ESTIMATING MULTI-CAUSE AVERAGE TREATMENT EFFECTS VIA PARTIAL CAUSE INTERVENTION

Hong Gao, Huazhen Lin*
Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics
Chengdu, China
{120071400002, linhz}@smail.swufe.edu.cn

Quanrun Chen
School of Statistics, University of International Business and Economics
Beijing, China
qchen@uibe.edu.cn

Yue Wang
Microsoft Research
Beijing, China
yuwang5@microsoft.com

Wei Chen
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
chenwei2022@ict.ac.cn

ABSTRACT

Treatment effect estimation is crucial for making reliable decisions and avoiding spurious correlations. However, estimating causal effects is harder in limited unbalanced observations, particularly in decision-making systems with multiple causes like healthcare, and politics. In this paper, we aim to enhance the estimation of the multi-cause conditional treatment effect (M-CATE) by augmenting limimted observational data with interventional data to alleviate the data unbalancing. One challenge is that the distribution of interventional data may not be close to the real data. We leverage the causal graph to consider the relationships among causes to solve this. Another challenge is that general identification conditions do not satisfy the realization of intervention. Thereby we give milder partial-cause conditions for identification to construct a Partial Cause Intervention (PCI) algorithm for M-CATE estimation. Specifically, we first intervene in part of the causes once at a time through causal regression which means only modeling the predicted variable using its parent variables, and then we combine the limited observational data with all the interventional data for M-CATE estimation. To support our approach, we prove that the estimation error can be upper bound by the empirical error and the distributional shift among treatments. The experimental results in simulations and real-world data applications validate our approach and theoretical findings.

1 INTRODUCTION

In variants of decision-making tasks, such as healthcare, political management, and IT industry decisions, the treatment effect estimation is crucial to make explainable and reliable decisions by avoiding spurious correlations (Dahabreh et al., 2016; Li et al., 2015; Yu et al., 2022). Usually, feasible actions/treatments in complex decision-making systems are multi-dimensional, outcome

*Correspondence to linhz@swufe.edu.cn

has multiple causes which are related. For example, a researcher may be interested in "If some drugs mutually relieved the patient's symptoms?" Estimating the conditional average *multi-cause* effect from limited observational data can help answer questions and make sound decisions. To ease the presentation, we call the traditional treatment effect estimation with single-cause "single-cause effect estimation." In Figure 1 we provide causal graphs under single-cause and multi-cause settings.

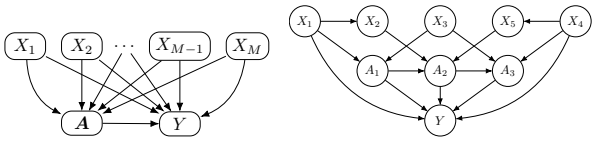


Figure 1: Example for causal graphs in single-cause and multi-cause treatment effect estimation problem. In the single-cause problem, only one cause is considered to affect the outcome and the causal relationships are constructed as triangles: covariates $\rightarrow$ cause, covariates $\rightarrow$ outcome, and cause $\rightarrow$ outcome. In multi-cause problem, outcome Y is influenced by covariates $X_1, \ldots, X_M$ and multiple causes $A_1, \ldots, A_K$. Apart from being affected by covariates, causes are also affected by part of these causes.

In recent years, with the success of deep neural networks (DNN) in important AI tasks like image recognition, machine translation, and speech generation (Li, 2022; Ranathunga et al., 2021; Li et al., 2022), researchers are investigating how to leverage the expressiveness of DNN to enhance the accuracy of treatment effect estimation. However, the limited observational data collected in the past decision-making systems are unbalanced that the covariates distribution is different in treat group and control group.

In the single-cause setting, the problem of unbalanced observations has been improved in many ways (Alaa and van der Schaar, 2017; Künzel et al., 2019; Shalit et al., 2017a; Huling and Mak, 2020; Chen et al., 2022; Wager and Athey, 2018; Lopez and Gutman, 2017). As the number of causes increases, some treatments may be observed only a few times and even not be observed from limited observational samples thus violating the *positivity* assumption in (Rubin, 2005). The multi-cause setting suffers from more complex confounding issues and more severe data unbalancing issues than single-cause. In Figure 1, if causes are binary variables, there will be two possible treatments in a single-cause setting, but 2^K treatments in the K-cause setting. Covariates balance also becomes harder to achieve when samples are sparsely distributed across the treatment groups. Generally in the multi-cause problems, each cause may have different parent covariates and multiple causes will exist causal relationships that affect the distribution of the treatment assignment and potential outcomes. Variational Sample Re-weighting (VSR) (Zou et al., 2020) and Deconfounder (Wang and Blei, 2019) followed this type of causal graph to estimate CATE and derived representation learning among causes. They assumed that the propensity score (PS) is determined by low-dimensional latent variables $\boldsymbol{Z}$, where $\boldsymbol{Z}$ are generated by initial covariates X. Qian et al. (2021) recognized the problem that some treatment groups may have no individuals and also found the traditional triangle causal graph ($\boldsymbol{X} \rightarrow \boldsymbol{A}$, $\boldsymbol{X} \rightarrow Y$ and $\boldsymbol{A} \rightarrow Y$) ignored the influences among the multiple causes. They developed single-cause perturbation (SCP) to augment datasets based on the causal order of the causes but they confused and missed some relationships among covariates, causes, and outcomes.

