arXiv link to full paper.

Proximal Causal Inference with Text Data

Jacob M. Chen†, Rohit Bhattacharya†, Katherine A. Keith†

jchen459@jhu.edu, *rb17*@williams.edu, *kak5*@williams.edu

†Dept. of Computer Science, Williams College

Williams College

Motivation

- Proximal causal inference (PCI) allows practitioners to identify the average causal effect (ACE) in the presence of unmeasured confounding, but essential conditions for identification are difficult to verify [6].
- Researchers have proposed using text data to infer proxies for confounders [7], but this requires ground-truth labels for a subset of instances, something that is often impractical due to privacy concerns.
- We propose a new causal inference method that uses unique instances of pre-treatment text data, infers two proxies with zero-shot models on the instances, and applies the proxies in the two-stage linear regression proximal g-formula [6].

Motivating Example

- We want to evaluate the effectiveness of clot busting medication to treat strokes.
- **Target of Inference:**

$$ACE = \mathbb{E}[Y \mid \mathrm{do}(A=1)] - \mathbb{E}[Y \mid \mathrm{do}(A=0)]$$

- **Problem:** (i) Atrial fibrillation (irregular heart rhythms) is an important confounder that is not recorded in the structured data. (ii) Atrial fibrillation is an unmeasured confounder; e.g., we do not have access to atrial fibrillation status for any individuals in the dataset.

Basics of Proximal Causal Inference

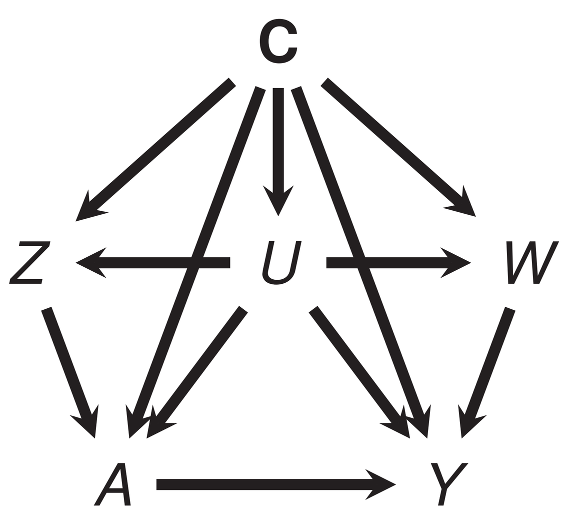

Figure: Canonical example of a DAG where (P1-P4) are fulfilled.

- When two proxies W and Z of the unmeasured confounder U relative to treatment A, outcome Y, and a set of baseline confounders $\mathbf{C}$ satisfy the following conditions:

(P1) $W \perp\!\!\!\perp Z \mid U, \mathbf{C}$

(P2) $W \perp\!\!\!\perp A \mid U, \mathbf{C}$

(P3) $Z \perp\!\!\!\perp Y \mid A, U, \mathbf{C}$

(P4) Completeness (intuition): W and Z are predictive of U and, if they are discrete, W and Z have the same number or more categories than U has.

The ACE is identified through the *proximal g-formula* [6].

- Throughout this work, we use the two-stage linear regression estimator for the proximal g-formula [5].
- **Problem:** How can we find two proxies W and Z among the structured variables that satisfy (P1-P4)?
- **Answer:** We cannot, at least not without a high degree of domain knowledge.

Designing Text-Based Proxies

- **Solution:** Infer our own proxies from text data using zero-shot text classifiers, but beware of pitfalls/gotchas.
- **Gotcha #1:** Using text-based inferences directly in backdoor adjustment. Subfigure (a) does not satisfy the backdoor criterion.
- **Gotcha #2:** Using post-treatment text. Subfigure (b) fails (P2) and (P3).
- **Gotcha #3:** Predicting both proxies from the same instance of text data. Subfigure (c) fails (P1).
- **Gotcha #4:** Using a single zero-shot classifier. In practice, we find that using two zero-shot classifiers works better.
- **Proposition.** If W and Z are inferred from two unique instances of pre-treatment text such that $\mathbf{T}_1^{\text{pre}} \perp\!\!\!\perp \mathbf{T}_2^{\text{pre}} \mid U, \mathbf{C}$, then these proxies satisfy (P1-P3). Additionally, if the proxies are predictive of U, i.e., $Z \not\perp\!\!\!\perp U \mid \mathbf{C}$ and $W \not\perp\!\!\!\perp U \mid \mathbf{C}$, then (P4) holds.
- **Problem:** How can we know that we inferred text-based proxies that fulfill (P1-P4)?

