and Y;
- $\mathbf{C}$ is the set of confounders with respect to X and Y.

Note that $\mathbf{C}$ or $\mathbf{M}$ can be the empty set, and confounding can exist among $\mathbf{C}$, $\mathbf{M}$, X, Y (Figure 1).

Structural Fairness Criteria Multiple criteria have been proposed for evaluating fairness from graphical structures. Here, we focus on a criterion for direct discrimination.

Definition 3 (Structural direct criterion (SDC), Plečko and Bareinboim 2024)**.** An SCM is fair with respect to direct discrimination if and only if the SDC evaluates to 0:

$$SDC = \begin{cases} 1 & \text{if } X \text{ is a parent of } Y, \\ 0 & \text{if } X \text{ is not a parent of } Y. \end{cases} \quad (1)$$

2.2 Controlled Direct Effect

The CDE can be used to test for direct discrimination, as the true value is non-zero if and only if there is a direct path from the protected attribute to the outcome (Zhang and Bareinboim 2018). Let $X = x$ and $X = x^*$ be the exposure values corresponding to treatment and no treatment, respectively.

Definition 4 (CDE, Pearl 2014)**.** The CDE measures the expected change in outcome as the exposure changes when mediators $\mathbf{M}$ are uniformly fixed to a constant value $\mathbf{m}$:

$$\text{CDE}(\mathbf{m}) \coloneqq \mathbb{E}[Y \mid do(x, \mathbf{m})] - \mathbb{E}[Y \mid do(x^*, \mathbf{m})]. \quad (2)$$

Definition 5 (Identifiability conditions of the CDE, Vanderweele 2011)**.** When we have access to a covariate set $\mathbf{S}$ that controls for observed confounding of both $\{X, Y\}$ and $\{\mathbf{M}, Y\}$, CDE is identifiable under the following conditions. Let $Y_{\check{x}, \check{\mathbf{m}}}$ denote the value of Y when $X = \check{x}, \mathbf{M} = \check{\mathbf{m}}$.

1. There is no latent confounding of the exposure and outcome given $\mathbf{S}$, i.e., $Y_{\check{x}, \check{\mathbf{m}}} \perp\!\!\!\perp X | \mathbf{S}$ for all $\check{x}, \check{\mathbf{m}}$.
2. There is no latent confounding of the mediators and outcome given $\{X, \mathbf{S}\}$, i.e., $Y_{\check{x}, \check{\mathbf{m}}} \perp\!\!\!\perp \mathbf{M} | X, \mathbf{S}$ for all $\check{x}, \check{\mathbf{m}}$.

Then, we obtain $\text{CDE}(\mathbf{m})$ by[2]

$$\sum_{\mathbf{s}} \big(\mathbb{E}[Y \mid x, \mathbf{s}, \mathbf{m}] - \mathbb{E}[Y \mid x^*, \mathbf{s}, \mathbf{m}]\big) P(\mathbf{s}). \quad (3)$$

Comparison to Alternative Measures Several estimands can capture direct effects, including CDE (Pearl 2001), natural direct effect (NDE; Pearl 2001), and counterfactual direct effect (Ctf-DE; Zhang and Bareinboim 2018). While CDE is an interventional quantity, NDE and Ctf-DE are counterfactual quantities. In this work, we favor CDE as it requires fewer untestable assumptions over the data generating process (Shpitser and VanderWeele 2011) and less structural knowledge than NDE and Ctf-DE (Pearl 2001; Vanderweele 2011; Zhang and Bareinboim 2018). The main objective of this work is to assess whether the protected attribute is a direct cause of the outcome. The CDE, NDE, and Ctf-DE are all non-zero if and only if X is a causal parent of Y. Thus, we advocate for this simpler estimand for practicality. See Appendix A for an extended comparison of estimands.

Remark 1 (Assumptions on Confounding)**.** Unlike the NDE, CDE identification does not forbid the existence of confounders for $\{\mathbf{M}, Y\}$ that are descended from X (Pearl 2014). As done previously (Vanderweele and Vansteelandt 2009), this work assumes structures in which adjusting for confounders of $\{\mathbf{M}, Y\}$ does not induce post-treatment bias, allowing us to identify the CDE with the expression provided in Equation 3. If this assumption does not hold, unbiased CDE estimates can still be obtained given alternative estimators (Petersen, Sinisi, and Van Der Laan 2006).

Weighted CDE When interaction between the mediator and exposure exists, the CDE can vary across different mediator values (Pearl 2014). To avoid assumptions about interaction while still obtaining a unique estimate, we define the *weighted CDE* (WCDE) as the following expectation over $\mathbf{M}$. Note that we use discrete $\mathbf{M}$ merely for notational simplicity. All results generalize to continuous variables.

Definition 6 (WCDE, Pearl 2000)**.** We define WCDE as

$$\sum_{\mathbf{m}'} \big(\mathbb{E}[Y \mid do(x, \mathbf{m}')] - \mathbb{E}[Y \mid do(x^*, \mathbf{m}')]\big) P(\mathbf{m}'), \quad (4)$$

where $\mathbf{M}' \subseteq \mathbf{M}$ are parents of Y. Per Equation 3, WCDE is identifiable as

$$\sum_{\mathbf{m}'} \sum_{\mathbf{s}} \big(\mathbb{E}[Y|x, \mathbf{s}, \mathbf{m}'] - \mathbb{E}[Y|x^*, \mathbf{s}, \mathbf{m}']\big) P(\mathbf{s}) P(\mathbf{m}'). \quad (5)$$

Definition 7 (VAS for the WCDE)**.** Given Definitions 5 and 6, a VAS for WCDE estimation blocks (1) all backdoor paths for $\{X, Y\}$, (2) all backdoor paths for $\{\mathbf{M}, Y\}$, and (3) all mediator paths for $\{X, Y\}$.

