Causal graphical models for target trial emulation

Jacqueline Maasch

Cornell Computer Science

maasch@cs.cornell.edu

20 February 2024

Table of Contents

Clinical trials: Essential but challenging

Randomized controlled trials (RCTs) are the gold standard for evaluating the treatment effects of medical interventions (e.g., efficacy and safety).

However, they pose significant challenges.

Clinical trials: Essential but challenging

- **Time:** Phase I, months; Phase II, months to 2 years; Phase III, 1–4 years [FDA].
- **Financial costs:**
  - On average, clinical trials cost $28M in Phase I, $65M in Phase II, $282M in Phase III.[1]
  - $0.8B to $2.3B in R&D spending per new drug, clinical and preclinical [US CBO].
  - For each successful drug, an average of $690M is spent on drugs that fail [US CBO].
- **Generalizability:** Participants are often not representative of the general population, due to stringent inclusion/exclusion criteria (favoring "ideal patients").

[1] Joseph A. DiMasi, Henry G. Grabowski, and Ronald W. Hansen, "Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs," Journal of Health Economics, vol. 47 (May 2016), pp. 23–24, https://doi.org/10.1016/j.jhealeco.2016.01.012.

We cannot replace RCTs entirely.

But can we complement R&D with *in silico* strategies?

Target trial emulation (TTE)

- **TTE** incorporates principles of clinical trial design into observational data analysis, using diverse strategies from causal inference, statistics, and machine learning.
- **Acceleration:** Can we use insights from observational data to probe hypotheses on faster timescales, and inform subsequent RCT design?
- **Automation:** Can we use algorithms to automate steps of R&D in a safe, principled, and data-driven manner? What theoretical guarantees can we provide?

Advantages of TTE:

- **Efficient:** Running an algorithm is faster than performing an RCT.
- **Cheap:** Analysis on existing observational data is less costly than performing an RCT.
- **Generalizable**: Optimally, we can obtain data for large-scale, real-world populations.

Challenges of TTE:

- **Incomplete information:** Not everything can be captured in the data (e.g., genetic markers for Alzheimer's Disease are not available in EHR).
- **Irregular follow-up:** Patient visits are temporally irregular.
- **No randomization:** Confounding poses major challenges.

Advantages of TTE:

- **Efficient:** Running an algorithm is faster than performing an RCT.
- **Cheap:** Analysis on existing observational data is less costly than performing an RCT.
- **Generalizable**: Optimally, we can obtain data for large-scale, real-world populations.

Challenges of TTE:

- **Incomplete information:** Not everything can be captured in the data (e.g., genetic markers for Alzheimer's Disease are not available in EHR).
- **Irregular follow-up:** Patient visits are temporally irregular.
- **No randomization:** Confounding poses major challenges.

Probabilistic graphical modeling is a branch of machine learning that uses probability distributions to

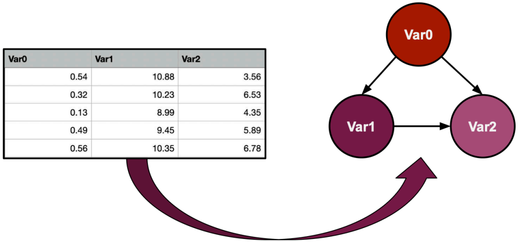
- Describe the world.
- Make predictions.
- Support decision-making under uncertainty.

PGMs concisely represent joint distributions over complex domains, like human disease or the economy. This framework yields powerful generative models for probabilistic and causal reasoning that can assist TTE.

Example PGMs include Bayesian networks (i.e., directed acyclic graphs, or DAGs) and variational autoencoders.

Causal discovery: A tool for insights

Causal DAGs are PGMs that impose a causal interpretation on directed edges, enabling causal reasoning.



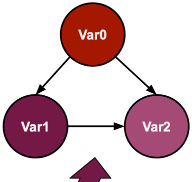
Var0	Var1	Var2
0.54	10.88	3.56
0.32	10.23	6.53
0.13	8.99	4.35
0.49	9.45	5.89
0.56	10.35	6.78

Causal discovery can be used to learn joint representations from data via diverse strategies.