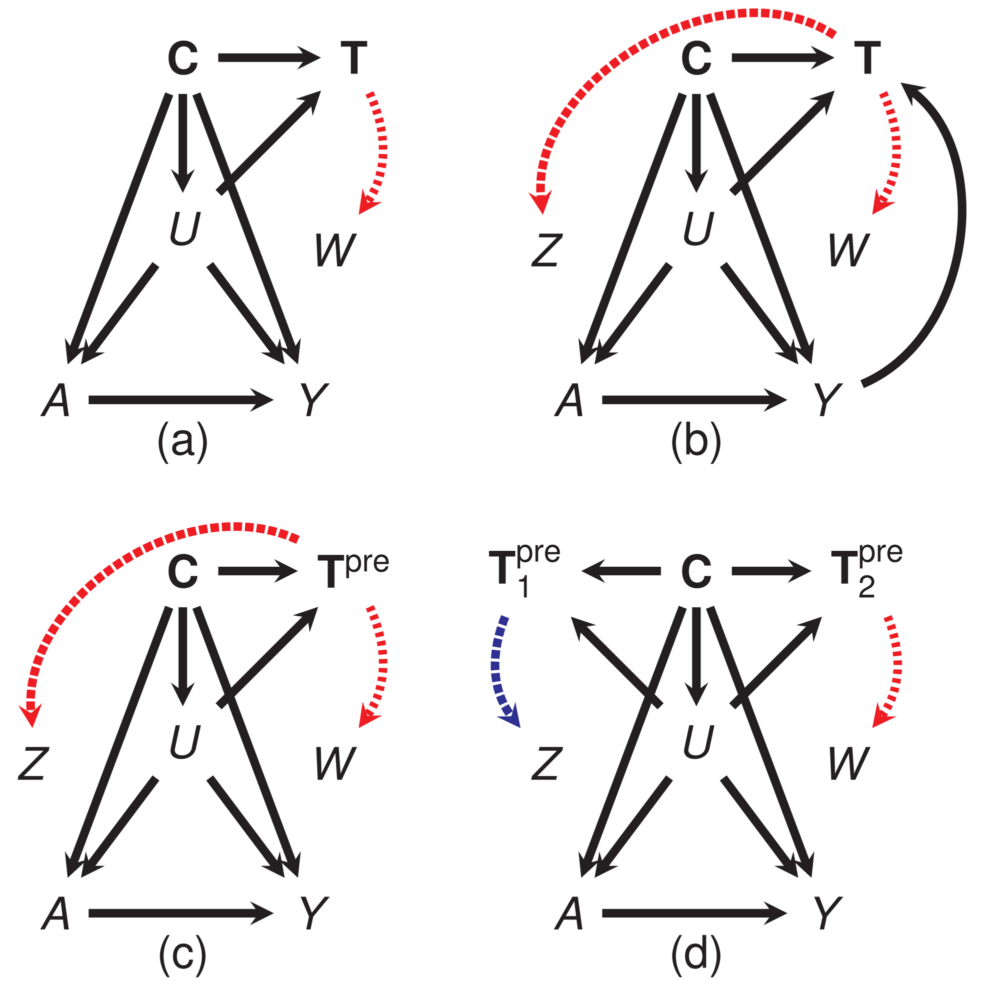

Figure: DAGs showing different scenarios for inferring text-based proxies. Dashed edges with different colors indicate that different zero-shot models that were used to infer proxies.

Odds Ratio Falsification Heuristic

- **Solution:** We propose a heuristic that warns us whenever (P1-P4) may be violated by the inferred proxies.
- We represent the odds ratio [1] – a measure of association between two variables – as a single free parameter, $\gamma_{WZ.\mathbf{C}}$, and estimate it under a linear parametric model for $p(W|Z, \mathbf{C})$. Algorithm 1 summarizes our procedure.

Algorithm 1 for inferring two text-based proxies

```
Inputs: Observed confounders C; Text T;
Zero-shot models M_1, M_2; Specified γ_high and γ_low
Extract two instances of pre-treatment text T_1^pre and T_2^pre from T
Z ← M_1(T_1^pre) and W ← M_2(T_2^pre)
// Odds Ratio Falsification Heuristic
if γ_low < γ_{WZ.C}^{CI low} and γ_{WZ.C}^{CI high} < γ_high then
    return W and Z
```

Fully Synthetic Experiments

Pipeline	$(\gamma_{WZ.\mathbf{C}}^{\text{CI low}}, \gamma_{WZ.\mathbf{C}}^{\text{CI high}})$	Est. ACE	Conf. Interval (CI)
P1M	$(1.35, 1.42)^{\checkmark}$	1.304	$(1.209, 1.394)$
P1M, same	$(10^{16}, 10^{16})$	1.430	$(1.405, 1.495)$
P2M	$(1.82, 1.94)^{\checkmark}$	1.343	$(1.273, 1.425)$
P2M, same	$(7.9, 8.41)$	1.407	$(1.376, 1.479)$

Table: **Fully synthetic results** with the true ACE equal to 1.3. Here, ✓ distinguishes settings that passed the odds ratio falsification heuristic from those that failed it. Corresponding to Gotcha #3, "same" means we use the same instance of synthetic text data to infer proxies. Proximal-1-Model (P1M) uses one zero-shot classifier for inference, and Proximal-2-Models (P2M) uses two zero-shot classifiers. We set $\gamma_{\text{low}=1}$ and $\gamma_{\text{high}=2}$.