As a Fairness Metric Similar to CDE, WCDE indicates direct discrimination when its value is non-zero. However, a zero value *does not* guarantee the absence of direct discrimination, as different CDE values may cancel each other out. Thus, we encourage caution when interpreting zero values.

[2]When $\mathbf{S}$ is not sufficient for valid control of confounding for both $\{X, Y\}$ and $\{\mathbf{M}, Y\}$, alternative formulae may be required (Vanderweele 2011; Pearl 2014).

Exhaustive, Disjoint Causal Partitions w.r.t. $\{X, Y\}$	
$\mathbf{Z}_1$	Confounders and their proxies.
$\mathbf{Z}_2$	Colliders and their proxies.
$\mathbf{Z}_3$	Mediators and their proxies.
$\mathbf{Z}_4$	Non-descendants of Y where $\mathbf{Z}_4 \perp\!\!\!\perp X$ and $\mathbf{Z}_4 \not\perp\!\!\!\perp X \mid Y$.
$\mathbf{Z}_5$	Instruments and their proxies.
$\mathbf{Z}_6$	Descendants of Y s.t. active paths with X are mediated by Y.
$\mathbf{Z}_7$	Descendants of X s.t. active paths with Y are mediated by X.
$\mathbf{Z}_8$	Nodes that share no active paths with X nor Y.

Table 1: Adapted from Maasch et al. (2024).

2.3 Mapping Causal Partitions to the SFM

The methods introduced in this work leverage the *causal partition* taxonomy defined in Maasch et al. 2024 (Table 1). Given an exposure-outcome pair $\{X, Y\}$, any arbitrary variable set $\mathbf{Z}$ can be uniquely partitioned into eight disjoint subsets (which may be empty) that are defined by the types of causal paths that they share with X and Y. By focusing structure learning on relationships that are *causally relevant* to the exposure and outcome, this partition taxonomy provides practical building blocks for local discovery. Mapping this partition taxonomy to the SFM, $\mathbf{Z}_1$ are confounders $\mathbf{C}$ (and their proxies) and $\mathbf{Z}_3$ are mediators $\mathbf{M}$ (and their proxies).[3] When referring to the union of multiple partitions, we use notation of the form $\mathbf{Z}_{1,3} \coloneqq \mathbf{Z}_1 \cup \mathbf{Z}_3$.

3 Local Discovery for Direct Discrimination

In CFA, we can translate the fairness query *"Is direct discrimination present?"* into the graphical query *"Is the protected attribute a parent of the outcome?"* To answer this graphical query, we focus on two indicators of parentage: the SDC and WCDE, where the former can be directly answered by LD3 and the latter can be estimated by using the VAS returned by LD3.

The input to LD3 (Algorithm 1) is a variable set $\mathbf{Z}$ of unknown causal relation to the protected attribute X and outcome Y. Instead of learning the causal graph, LD3 learns *causal partition labels* (Table 1). This local learning approach abstracts away structural information that is impertinent to direct discrimination detection, resulting in a computationally efficient discovery procedure. LD3 performs sequential CI tests to iteratively discover the partition label of each variable in $\mathbf{Z}$. Leveraging these labels ($\widehat{\mathbf{Z}}$), LD3 uses a test of d-separation to evaluate SDC (Lines 12–13). By conditioning only on $\widehat{\mathbf{Z}}_{1,3 \in pa(Y)}$, this CI test offers sample efficiency benefits relative to conditioning on all $\mathbf{Z}$. Additionally, LD3 returns a VAS for WCDE containing $pa(Y) \setminus X$.

Time Complexity Constraint-based discovery methods are typically analyzed by the number of CI tests performed (Spirtes, Glymour, and Scheines 2001). For-loops at Lines 2–5, 7–8, and 9–10 of Algorithm 1 perform $O(|\mathbf{Z}|)$ tests each. All remaining lines perform a constant number of op-

[3]See Maasch et al. (2024) for more formal partition definitions. Proxy variables are not relevant in this setting, as these cannot be parents of Y and are not returned by LD3.

Algorithm 1: *LD3*

Input: Exposure X, outcome Y, variable set $\mathbf{Z}$, CI test of choice, significance level α.
Output: Adjustment set $\mathbf{A}_{\text{DE}}$, SDC results.
Assumptions: Sufficient conditions A1 and A2.

$\mathbf{Z}' \leftarrow \mathbf{Z}$
for $\forall\, Z \in \mathbf{Z}'$ **do**
  if $Z \perp\!\!\!\perp X \wedge Z \perp\!\!\!\perp Y$ **then** $Z \in \widehat{\mathbf{Z}}_8$
  if $Z \not\perp\!\!\!\perp Y \wedge Z \perp\!\!\!\perp Y | X$ **then** $Z \in \widehat{\mathbf{Z}}_{5,7}$
  if $Z \perp\!\!\!\perp X \wedge Z \not\perp\!\!\!\perp X | Y$ **then** $Z \in \widehat{\mathbf{Z}}_4$
$\mathbf{Z}' \leftarrow \mathbf{Z}' \setminus \widehat{\mathbf{Z}}_8 \cup \widehat{\mathbf{Z}}_{5,7} \cup \widehat{\mathbf{Z}}_4$
for $\forall\, Z \in \mathbf{Z}'$ **do**
  if $Z \not\perp\!\!\!\perp Y | X \cup \widehat{\mathbf{Z}}_4 \cup \{\mathbf{Z}' \setminus Z\}$ **then** $Z \in \widehat{\mathbf{Z}}_{1 \in pa(Y)} \cup \widehat{\mathbf{Z}}_{3 \in pa(Y)}$
for $\forall\, \widehat{Z}_4 \in \widehat{\mathbf{Z}}_4$ **do**
  if $\widehat{Z}_4 \not\perp\!\!\!\perp Y | X \cup \widehat{\mathbf{Z}}_{1 \in pa(Y)} \cup \widehat{\mathbf{Z}}_{3 \in pa(Y)} \cup \{\widehat{\mathbf{Z}}_4 \setminus \widehat{Z}_4\}$ **then** $\widehat{Z}_4 \in \widehat{\mathbf{Z}}_{4 \in pa(Y)}$
$\mathbf{A}_{\text{DE}} \leftarrow \widehat{\mathbf{Z}}_{1 \in pa(Y)} \cup \widehat{\mathbf{Z}}_{3 \in pa(Y)} \cup \widehat{\mathbf{Z}}_{4 \in pa(Y)}$
if $X \perp\!\!\!\perp Y | \widehat{\mathbf{Z}}_{1 \in pa(Y)} \cup \widehat{\mathbf{Z}}_{3 \in pa(Y)}$ **then** $SDC \leftarrow 0$
else $SDC \leftarrow 1$
return $\mathbf{A}_{\text{DE}}, SDC$