Causal discovery: A tool for insights
1 Background

Many avenues for inference open up once the structure of the data is known.

Var0	Var1	Var2
0.54	10.88	3.56
0.32	10.23	6.53
0.13	8.99	4.35
0.49	9.45	5.89
0.56	10.35	6.78



For example, discovering confounder $Var0$ can enable causal effect estimation for $Var1 \rightarrow Var2$.

My work: Methods ⟷ Applications

1 Background

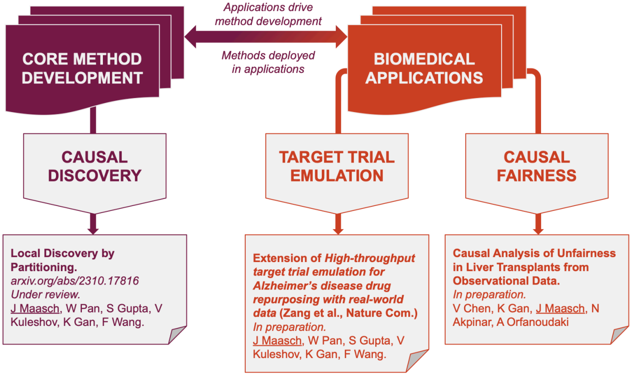


Table of Contents

2 Local Discovery by Partitioning (LDP)

Local Discovery by Partitioning: Polynomial-Time Causal Discovery Around Exposure-Outcome Pairs

Jacqueline R. M. A. Maasch[1] **Weishen Pan**[2] **Shantanu Gupta**[3] **Volodymyr Kuleshov**[1] **Kyra Gan**[4] **Fei Wang**[2]

[1]Department of Computer Science, Cornell Tech, New York, NY
[2]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY
[3]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA
[4]Department of Operations Research and Information Engineering, Cornell Tech, New York, NY

Motivation: Covariate adjustment in TTE

2 Local Discovery by Partitioning (LDP)

nature communications

Article https://doi.org/10.1038/s41467-023-43929-1

High-throughput target trial emulation for Alzheimer's disease drug repurposing with real-world data

Received: 13 February 2022

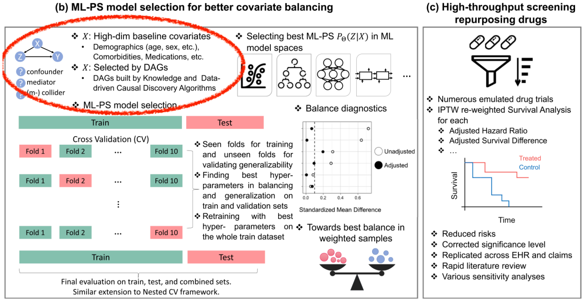
Accepted: 24 November 2023

Chengxi Zang[1,2], Hao Zhang[1], Jie Xu[3], Hansi Zhang[3], Sajjad Fouladvand[4], Shreyas Havaldar[5], Feixiong Cheng[6,7,8], Kun Chen[9], Yong Chen[10], Benjamin S. Glicksberg[5], Jin Chen[4], Jiang Bian[3] & Fei Wang[1,2]

This work emulated trials for thousands of medications from two large-scale real-world data warehouses (10+ years of clinical records, 170M+ patients), using propensity score models under the IPTW framework and causal discovery for covariate selection.