- As expected, both P1M and P2M yield valid inferences under synthetic, ideal conditions.

Semi-Synthetic Experiments

- We generate semi-synthetic data from the MIMIC-III dataset [4] and use Echocardiogram, Radiology, and Nursing notes to infer proxies with instruction-tuned large language models Flan-T5 [2] and OLMo [3].
- Corresponding to Gotcha #1, we compare our text-based proximal causal inference estimators to using one of the inferred proxies directly in backdoor adjustment.
- Our odds ratio falsification heuristic correctly identifies invalid proxies, and we find that P2M is more likely to generate valid proxies in practice.

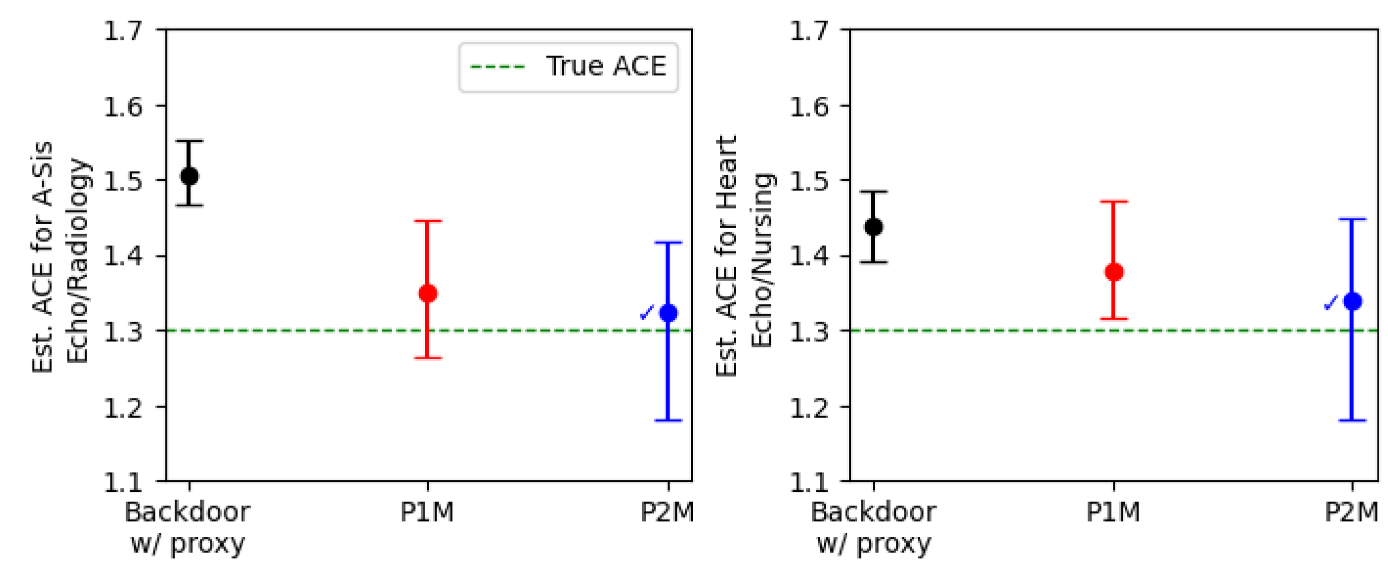

Figure: Estimates and bootstrap confidence intervals for the ACE when the unmeasured confounder is coronary atherosclerosis of the native coronary artery (A-Sis) and congestive heart failure (Heart). Blue and red distinguish passing and failing the odds ratio heuristic, respectively. We set $\gamma_{\text{low}} = 1$ and $\gamma_{\text{high}} = 2$.

Future Work

- How can we integrate non-linear proximal estimation?
- Can we extend our semi-synthetic studies to social science settings such as social media and education?
- Can we incorporate categorical U, W, and Z?
- What is the efficacy of using soft probabilistic outputs from the zero-shot classifiers?

References

[1] H. Y. Chen. A semiparametric odds ratio model for measuring association. *Biometrics*, 63:413–421, 2007.

[2] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[3] D. Groeneveld, I. Beltagy, P. Walsh, A. Bhagia, R. Kinney, O. Tafjord, A. H. Jha, H. Ivison, I. Magnusson, Y. Wang, et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

[4] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

[5] J. Liu, P. Chan, K. Li, and E. J. T. Tchetgen. Regression-based proximal causal inference. *arXiv preprint arXiv:2402.00335*, 2024.

[6] E. J. Tchetgen Tchetgen, A. Ying, Y. Cui, X. Shi, and W. Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

[7] Z. Wood-Doughty, I. Shpitser, and M. Dredze. Challenges of using text classifiers for causal inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4586–4598, 2018.