erations. Thus, the total number of CI tests is of $O(|\mathbf{Z}|)$, ensuring scalability in real-world CFA.

Sufficient Conditions for Structure Learning We assume causal Markov, faithfulness, and acyclicity. We do not impose parametric assumptions on causal functions nor distributional forms. As for all constraint-based methods, the independence test selected may impose its own parametric assumptions. When these are not well-justified, we recommend nonparametric tests (e.g., Gretton et al. 2005, 2007).

Theorem 1. *Asymptotic guarantees on partitioning and SDC correctness hold under Assumptions A1 and A2. Given WCDE identifiability by Equation 5 (Remark 1), A1 and A2 are also sufficient for VAS discovery.*

A1 *Y has no descendants in the observed variable set.* This is satisfied when Y is a terminal variable in the temporal ordering (e.g., when outcome is death, a policy or algorithmic decision made at a known time point, etc.).

A2 *All parents of Y are observed.* Latent variables that are not parents of Y are permissible. Thus, this is a milder condition than assuming causal sufficiency.

Proof of Theorem 1 is in Appendix B. Note that assumptions A1 and A2 are sufficient but not necessary. A1 has been previously used to facilitate parent and ancestor learning (Soleymani et al. 2022; Cai et al. 2023). LD3 learns causal partitions directly from data, without assumptions on temporal ordering except (1) assumption A1 and (2) Y cannot cause X. Assumption A2 is a consequence of the sufficient conditions for CDE identifiability, as Definition 5 requires blocking all spurious and indirect paths into Y. A2 allows for unobserved variables that are not in $pa(Y)$, a more relaxed assumption than the causal sufficiency typically required in discovery (e.g., Spirtes, Glymour, and Scheines 2001; Zheng et al. 2018; etc.). We exploit A2 to evaluate the SDC, as it ensures that X is conditionally d-separable from Y when there is no direct path from X to Y. It also ensures colliders ($\mathbf{Z}_{2 \notin de(Y)}$) are removed from the adjustment set.

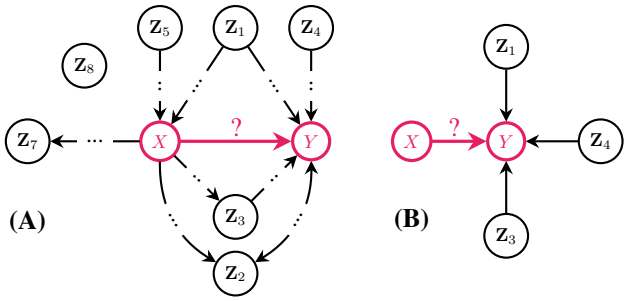

Figure 2: LD3 assesses whether the edge $X \to Y$ exists. **(A)** Allowable partitions under A1 and A2. **(B)** Parents of Y returned by LD3. Nodes are partition sets or subsets. Partition interrelations and latent confounding are abstracted away. Bidirected edge $Y \leftrightarrow \mathbf{Z}_2$ signifies $\mathbf{Z}_{2 \notin de(Y)}$. Edges with $\cdots$ are paths of arbitrary length. Solid edges are adjacencies.

We empirically demonstrate failure modes and robustness to violations of A2 in Appendix D.1, showing that A2 is sufficient but not necessary. We prove in Appendix B that latent variables not adjacent to Y do not impact correctness.

Theorem 2. *Latent variables that are not parents of Y do not affect Algorithm 1.*

Causal Partitions in this Setting Given A1, there are no descendants of Y in $\mathcal{G}$ and therefore no $\mathbf{Z}_6$ nor $\mathbf{Z}_{2 \in de(Y)}$. However, there may be $\mathbf{Z}_{2 \notin de(Y)}$. Thus, $\mathcal{G}$ can contain the following seven (potentially empty) causal partitions with respect to $\{X, Y\}$: $\mathbf{Z}_1$, $\mathbf{Z}_{2 \notin de(Y)}$, $\mathbf{Z}_3$, $\mathbf{Z}_4$, $\mathbf{Z}_5$, $\mathbf{Z}_7$, and $\mathbf{Z}_8$ (Figure 2.A). LD3 returns the partition *subsets* in Figure 2.B: all $\mathbf{Z}_1$, $\mathbf{Z}_3$, and $\mathbf{Z}_4$ that are directly adjacent to Y.

Remark 2 (The observed WCDE under violations of A2)**.** While A2 is sufficient but not necessary for WCDE identifiability, each backdoor and frontdoor path must be blocked by at least one observable variable. If no variables on such a path are measured, then no algorithm can identify the true WCDE. Users should note that the identifiability of any direct effect measure is fundamentally limited by the observability of backdoor and frontdoor paths.

3.1 A Graphical Criterion for the Weighted CDE

Under conditions where the WCDE is identifiable by Equation 5 (Remark 1), we propose the following criterion.