In this paper, we aim to improve the multi-cause treatment effect estimation by augmenting limited observational data with interventional data to alleviate the data unbalancing issue. When part of the causes are intervened, they may affect both their descendant causes and outcomes. The augmented data brings more combinations among causes than the limited observational data and if the augmented data is by the true distribution derived from the causal graph, the enlarged dataset will be more balanced. One difficulty arises when there is a potential disparity between the distribution of interventional data and the actual data. To address this, we utilize the causal graph to examine the connections among causes by causal regression which means only modeling the predicted variable using its parent variables. First of all, we prove that the multi-cause treatment effect can be identified under "the milder partial-cause conditions" in Proposition 1 and 2. Then, we propose the Partial Cause Intervention (PCI) algorithm Algorithm 1) and we assume there exists a pre-discovered causal graph [1]. For theoretical support of our approach, we prove the upper bound of outcome and causal

[1]There are variants of methods to discover the causal graph (Spirtes and Zhang, 2016; Zeng et al., 2022)

effect estimation error in Theorem 1. We are the first to prove that the estimation error can be upper bounded by the empirical error and the distributional shift (among treatments) over the augmented data. Finally, we conduct experiments on medical and political domains to validate our method.

2 PARTIAL-CAUSE INTERVENTION AND IDENTIFICATION

We focus on M-CATE estimation with K binary causes. The causes $\boldsymbol{A} = (A_1, \ldots, A_K) \in \mathcal{A}$ be a multi-dimansional random variable with sample space $\Omega = \{0,1\}^K$, where A_k is the kth cause. $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^d$ are pre-treatment covariates which are not affected by causes and $\boldsymbol{Y} \in \mathbb{R}$ are observed outcomes affected by parts of pre-treatment covariates and all causes. The observational dataset $\mathcal{D}_0 = \{\boldsymbol{x}_i, y_i, \boldsymbol{a}_i\}_{i\in[N]}$ with N independent samples. We assume the causal relationship between these variables is known. If not, causal recovery methods can be employed to guarantee the true graph is contained in the estimated graph. In the following, we give assumptions and conditions for partial-cause identification as well as multi-cause potential outcomes identification under partial-cause intervention.

Because of the weakness of the traditional graph setting, we consider a precise causal graph model by the partial-cause intervention (PCI). PCI intervenes p causes among all K causes once at a time, so there exist C_K^p combinations to intervene. We denote the sth combination of intervened partial-cause as $\boldsymbol{A}_{s,p} = (A_{s_1}, \ldots, A_{s_p})$, where $s = 1, \ldots, C_K^p$ and $1 \leq s_1 < \cdots < s_p \leq K$. The lower case $\boldsymbol{a}_{s,p}$ is the value of the partial-cause $\boldsymbol{A}_{s,p}$. Based on graph knowledge, we partition the rest of the $K-p$ causes $\boldsymbol{A}_{-s,-p}$ into $\boldsymbol{A}_{s,p}$'s causal descendants $\boldsymbol{A}^{\downarrow}_{-s,-p}$ and their non-descendants $\boldsymbol{A}^{\uparrow}_{-s,-p}$. We denote the cause in descendants $\boldsymbol{A}^{\downarrow}_{-s,-p}$ by A_j. Instead of fit $\boldsymbol{A}^{\downarrow}_{-s,-p}$ jointly, we separately fit every cause A_j in $\boldsymbol{A}^{\downarrow}_{-s,-p}$ by all A_js' causal order using A_j's parent variables $Pa_{\boldsymbol{X}}(A_j)$ and $Pa_{\boldsymbol{A}}(A_j)$ where $Pa_{\boldsymbol{X}}(A_j)$ and $Pa_{\boldsymbol{A}}(A_j)$ means A_j's parent covariates set and parent causes set. Once we intervene in the partial-cause $\boldsymbol{A}_{s,p}$, we want to estimate the potential outcome Y and the elements in $\boldsymbol{A}^{\downarrow}_{-s,-p}$ with the non-descendant causes of $\boldsymbol{A}_{s,p}$ fixed.