Motivation: Covariate adjustment in TTE

2 Local Discovery by Partitioning (LDP)

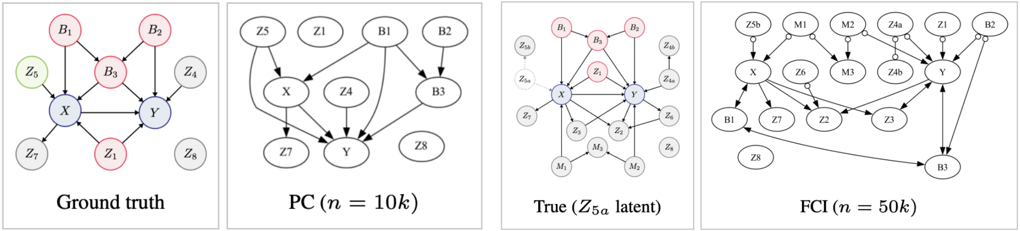


Covariates were selected using a variant of the PC algorithm [1]. Inferred colliders and mediators were removed. The remaining variables formed the adjustment set.

Why is this approach problematic?

- **The forbidden set:** Valid adjustment sets (VAS) cannot contain descendants of the exposure, but mediators and colliders are only a fraction of this forbidden set.
- **Ambiguity:** PC returns the Markov equivalence class (MEC) of the true DAG, leaving some paths ambiguous.
- **Sample complexity:** PC only provides asymptotic guarantees, and displays notorious failure modes under finite samples due to high sample complexity.
- **Latent variables:** The possibility of latent confounding (i.e., *causal insufficiency*) is not addressed.

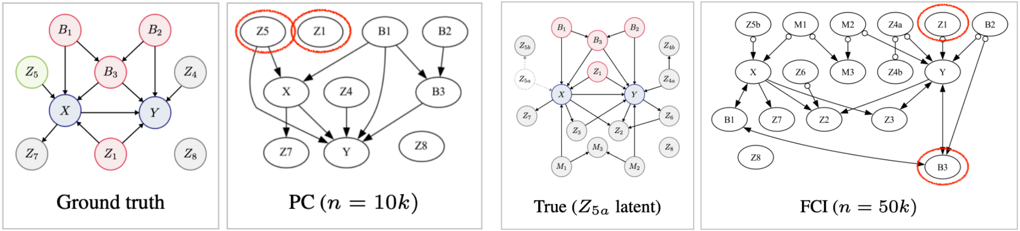
Motivation: Is global discovery effective?



Classic constraint-based global discovery [1] can fail even on simple DAGs with moderately large sample sizes, in causally sufficient and insufficient settings.

Motivation: Is global discovery effective?

2 Local Discovery by Partitioning (LDP)



Classic constraint-based global discovery [1] can fail even on simple DAGs with moderately large sample sizes, in causally sufficient and insufficient settings.

Contributions of LDP

1. **Partition taxonomy:** A taxonomy of eight exhaustive, mutually exclusive *causal partitions* that are *universal properties* of any dataset w.r.t. an exposure-outcome pair.

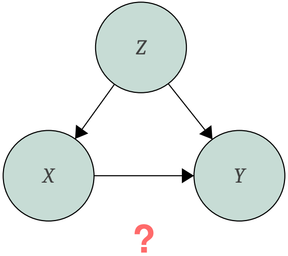
2. **Partition discovery method:** A nonparametric, polynomial-time procedure for learning these partitions.

3. **VAS discovery method:** LDP returns a VAS under causal insufficiency and mild graphical conditions (relative to existing automated covariate selection methods).

Contributions of LDP

- **Time efficiency:** Total independence tests is worst-case quadratic with respect to total variables, versus exponential for common baselines. On a community benchmark, LDP ran at least $\mathbf{1300}\times$ **faster than baselines.**
- **Sample efficiency:** The majority of CI tests defined in LDP use conditioning sets of size one or two, contributing to more favorable sample efficiency.
- **Flexibility:** LDP is nonparametric and does not assume the magnitude of the exposure-outcome effect (which may be null). We replace the conventional *pretreatment assumption* with a milder graphical requirement.

Does X cause Y? If so, how strong is the effect?
How can we achieve *conditional exchangeability*, such that this effect is *identifiable*?



Preliminaries: Why not adjust for *everything*?