Definition 8 (Graphical criterion for identifying the WCDE)**.** Under the causal partition taxonomy defined in Maasch et al. (2024), we define the set $\mathbf{A}_{\text{DE}}$ that contains all parents of the outcome:

$$\mathbf{A}_{\text{DE}} \coloneqq \mathbf{Z}_{1 \in pa(Y)} \cup \mathbf{Z}_{3 \in pa(Y)} \cup \mathbf{Z}_{4 \in pa(Y)}. \tag{6}$$

Theorem 3 ($\mathbf{A}_{\text{DE}}$ is a VAS for the WCDE)**.** $\mathbf{A}_{\text{DE}}$ *is a valid adjustment set for the WCDE (Equation 5), satisfying the identification conditions in Definition 5.*

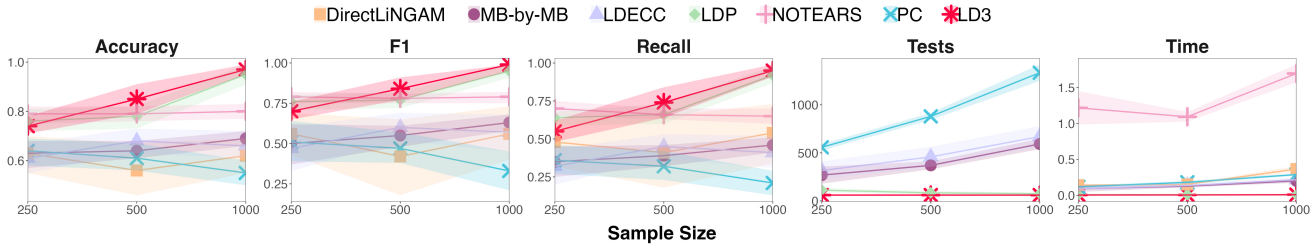

Figure 3: Baseline results for parent discovery on the SANGIOVESE benchmark. Independence test count (Tests) is reported for constraint-based methods. Time is in seconds. Shaded regions denote 95% confidence intervals over ten replicates.

Intuition. $\mathbf{A}_{\text{DE}}$ contains exactly all the parents of Y: confounders of $\{X, Y\}$ adjacent to Y ($\mathbf{Z}_{1 \in pa(Y)}$), mediators of $\{X, Y\}$ adjacent to Y ($\mathbf{Z}_{3 \in pa(Y)}$), and all parents of Y that are marginally independent of X ($\mathbf{Z}_{4 \in pa(Y)}$). Thus, at least one member of every backdoor and frontdoor path is in $\mathbf{A}_{\text{DE}}$. This provides conditional d-separation of X and Y if and only if there is no edge $X \to Y$. Proof is in Appendix B.

Remark 3 (The role of $\mathbf{Z}_4$)**.** Note that $\mathbf{Z}_3$ and Y can be confounded by $\mathbf{Z}_1$, $\mathbf{Z}_3$, or $\mathbf{Z}_4$ (Figure B.1). Including $\mathbf{Z}_{4 \in pa(Y)}$ in $\mathbf{A}_{\text{DE}}$ helps guarantee the identifiability of WCDE without requiring exact knowledge of confounding for $\mathbf{Z}_3$ and Y.

Remark 4 (Variance and Actionability)**.** Definition 8 defines a VAS for the WCDE in the general setting, irrespective of A1 and A2. In settings where LD3 is used to obtain $\mathbf{A}_{\text{DE}}$, $\mathbf{Z}$ itself constitutes a VAS (per A1 and A2). However, adjusting for all $\mathbf{Z}$ is not advised for two primary reasons:

1. Adjusting for all $\mathbf{Z}$ risks *unnecessary adjustment*, which can inflate estimator variance under finite data (Schisterman, Cole, and Platt 2009). To support statistical efficiency, we follow intuition provided by prior theorems on VAS optimality with respect to asymptotic variance, which dictate exclusion of $\mathbf{Z}_5$, inclusion of $\mathbf{Z}_4$, and generally favor control for parents of Y (Rotnitzky and Smucler 2020; Henckel, Perković, and Maathuis 2022).
2. Ignoring causal structure limits the informativeness and actionability of fairness conclusions. Forgoing structure learning is a missed opportunity to identify potential structural mechanisms that could be redressed through interventions (e.g., policy change).

A full discussion of this topic is in Appendix C. Sections 5 and 6 provide empirical support for points (1) and (2).

4 Related Works

Discovery for CFA Few works in causal discovery have centered on fairness objectives (Binkytė et al. 2023). These include learning Suppes-Bayes causal networks with maximum likelihood estimation (Bonchi et al. 2017) and applying PC Algorithm (Zhang, Wu, and Wu 2017). To our knowledge, this work presents the first *local* causal discovery method that is specifically tailored for CFA.

Local Discovery of Direct Causes Learning the direct causes of a target has primarily garnered interest in causal feature selection (Soleymani et al. 2022). Various Markov blanket (MB) learners have been proposed, though many cannot distinguish parents from children and/or spouses (Yu et al. 2020). MB-by-MB (Wang et al. 2014), Causal Markov Blanket (CMB; Gao and Ji 2015), and Local Discovery Using Eager Collider Checks (LDECC; Gupta, Childers, and Lipton 2023) are constraint-based methods that differentiate parents and children when unambiguous over the Markov equivalence class (MEC). Like the global algorithm PC (Spirtes, Glymour, and Scheines 2001), MB-by-MB, CMB, and LDECC have worst-case exponential time complexity with respect to variable set size. Local Discovery by Partitioning (LDP; Maasch et al. 2024) causally partitions variables around $\{X, Y\}$ using a quadratic number of CI tests with respect to variable set size; results can be post-processed to identify parents of Y if A1 is imposed.