For potential outcomes identification, we first make three standard assumptions on partial-cause consistency, unconfoundedness, and positivity: (1) Partial-cause Consistency: $\forall s \leq C_K^p, \forall \boldsymbol{A}_{s,p} \in \{0,1\}^p$ and $\forall A_j \in \boldsymbol{A}^{\downarrow}_{-s,-p}$, if $\boldsymbol{A}_{s,p} = \boldsymbol{a}_{s,p}$, then $Y(\boldsymbol{a}_{s,p}) = Y$ and $A_j(\boldsymbol{a}_{s,p}) = A_j$; (2) Partial-cause Unconfoundedness: $\boldsymbol{A}^{\downarrow}_{-s,-p}(\boldsymbol{a}_{s,p}) \perp\!\!\!\perp \boldsymbol{A}_{s,p} \mid Pa_{\boldsymbol{X}}(\boldsymbol{A}^{\downarrow}_{-s,-p}), \boldsymbol{A}^{\uparrow}_{-s,-p}$ and $Y(\boldsymbol{a}_{s,p}) \perp\!\!\!\perp \boldsymbol{A}_{s,p} | Pa_{\boldsymbol{X}}(Y), \boldsymbol{A}^{\uparrow}_{-s,-p}, \boldsymbol{A}^{\downarrow}_{-s,-p}(\boldsymbol{a}_{s,p})$, $\forall \boldsymbol{a}_{s,p} \in \{0,1\}^p, \forall s \leq C_K^p$; (3) Partial-cause Positivity: $\mathbb{P}(\boldsymbol{A}_{s,p} = \boldsymbol{a}_{s,p} | Pa_{\boldsymbol{X}}(\boldsymbol{A}_{s,p}), \boldsymbol{A}^{\uparrow}_{s,p}) > 0$, $\forall \boldsymbol{a}_{s,p} \in \{0,1\}^p$, if $\mathbb{P}(Pa_{\boldsymbol{X}}(\boldsymbol{A}_{s,p})) \geq 0$. And we assume outcomes and treatments model to be: $Y = g_1(Pa_{\boldsymbol{X}}(Y), \boldsymbol{A}) + \varepsilon$, and $A_k \sim B[g_2(Pa_{\boldsymbol{X}}(A_k), Pa_{\boldsymbol{A}}(A_k)) + \epsilon_k]$ where g_1 and g_2 are unknown functions, ε and ϵ are noises, and $B[]$ denotes Bernoulli distribution.

Under these assumptions, in Proposition 1, we prove that the descendant causes and partial-cause potential outcomes affected by the intervened partial-cause can be identified separately from observational data, precisely, from their parent variables based on the causal graph.

Proposition 1. *Under partial-cause assumptions (1')-(3'), we can identify the $Y(\boldsymbol{a}_{s,p})$ from observational data as: $\forall s \leq C_K^p, \forall \boldsymbol{a}_{s,p} \in \{0,1\}^p$,*

$$\begin{aligned}\mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}\right), \boldsymbol{A}^{\downarrow}_{-s,-p}(\boldsymbol{a}_{s,p}) \mid \boldsymbol{X}, \boldsymbol{A}^{\uparrow}_{-s,-p}\right) =& \mathbb{P}\left(\boldsymbol{A}^{\downarrow}_{-s,-p} \mid Pa_{\boldsymbol{X}}(\boldsymbol{A}^{\downarrow}_{-s,-p}), \boldsymbol{A}^{\uparrow}_{-s,-p}, \boldsymbol{A}_{s,p} = \boldsymbol{a}_{s,p}\right) \\ &\times \mathbb{P}\left(Y \mid Pa_{\boldsymbol{X}}(Y), \boldsymbol{A}^{\uparrow}_{-s,-p}, \boldsymbol{A}^{\downarrow}_{-s,-p}, \boldsymbol{A}_{s,p} = \boldsymbol{a}_{s,p}\right).\end{aligned} \tag{1}$$

Proposition 2. *Under the sequential ignorability assumption (Robins and Greenland, 1992), $\forall s \leq C_K^p$,*

$$\mathbb{E}\left(Y\left(\boldsymbol{a}\right) \mid \mathbf{X}\right) = \mathbb{E}\left(Y\left(\boldsymbol{a}_{s,p}\right) \mid Pa_{\boldsymbol{X}}(Y), \boldsymbol{A}_{-s,-p}\left(\boldsymbol{a}_{s,p}\right) = \boldsymbol{a}_{-s,-p}\right). \tag{2}$$