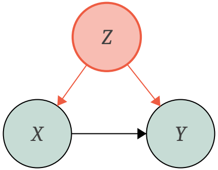
- **Bias:** Multiple variable types can induce bias when retained for adjustment [2, 3].
  1. Colliders induce selection bias [4–6].
  2. Mediators bias total effects by controlling for indirect effects [7].
  3. Instruments can amplify existing bias or introduce new bias in some settings [8].

- **Variance:** Unnecessary adjustment can inflate the variance of effect estimates [3].

- **Curse of dimensionality:** Unnecessary adjustment can undermine model fitting [9].

Preliminaries: Non-causal associations

Definition 2.3 (Backdoor path, Pearl 2009)**.** Any non-causal path between exposure X and outcome Y with an edge pointing into X $(\cdots \to X)$.

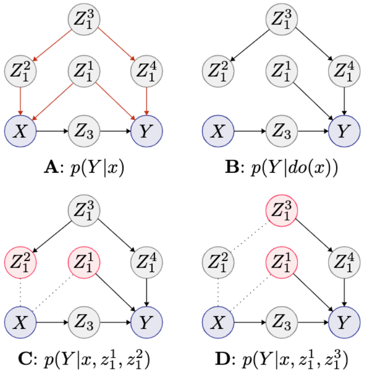


Definition 2.4 (Valid adjustment under the backdoor criterion, Peters et al. 2017). Let $\mathbf{A}_{XY}$ be an adjustment set for $\{X, Y\}$ that does not contain $\{X, Y\}$. $\mathbf{A}_{XY}$ is valid if

1. $\mathbf{A}_{XY}$ contains no descendants of X and
2. $\mathbf{A}_{XY}$ blocks all backdoor paths from X to Y.

Preliminaries: Adjustment blocks backdoor paths

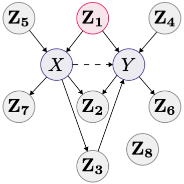
2 Local Discovery by Partitioning (LDP)



Method: Partitions are universal properties

2 Local Discovery by Partitioning (LDP)

Mutually Exclusive Partitions of Arbitrary $\mathbf{Z}$	
$\mathbf{Z}_1$	*Confounders and their proxies*: Non-descendants of X that lie on an active backdoor path between X and Y (Definition 2.5), and their proxies (Definition B.9).
$\mathbf{Z}_2$	*Colliders and their proxies*: Non-ancestors of $\{X, Y\}$ with at least one active path to X not mediated by Y and at least one active path to Y not mediated by X.
$\mathbf{Z}_3$	*Mediators and their proxies*: Descendants of X that are ancestors of Y, and their proxies (Definition B.9).
$\mathbf{Z}_4$	Non-descendants of Y that are marginally dependent on Y but marginally independent of X (Definition B.4).
$\mathbf{Z}_5$	*Instruments and their proxies*: Non-descendants of X whose causal effect on Y is fully mediated by X, and that share no confounders with Y (Definition B.2).
$\mathbf{Z}_6$	Descendants of Y where all active paths shared with X are mediated by Y.
$\mathbf{Z}_7$	Descendants of X where all active paths shared with Y are mediated by X.
$\mathbf{Z}_8$	All nodes that share no active paths with X nor Y.



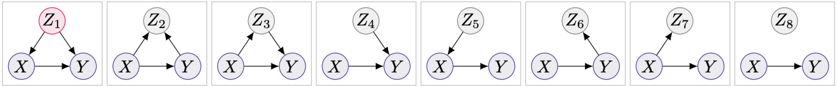


Figure 2: All potential acyclic triples that can be induced by X, Y, and a single Z when paths are restricted to a length of 1.