5 Empirical Validation on Synthetic Data

Baselines Baselines represent a range of approaches that are available as open-source Python implementations. Local baselines are MB-by-MB (Wang et al. 2014), LDECC (Gupta, Childers, and Lipton 2023), and LDP (Maasch et al. 2024). Global baselines are PC (Spirtes, Glymour, and Scheines 2001), DirectLiNGAM (Shimizu et al. 2011), and NOTEARS (Zheng et al. 2018). Extended baseline descriptions and post-processing procedures are given in Appendix D.2. Besides LDP, all baselines assume causal sufficiency. PC, MB-by-MB, and LDECC return results in terms of the MEC. DirectLiNGAM assumes an additive noise model. DirectLiNGAM and NOTEARS assume linearity. All experiments used an Apple MacBook (M2 Pro Chip).

Parent Discovery We evaluated whether LD3 can recover true parent sets using an oracle independence test on random exposure-outcome pairs in 90 unique directed acyclic Erdős-Rényi graphs (node counts in $[5 \ldots 500]$). Parent F1, recall, and precision were 100%. Runtimes and total CI tests as node and edge cardinality scale are shown in Figure D.2. In Appendix D.4, we show for a linear-Gaussian SCM that WCDE estimates converged toward the true direct effect with low variance when adjusting for $\mathbf{A}_{\text{DE}}$ discovered with LD3 (Figures D.3, D.4).

All baselines were assessed on the SANGIOVESE benchmark from the `bnlearn` repository (Scutari 2010), a linear-Gaussian model of Tuscan grape production (Magrini, Di Blasi, and Stefanini 2017). Ten replicate datasets were

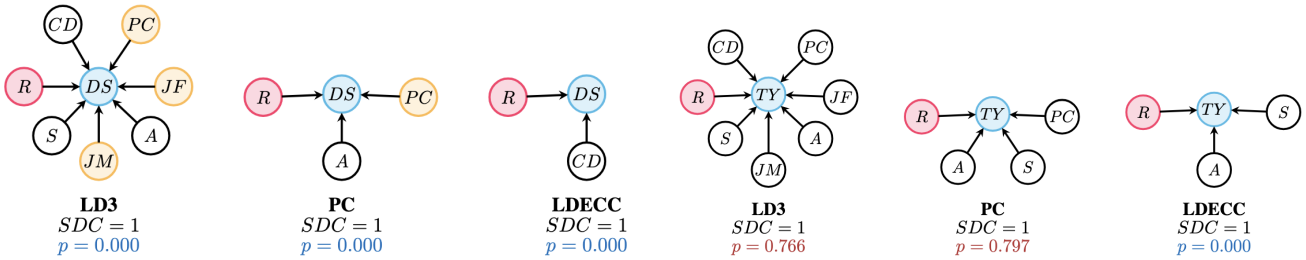

Figure 4: Predicted parent sets, SDC, and WCDE p-values for COMPAS. Exposure is race (*R*; red) and outcome is general recidivism risk decile score (*DS*; blue). Known parents of *DS* are in yellow. *A* = age; *CD* = charge degree; *JF* = juvenile felonies; *JM* = juvenile misdemeanors; *PC* = priors count; *S* = sex. All methods used χ^2 CI tests ($\alpha = 0.05$).

sampled at $n = [250, 500, 1000]$. All constraint-based methods used Fisher-z tests ($\alpha = 0.01$). DirectLiNGAM assumes non-Gaussian noise and was expected to underperform. LD3 was generally most performant across metrics, with LDP performing similarly (Figure 3, Table D.5). NOTEARS was significantly slower than other methods. Comparisons of LD3 to LDECC, LDP, and PC on ASIA (Lauritzen and Spiegelhalter 1988) and SACHS (Sachs et al. 2005) benchmarks are in Tables D.6 and D.8. Runtime comparisons are in Figure D.2 and Tables D.7, D.9.

Estimator Variance in Finite Samples As discussed in Remark 4, $\mathbf{Z}$ itself is a VAS under A1 and A2. However, adjusting for $\mathbf{Z}$ risks unnecessary adjustment, which can inflate the asymptotic variance of the causal effect estimator. We demonstrate the impacts of variance inflation as sample size scales in both a linear and nonlinear SCM (Figure D.5). Estimate variance using all $\mathbf{Z}$ was $7.8\times$ to $9.6\times$ higher than using $\mathbf{A}_{\text{DE}}$ in the linear SCM (Table D.2) and at least $12.6\times$ higher in the nonlinear SCM (Table D.3).

VAS Interpretability Structure learning can improve the interpretability of the VAS by removing irrelevant variables. In some structures, this removal can substantially reduce VAS size. On two `bnlearn` benchmarks with moderate to large DAGs and small $\mathbf{A}_{\text{DE}}$ cardinality, discovery with LD3 reduced adjustment set cardinality by at least 95.4% relative to retaining all $\mathbf{Z}$ (Table D.4).

6 Real-World Causal Fairness Analyses

We deploy LD3 for two causal fairness settings: (1) racial discrimination in criminal recidivism prediction and (2) sex-based discrimination in healthcare. We compare the results of LD3 to PC and LDECC on the basis of $\mathbf{A}_{\text{DE}}$ quality and computational efficiency. Causal discovery used χ^2 CI tests and WCDE estimation used double machine learning (Chernozhukov et al. 2018).[4] Estimators used random forest classifiers with a 70% / 30% train-test split. We assumed A1 and A2. Data preprocessing is described in Appendix E. Note that all CFA results are only preliminary qualitative indicators, and further analyses should take place.

[4] `https://econml.azurewebsites.net`

6.1 Race and COMPAS Recidivism Prediction

Background We assessed the ability of LD3 to facilitate CFA on the ProPublica COMPAS dataset. Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a commercial algorithm for decision support used by the US criminal justice system to assess risk of recidivism (NorthPointe 2015). ProPublica's landmark exposé on COMPAS found that African Americans were "almost twice as likely as whites to be labeled a higher risk but not actually re-offend" (Angwin et al. 2022).