After the intervention, we enrich the causes that individuals received. Also, the partial-cause positivity must be satisfied within the enlarged datasets for identification. The partial and multi-cause potential outcomes are equal in expectation under appropriate conditioning. With Proposition 2, we can augment the initial dataset by estimating the partial-cause potential outcomes on the r.h.s of equation A.3 and pooling them into one enlarged dataset to estimate the multi-cause potential outcomes on the l.h.s. When estimating the r.h.s of equation 2, the partial-cause potential outcomes

are estimated by parent variables based on the causal graph. The increased sample size mitigates the data scarcity issue and allows the estimator to generalize better.

3 INTERVENTION ON PARTIAL CAUSES FOR M-CATE ESTIMATION

3.1 THE ALGORITHM

To balance the distribution shift, we hope to enlarge the limited observations by intervention. Intervention requires us to specify which cause(s) need to be intervened and how to intervene it/them. Thus we make our approach into three steps and the algorithm is shown in Appendix A.1.

Step One: train causal models Based on causal regression idea that we estimate a variable only using its parents regardless of other variables, for p causes, we train models for $\boldsymbol{A}_{s,p}$'s descendant causes and potential outcomes $Y(\boldsymbol{a}_{s,p})$ on the initial observational data $\mathcal{D}_0$, where $\boldsymbol{a}_{s,p} \in \{0,1\}^p$. For each model, we only input parent variables of the output variable to train [2]:

$$\hat{Y}(\boldsymbol{a}_{s,p}) = f_Y(Pa_{\boldsymbol{X}}(Y), \boldsymbol{A}), \quad \hat{\boldsymbol{A}}^{\downarrow}_{s,p} = f_{\boldsymbol{A}^{\downarrow}_{s,p}}(Pa_{\boldsymbol{X}}(\boldsymbol{A}^{\downarrow}_{s,p}), Pa_{\boldsymbol{A}}(\boldsymbol{A}^{\downarrow}_{s,p})). \tag{3}$$

Step Two: intervene p causes For intervened causes $\boldsymbol{A}_{s,p}$, we specify $\boldsymbol{A}^{\downarrow}_{s,p}$'s parent variables and perturb causes $\boldsymbol{A}_{s,p}$ by setting $\boldsymbol{A}_{s,p} = \mathbf{1} - \boldsymbol{a}_{s,p}$, their opposite values, where $\mathbf{1}$ indicates vectors of length p whose elements are all equal to 1. Then we obtain the potential descendant causes and potential outcomes from the estimated models in step one: $\hat{A}_j = f_{A_j}(Pa_{\boldsymbol{X}}(A_j), Pa_{\boldsymbol{A}_{-s,-p}}(A_j), \boldsymbol{A}_{s,p} = 1 - \boldsymbol{a}_{s,p})$, $\hat{Y}(\mathbf{1} - \boldsymbol{a}_{s,p}) = f_Y(Pa_{\boldsymbol{X}}(Y), \boldsymbol{A}^{\uparrow}_{-s,-p}, \boldsymbol{A}_{s,p} = \mathbf{1} - \boldsymbol{a}_{s,p}, \hat{\boldsymbol{A}}^{\downarrow}_{-s,-p})$, for $A_j \in \boldsymbol{A}^{\downarrow}_{-s,-p}$. Finally, we obtain C^p_K new dataset $\mathcal{D}_1, \ldots, \mathcal{D}_{C^p_K}$ where each interventional dataset contains: (1) (unchanged) $\boldsymbol{X}$, $\boldsymbol{A}_{s,p}$'s non-descendant causes; (2) (changed) $\boldsymbol{A}_{s,p} = \mathbf{1} - \boldsymbol{a}_{s,p}$, $\boldsymbol{A}_{s,p}$'s potential descendant causes and potential outcomes $Y(\mathbf{1} - \boldsymbol{a}_{s,p})$.

Step Three: estimate potential outcome on the augmented dataset. After data augmentation in step two, we can merge observational data $\mathcal{D}_0$ and interventional datasets $\mathcal{D}_1, \ldots, \mathcal{D}_{C^p_K}$ to train final outcome model for potential outcome prediction.