Method: LDP learns partitions progressively

2 Local Discovery by Partitioning (LDP)

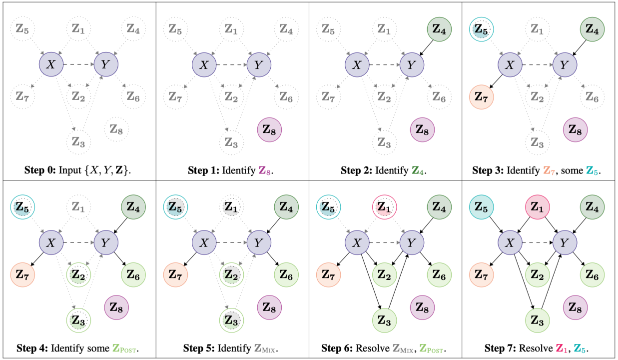


Table D.1: Schematic of Algorithm 1. The exposure-outcome pair $\{X, Y\}$ serves as a nucleus around which LDP assembles a partial causal graph. Each step reveals additional information about the partitions of $\mathbf{Z}$. Nodes that are fully colored are fully discovered by Algorithm 1. Partial coloring denotes partial knowledge, and no coloring denotes no knowledge.

Algorithm 1 *Local Discovery by Partitioning (LDP)*

input $X, Y, \mathbf{Z}$, independence test, p-value threshold.
output Partitions of $\mathbf{Z}$: $\mathbf{Z}_1, \mathbf{Z}_4, \mathbf{Z}_5, \mathbf{Z}_7, \mathbf{Z}_8, \mathbf{Z}_{\text{POST}}$.

Copy $\mathbf{Z}' \leftarrow \mathbf{Z}$
for all $Z \in \mathbf{Z}'$ **do**
▷ STEP 1: TEST FOR $\mathbf{Z}_8$
  if $X \perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z$ **then**
    $Z \in \mathbf{Z}_8, \mathbf{Z}' \leftarrow \mathbf{Z}' \setminus Z$
▷ STEP 2: TEST FOR $\mathbf{Z}_4$
  if $X \perp\!\!\!\perp Z$ and $X \not\perp\!\!\!\perp Z | Y$ **then**
    $Z \in \mathbf{Z}_4, \mathbf{Z}' \leftarrow \mathbf{Z}' \setminus Z$
▷ STEP 3: TEST FOR $\mathbf{Z}_{5,7}$
  if $Y \not\perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z | X$ **then**
    $Z \in \mathbf{Z}_{5A,7}, \mathbf{Z}' \leftarrow \mathbf{Z}' \setminus Z$

▷ STEP 4: TEST FOR $\mathbf{Z}_{\text{POST}}$
if $|\mathbf{Z}_4| > 0$ **then**
  for all $Z \in \mathbf{Z}'$ **do**
    if $\exists\, Z_4$: $Z \not\perp\!\!\!\perp Z_4$ or $Z \perp\!\!\!\perp Z_4 | X \cup Y$ **then**
      $Z \in \mathbf{Z}_{\text{POST}}$
$\mathbf{Z}' \leftarrow \mathbf{Z}' \setminus \mathbf{Z}_{\text{POST}}$

▷ STEP 5: TEST FOR $\mathbf{Z}_{\text{MIX}}$
for all $Z \in \mathbf{Z}'$ **do**
  if $Y \not\perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z | X \cup \mathbf{Z}' \setminus Z$ **then**
    $Z \in \mathbf{Z}_{1,2,3,5} \in \mathbf{Z}_{\text{MIX}}$
$\mathbf{Z}' \leftarrow \mathbf{Z}' \setminus \mathbf{Z}_{\text{MIX}}$