We examined racial bias in the COMPAS General Recidivism Risk model. Due to data availability, we limited our analyses to the most represented racial groups (black and white). The algorithm's developer states that the model directly considers prior criminal history and juvenile delinquency, among other factors (NorthPointe 2015, p. 27). Our data contained multiple indicators of criminal history and juvenile delinquency, so these were used to assess the quality of causal discovery in lieu of complete ground truth. We used three significance levels for independence testing ($\alpha = 0.005, 0.01, 0.05$) to assess stability of results (Figures 4, E.3–E.5). We selected 11 features with $n = 6150$ observations (2454 white, 3696 black; see Appendix E). We explored two outcomes: (1) *general recidivism decile score* to probe bias in the COMPAS algorithm, and (2) *actual two-year recidivism*, to examine factors in real outcomes.

Results At all significance levels, results from LD3, PC, and LDECC qualitatively agree that the effects of race on decile score are not fully explained by observed[5] confounding and mediation (SDC = 1) and that the WCDE is significantly different from zero ($p = 0.000$). LD3 successfully predicted that juvenile delinquency is a parent of decile score at all significance levels, while PC and LDECC never did. Age, priors count, and charge degree were nearly always predicted to be parents across methods and significance levels. For two-year recidivism, LD3 stably predicted that the WCDE of race was not significant ($p = 0.87, 0.68, 0.77$).

In general, results for PC and LDECC were less stable than LD3. Both methods wavered between strong significance ($p = 0.000$) and no significance ($p > 0.4$) for the WCDE of race on two-year recidivism. LDECC parent sets

[5] Note that if A2 was violated, some of the observed effects might be due to unobserved variables. See Remark 2.

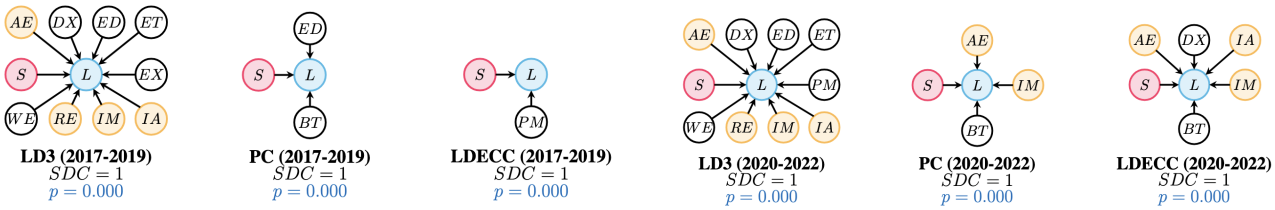

Figure 5: Predicted parent sets, SDC, and WCDE p-values for STAR liver data. Exposure is patient sex (S; red) and outcome is receiving a liver (L; blue). Known parents of L are in yellow. *AE* = active exception case; *BT* = recipient blood type; *DX* = diagnosis; *ED* = education; *ET* = ethnicity; *EX* = exception type; *IA* = initial age; *IM* = initial MELD; *PM* = payment method; *RE* = region; *WE* = weight. All methods used χ^2 CI tests ($\alpha = 0.01$). Additional results are in Tables E.2–E.4.

were inconsistent depending on whether they were taken as the intersection or union across graphs in the MEC, requiring additional interpretation by the user. On average, PC and LDECC took 46× longer to run and performed at least 11.7× more tests than LD3 across experiments (Figure E.2).

6.2 Sex and Liver Transplant Allocation

Background We applied LD3 to a case study in the US healthcare system: fairness in liver transplant allocation. Liver transplantation is a critical therapeutic option for patients with end-stage chronic liver disease and acute liver failure. Demand significantly surpasses supply for donor livers (SRTR 2020), and patients are placed on a national waiting list managed by the United Network for Organ Sharing (UNOS). The distribution policy sorts waitlisted patients by multiple criteria, such as medical urgency, compatibility, and location (Latt, Niazi, and Pyrsopoulos 2022). Several key policy changes have sought to optimize distribution and improve patient outcomes (Papalexopoulos et al. 2023; see Appendix E.2), including the *model for end-stage liver disease* (MELD; Malinchoc et al. 2000). Despite efforts to increase fairness (Kim et al. 2022), it is widely recognized that US organ allocation suffers from disparities (Zhang et al. 2018).

We explored potential sex-based discrimination in liver allocation. Sex-based disparities have been observed as statistical associations (Oloruntoba and Moylan 2015; Allen et al. 2018; Nephew and Serper 2021), but have not been explored through a causal lens. We use the National Standard Transplant Analysis and Research (STAR) dataset (OPTN 2024) for adult patients during 2017-2019 ($n = 21,101$) and 2020-2022 ($n = 22,807$). See Appendix E.2 for feature selection and summary statistics.

Results Under all experimental settings and time frames, results from LD3 suggest that the effects of sex on receiving a liver were not fully explained by observed confounding and mediation (Figure 5, Table E.2). Causal evidence of direct discrimination was detected (SDC = 1) and corroborated by non-zero WCDE with significant p-values (< 0.005). In all settings, $\mathbf{A}_{\text{DE}}$ contains key factors used in the liver distribution policy, including initial MELD, initial age, region, and active exception case (OPTN 2024). Our results are concordant with prior observations that differences in body size and medical condition contribute to sex-based disparities (Nephew et al. 2017), as weight and diagnosis are in $\mathbf{A}_{\text{DE}}$ in all settings. Predicted $\mathbf{A}_{\text{DE}}$ also indicate potential discrimination with respect to ethnicity, education, and payment method, which may warrant further investigation.

PC and LDECC qualitatively agree with LD3 on both discrimination metrics (Figure 5, Tables E.3, E.4). However, PC, LDECC, and LD3 had relatively low agreement in terms of $\mathbf{A}_{\text{DE}}$. Adjustment set cardinality was lower (and at times zero) for PC and LDECC. Most settings for PC and LDECC returned $\mathbf{A}_{\text{DE}}$ that omitted key expected variables, such as known policy criteria (e.g., initial MELD score, active exception case, and initial age). At both significance levels, PC returned multiple untrustworthy edges (e.g., weight $\to$ sex, height $\to$ sex, blood type $\to$ age, education $\to$ height, body mass index $\to$ height), undermining the credibility of results. Likewise, LDECC predicts that the outcome has children, which is known to be untrue. PC performed 479–1021× more tests and took 2295–5870× longer to execute than LD3 (Table E.5). LDECC performed 42–984× more tests and took 197–5774× longer to run.

7 Conclusion

This work advocates for increased practicality in causal fairness pipelines. We propose a time-efficient and asymptotically correct local discovery method for identifying two qualitative indicators of direct discrimination: the SDC and WCDE. For two real-world causal fairness analyses, LD3 returned more stable and plausible predictions with signficantly better computational efficiency relative to baselines.

Limitations and Future Directions Future work could extend LD3 to allow for Y with descendants. The assumption that all parents of Y are observed is sufficient but not necessary for CDE identification and could be replaced with a different criterion. LD3 cannot differentiate $\mathbf{Z}_1$ from $\mathbf{Z}_3$, which would enable analysis of indirect and spurious discrimination. Future work could consider discovery for other fairness estimands, such as the natural effects (Pearl 2014).

Acknowledgments

This research was supported by NSF 2212175; NIH RF1AG084178, R01AG076448, R01AG080624, R01AG076234, R01AG080991 and RF1AG072449; and the NSF GRFP under Grant No. DGE – 2139899.

References

Allen, A. M.; Heimbach, J. K.; Larson, J. J.; Mara, K. C.; Kim, W. R.; Kamath, P. S.; and Therneau, T. M. 2018. Reduced access to liver transplantation in women: role of height, MELD exception scores, and renal function underestimation. *Transplantation*, 102(10): 1710–1716.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2022. Machine bias. In *Ethics of data and analytics*, 254–264. Auerbach Publications.

Bareinboim, E.; Correa, J. D.; Ibeling, D.; and Icard, T. 2022. On Pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: the works of judea pearl*, 507–556. ACM.

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *California Law Review*, 104: 671.

Binkytė, R.; Makhlouf, K.; Pinzón, C.; Zhioua, S.; and Palamidessi, C. 2023. Causal discovery for fairness. In *Workshop on Algorithmic Fairness through the Lens of Causality and Privacy*, 7–22. PMLR.

Bonchi, F.; Hajian, S.; Mishra, B.; and Ramazzotti, D. 2017. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics*.

Cai, H.; Wang, Y.; Jordan, M.; and Song, R. 2023. On learning necessary and sufficient causal graphs. *Advances in Neural Information Processing Systems*, 36.

Carey, A. N.; and Wu, X. 2022. The causal fairness field guide: Perspectives from social and formal sciences. *Frontiers in big Data*, 5: 892837.

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

Chickering, D. M.; Heckerman, D.; and Meek, C. 2004. Large-Sample Learning of Bayesian Networks is NP-Hard. *Journal of Machine Learning Research*, 5: 1287–1330.

Claassen, T.; Mooij, J. M.; and Heskes, T. 2013. Learning Sparse Causal Models is not NP-hard. *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI2013)*.

Corbett-Davies, S.; Gaebler, J. D.; Nilforoshan, H.; Shroff, R.; and Goel, S. 2023. The measure and mismeasure of fairness. *The Journal of Machine Learning Research*.

Gao, T.; and Ji, Q. 2015. Local Causal Discovery of Direct Causes and Effects. In Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, 63–77. Springer.

Gretton, A.; Fukumizu, K.; Teo, C.; Song, L.; Schölkopf, B.; and Smola, A. 2007. A kernel statistical test of independence. *Advances in neural information processing systems*.

Gupta, S.; Childers, D.; and Lipton, Z. C. 2023. Local Causal Discovery for Estimating Causal Effects. In *Proceedings of the 2nd Conference on Causal Learning and Reasoning (CLeaR)*. arXiv. ArXiv:2302.08070 [cs, stat].

Henckel, L.; Perković, E.; and Maathuis, M. H. 2022. Graphical Criteria for Efficient Total Effect Estimation via Adjustment in Causal Linear Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

Kilbertus, N.; Rojas Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

Kim, I. K.; Martins, P. N.; Pavlakis, M.; Eneanya, N. D.; and Patzer, R. E. 2022. Past and Present Policy Efforts in Achieving Racial Equity in Kidney Transplantation. *Current Transplantation Reports*, 9(2): 114–118.

Latt, N. L.; Niazi, M.; and Pyrsopoulos, N. T. 2022. Liver transplant allocation policies and outcomes in United States: A comprehensive review. *World Journal of Methodology*.

Lauritzen, S. L.; and Spiegelhalter, D. J. 1988. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50.

Maasch, J.; Pan, W.; Gupta, S.; Kuleshov, V.; Gan, K.; and Wang, F. 2024. Local Discovery by Partitioning: Polynomial-Time Causal Discovery Around Exposure-Outcome Pairs. In *Proceedings of the 40th Conference on Uncertainy in Artificial Intelligence*.

Magrini, A.; Di Blasi, S.; and Stefanini, F. M. 2017. A conditional linear Gaussian network to assess the impact of several agronomic settings on the quality of Tuscan Sangiovese grapes. *Biometrical Letters*, 54(1): 25–42.

Makhlouf, K.; Zhioua, S.; and Palamidessi, C. 2020. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*.

Malinchoc, M.; Kamath, P. S.; Gordon, F. D.; Peine, C. J.; Rank, J.; and Ter Borg, P. C. 2000. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. *Hepatology*, 31(4): 864–871.

Nephew, L. D.; Goldberg, D. S.; Lewis, J. D.; Abt, P.; Bryan, M.; and Forde, K. A. 2017. Exception points and body size contribute to gender disparity in liver transplantation. *Clinical gastroenterology and hepatology*, 15(8): 1286–1293.