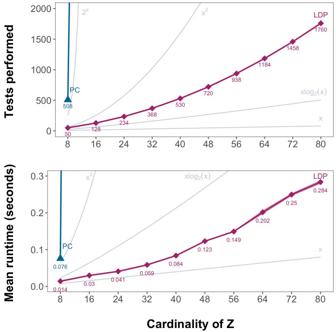
▷ STEP 6: SPLIT $\mathbf{Z}_{\text{MIX}}$ BETWEEN $\mathbf{Z}_{1,5}$, $\mathbf{Z}_7$, $\mathbf{Z}_{\text{POST}}$
$\mathbf{Z}_{\text{MIX}} \leftarrow \mathbf{Z}_{\text{MIX}} \cup \mathbf{Z}_{5,7}$
if $|\mathbf{Z}_{\text{MIX}}| > 0$ **then**
  for all $Z \in \mathbf{Z}'$ **do**
    if $\exists\, Z_{\text{MIX}}$: $Z_{\text{MIX}} \perp\!\!\!\perp Z$ and $Z_{\text{MIX}} \not\perp\!\!\!\perp Z | X$ **then**
      $Z \in \mathbf{Z}_1$, $Z_{\text{MIX}} \in \mathbf{Z}_{1,5} \notin \mathbf{Z}_{\text{MIX}}$
    else
      $Z \in \mathbf{Z}_{\text{POST}}$
  for all $Z_{\text{MIX}} \in \mathbf{Z}_{\text{MIX}}$ **do**
    if $\exists\, Z_{1,5}$: $Z_{1,5} \perp\!\!\!\perp Z_{\text{MIX}}$ **then**
      $Z_{\text{MIX}} \in \mathbf{Z}_1$
    else
      $Z_{\text{MIX}} \in \mathbf{Z}_{\text{POST}}$

▷ STEP 7: FINALIZE $\mathbf{Z}_1$ AND $\mathbf{Z}_5$
if $|\mathbf{Z}_{1,5}| > 0$ and $|\mathbf{Z}_1| > 0$ **then**
  for all $Z_{1,5} \in \mathbf{Z}_{1,5}$ **do**
    if $\exists\, Z_1 \in \mathbf{Z}_1$: $Z_{1,5} \not\perp\!\!\!\perp Z_1$ **then**
      $Z_{1,5} \in \mathbf{Z}_1$
    else
      $Z_{1,5} \in \mathbf{Z}_5$
if $|\mathbf{Z}_5| > 0$ **then**
  for $Z_5 \in \mathbf{Z}_5$ **do**
    if $Z_5 \not\perp\!\!\!\perp X | \mathbf{Z}_5 \cup \mathbf{Z}_{\text{POST}} \setminus Z_5$ **then**
      $Z_5 \in \mathbf{Z}_{5 \in adj(X)}$ and $\mathbf{Z}_1$ is a VAS

$\{\text{not identifiable}\} \leftarrow \mathbf{Z}'$
return Partitions of $\mathbf{Z}$ and $\{\text{not identifiable}\}$.

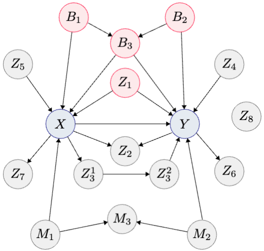
Results: Fewer tests, faster runtimes

2 Local Discovery by Partitioning (LDP)



Results with an oracle. On a community benchmark, LDP ran **1400× to 2500× faster** than PC.

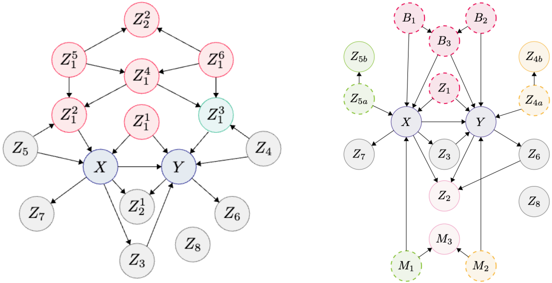
Results: Partition correctness



LDP correctly partitions $98.7\%[97.6, 99.9]$ of linear-Bernoulli instantiations and $98.7\%[98.0, 99.4]$ of quadratic hypergeometric instantiations of this DAG (100 replicates each, $n = 20k$).