Nephew, L. D.; and Serper, M. 2021. Racial, gender, and socioeconomic disparities in liver transplantation. *Liver Transplantation*, 27(6): 900–912.

NorthPointe. 2015. Practitioner's Guide to COMPAS Core.

Oloruntoba, O. O.; and Moylan, C. A. 2015. Gender-based disparities in access to and outcomes of liver transplantation. *World journal of hepatology*, 7(3): 460.

OPTN. 2024. Data Request Instructions.

Papalexopoulos, T.; Alcorn, J.; Bertsimas, D.; Goff, R.; Stewart, D.; and Trichakis, N. 2023. Reshaping national organ allocation policy. *Operations Research*.

Pearl, J. 2000. *Causality: Models, reasoning and inference*. Cambridge University Press. ISBN 978-0-521-77362-1.

Pearl, J. 2001. Direct and Indirect Effects. In *Proceedings of the 17th Conference on Uncertainy in Articial Intelligence*.

Pearl, J. 2014. Interpretation and identification of causal mediation. *Psychological Methods*, 19(4): 459–481.

Petersen, A. H.; Ekstrøm, C. T.; Spirtes, P.; and Osler, M. 2023. Constructing Causal Life-Course Models: Comparative Study of Data-Driven and Theory-Driven Approaches. *American Journal of Epidemiology*, 192(11): 1917–1927.

Petersen, M. L.; Sinisi, S. E.; and Van Der Laan, M. J. 2006. Estimation of Direct Causal Effects. *Epidemiology*, 17.

Plecko, D.; and Bareinboim, E. 2023. Causal Fairness for Outcome Control. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*.

Plečko, D.; and Bareinboim, E. 2024. Causal Fairness Analysis: A Causal Toolkit for Fair Machine Learning. *Foundations and Trends® in Machine Learning*, 17(3): 304–589.

Rotnitzky, A.; and Smucler, E. 2020. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188): 1–86.

Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D. A.; and Nolan, G. P. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*.

Schisterman, E. F.; Cole, S. R.; and Platt, R. W. 2009. Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies. *Epidemiology*, 20(4): 488–495.

Scutari, M. 2010. Learning Bayesian networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3).

Shah, A.; Shanmugam, K.; and Kocaoglu, M. 2024. Frontdoor Adjustment Beyond Markov Equivalence with Limited Graph Knowledge. *Advances in Neural Information Processing Systems*, 36.

Shen, X.; Ma, S.; Vemuri, P.; Simon, G.; and The Alzheimer's Disease neuroimaging initiative. 2020. Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology. *Scientific Reports*, 10(2975).

Shimizu, S.; Inazumi, T.; Sogawa, Y.; Hyvarinen, A.; Kawahara, Y.; Washio, T.; Hoyer, P. O.; Bollen, K.; and Hoyer, P. 2011. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr): 1225–1248.

Shpitser, I.; and VanderWeele, T. J. 2011. A Complete Graphical Criterion for the Adjustment Formula in Mediation Analysis. *The International Journal of Biostatistics*, 7.

Soleymani, A.; Raj, A.; Bauer, S.; Schölkopf, B.; and Besserve, M. 2022. Causal Feature Selection via Orthogonal Search. *Transactions on Machine Learning Research*. ArXiv:2007.02938 [cs, math, stat].

Spirtes, P.; Glymour, C.; and Scheines, R. 2001. *Causation, prediction, and search*. MIT press.

SRTR. 2020. OPTN/SRTR 2020 Annual Data Report: Liver. Accessed: 2024-05-20.

Starke, C.; Baleis, J.; Keller, B.; and Marcinkowski, F. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, 9(2): 20539517221115189.

Vanderweele, T. J. 2011. Controlled Direct and Mediated Effects: Definition, Identification and Bounds. *Scandinavian Journal of Statistics*, 38(3): 551–563.

VanderWeele, T. J. 2013. Policy-relevant proportions for direct effects. *Epidemiology*, 24(1): 175–176.

VanderWeele, T. J. 2016. Mediation analysis: a practitioner's guide. *Annual review of public health*, 37: 17–32.

Vanderweele, T. J.; and Vansteelandt, S. 2009. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2(4): 457–468.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, 1–7.

Wang, C.; Zhou, Y.; Zhao, Q.; and Geng, Z. 2014. Discovering and orienting the edges connected to a target variable in a DAG via a sequential local learning approach. *Computational Statistics & Data Analysis*, 77: 252–266.

Yu, K.; Guo, X.; Liu, L.; Li, J.; Wang, H.; Ling, Z.; and Wu, X. 2020. Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys (CSUR)*, 53(5): 1–36.

Yu, K.; Guo, X.; Liu, L.; Li, J.; Wang, H.; Ling, Z.; and Wu, X. 2021. Causality-based Feature Selection: Methods and Evaluations. *ACM Computing Surveys*, 53(5): 1–36.

Yu, K.; Liu, L.; and Li, J. 2021. A Unified View of Causal and Non-causal Feature Selection. *ACM Transactions on Knowledge Discovery from Data*, 15(4): 1–46.

Zhang, J.; and Bareinboim, E. 2018. Fairness in Decision-Making — The Causal Explanation Formula. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Zhang, L.; Wu, Y.; and Wu, X. 2017. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 3929–3935.

Zhang, X.; Melanson, T. A.; Plantinga, L. C.; Basu, M.; Pastan, S. O.; Mohan, S.; Howard, D. H.; Hockenberry, J. M.; Garber, M. D.; and Patzer, R. E. 2018. Racial/ethnic disparities in waitlisting for deceased donor kidney transplantation 1 year after implementation of the new national kidney allocation system. *American Journal of Transplantation*.

Zheng, X.; Aragam, B.; Ravikumar, P. K.; and Xing, E. P. 2018. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31.