Results: VAS with and without latent variables

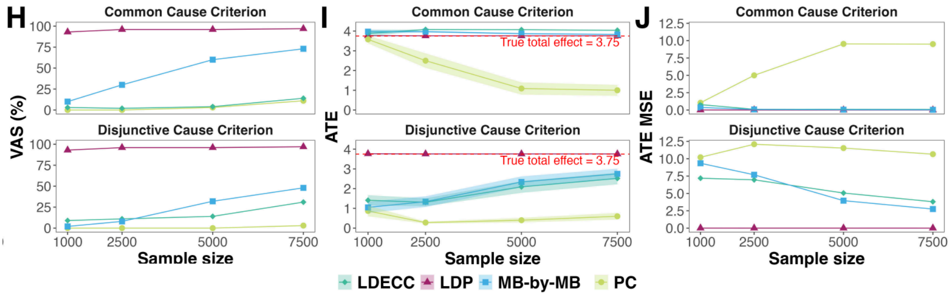
2 Local Discovery by Partitioning (LDP)



LDP provides VAS for these structures 100% of the time with an oracle and $\geq 97\%$ on finite samples. At right, each variable with a dashed perimeter was iteratively dropped to simulate latent confounding.

Results: Less biased, more precise ATE

2 Local Discovery by Partitioning (LDP)



Results on a 10-node linear-Gaussian DAG suggest that LDP offers greater sample and statistical efficiency for average treatment effect (ATE) estimation relative to baseline causal discovery algorithms.

Next steps

2 Local Discovery by Partitioning (LDP)

1. **Target trial emulation**
   - **Research question:** Does more rigorous covariate selection improve causal inference for drug repurposing?
   - We will reperform the large-scale Alzheimer's drug repurposing TTE proposed in Zang et al. 2023 [10] using LDP for covariate selection.

2. **Causal fairness and heterogeneous treatment effects**
   - **Research question:** Do sex, ethnicity, or other protected attributes cause differential patient outcomes in liver transplantation in the US health system?
   - Drawing principles from the Standard Fairness Model [11], we use LDP to identify VAS for heterogeneous treatment effect estimation.

Thank you! Any questions?

maasch@cs.cornell.edu

References

[1] P. Spirtes et al. *Causation, Prediction, and Search*. en. Ed. by J. Berger et al. Vol. 81. Lecture Notes in Statistics. New York, NY: Springer New York, 2000. DOI: 10.1007/978-1-4612-2748-9.

[2] H. Lu et al. "Revisiting Overadjustment Bias". en. In: *Epidemiology* 32.5 (2021), e22–e23. DOI: 10.1097/EDE.0000000000001377.

[3] E. F. Schisterman et al. "Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies". en. In: *Epidemiology* 20.4 (2009), pp. 488–495. DOI: 10.1097/EDE.0b013e3181a819a1.

[4] M. A. Hernán et al. "A Structural Approach to Selection Bias". en. In: *Epidemiology* 15.5 (2004), pp. 615–625. DOI: 10.1097/01.ede.0000135174.63482.43.

[5] F. Elwert et al. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable". In: *Annual Review of Sociology* 40.1 (2014), pp. 31–53. DOI: 10.1146/annurev-soc-071913-043455.

[6] M. J. Holmberg et al. "Collider bias". In: *JAMA Guide to Statistics and Methods* 327.13 (2022).

[7] J. Pearl. "Direct and Indirect Effects". In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. 2001.

[8] J. Pearl. "On a Class of Bias-Amplifying Variables that Endanger Effect Estimates". en. In: (2012).

[9] M. E. Schnitzer et al. "Variable Selection for Confounder Control, Flexible Modeling and Collaborative Targeted Minimum Loss-Based Estimation in Causal Inference". In: *The International Journal of Biostatistics* 12.1 (2016), pp. 97–115. DOI: 10.1515/ijb-2015-0017.

[10] C. Zang et al. "High-throughput target trial emulation for Alzheimer's disease drug repurposing with real-world data". In: *Nature Communications* 14.1 (2023), p. 8180.

[11] D. Plečko et al. "Causal Fairness Analysis: A Causal Toolkit for Fair Machine Learning". In: *Foundations and Trends® in Machine Learning* 17.3 (2024), pp. 304–589.