3.2 BOUNDS FOR ESTIMATING POTENTIAL OUTCOMES AND MULTI-CAUSE CATE

With the observational data, we hope to learn a hypothesis $f_{\boldsymbol{a}} : \mathcal{X} \to \mathbb{R}$ which predicts the outcome based on the covariates and multiple causes. To bound the risk of $f_{\boldsymbol{a}}$ on the whole population under multi-cause setting, denoted as $R(f_{\boldsymbol{a}})$, we give Theorem 1 as follows and the details of Theorem 1 are shown in appendix.

Theorem 1. *Assume that weak unconfoundedness (Assumption (2') holds w.r.t $\boldsymbol{X}$. Given samples $(\boldsymbol{x}_1, \boldsymbol{a}_1, y_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{a}_n, y_n) \overset{i.i.d.}{\sim} p(\boldsymbol{X}, \boldsymbol{A}, Y)$ with empirical measure $\hat{p}^n$, let $f_{\boldsymbol{a}}(\boldsymbol{x}) \in \mathcal{H}$ be a hypothesis of $\mathbb{E}_{Y(\boldsymbol{a})|\boldsymbol{X}}[Y(\boldsymbol{a})|\boldsymbol{X} = \boldsymbol{x}]$. With probability at least $1 - 2\delta$,*

$$R(f_{\boldsymbol{a}}) \leq \mathcal{A}(f_{\boldsymbol{a}}) + \mathcal{B}_{\boldsymbol{a}} + \sigma^2_{Y(\boldsymbol{a})}, \tag{4}$$

where $\mathcal{A}(f_{\boldsymbol{a}})$ contains the empirical factual risk on the treatment $\boldsymbol{a}$ and the gap between population and empirical factual risk, $\mathcal{B}_{\boldsymbol{a}}$ contains the empirical distribution distance and the gap between its population and empirical form, $\sigma^2_{Y(\boldsymbol{a})}$ is the expected variance in $Y(\boldsymbol{a})$.

4 EXPERIMENTS

4.1 SIMULATION STUDY

Data setting and Benchmarks Each dataset contains 500 samples for training, 200 samples for validation, and 500 samples for testing. The model that data generation follows is shown in ap-

[2] Some work (VanderWeele, 2019) highlighted that when using instrumental variables can worsen outcome prediction because instrumental variables have no edges to outcomes. For treatment models, adjustment variables on outcome prediction are unnecessary to remove bias but can reduce variance in treatment effect estimation (Sauer et al., 2013; Kuang et al., 2019)

pendix. We evaluate the models using RMSE on all potential outcomes and PEHE. We use Wasserstein distance to measure the balance of covariates and achieving smaller distance is more balanced. We included seven benchmarks to compare with our method. As a baseline, we used covariates adjustment with feed-forward neural networks (NN). We compared with VSR and Deconfounder (DEC) (Zou et al., 2020; Wang and Blei, 2019). We also included Counterfactual Regression (CFR) and DR-CFR (Shalit et al., 2017b; Hassanpour and Greiner, 2020), the propensity score (NN-IPW) and overlap score (OP) methods from the ATE literature (Hirano et al., 2003; Li and Li, 2019) as well as SCP(Single-cause-perturbation)(Qian et al., 2021) which also used data augmentation.

Table 1: Prediction error on treatment effect and data balancing on simulation data

	PEHE (std)		
Method	$K = 2$	5	7
NN	0.18 (.006)	0.29 (.016)	6.48 (.056)
NN_IPW	0.20 (.004)	0.37 (.021)	7.88 (.210)
OP	0.20 (.005)	0.43 (.020)	15.1 (.500)
VSR	0.25 (.037)	0.28 (.016)	8.15 (.068)
DEC	0.28 (.026)	0.25 (.012)	4.99 (.033)
CFR	0.15 (.006)	0.51 (.023)	14.5 (.523)
DR-CFR	0.18 (.008)	0.77 (.034)	14.3 (.516)
SCP	0.12 (.008)	0.15 (.008)	4.77 (.062)
PCI (Ours)	**0.06 (.002)**	**0.10 (.015)**	**1.06 (.024)**
	Wasserstein distance ($\text{IPM}_{\mathcal{L}}$)		
Dataset	$K = 2$	5	7
Dataset0	13.5	177	1947
SCP	6.26	145	1871
PCI (Ours)	**5.96**	**137**	**1810**

Results Based on the validation error, we choose the intervened partial-cause number to be $p = 1$ when cause number $K = 2, 3, 7$ and $p = 2$ when $K = 5$. As shown in Table 1, Our PCI method gains the smallest RMSE and PEHE and consistently outperforms the benchmarks across the different number of causes K with covariate dimensionality $d = 4 \times K$. The performance gain becomes more pronounced as the number of causes increases, e.g. $K = 7$. Based on Theorem 1, we compare the covariate distribution shift among different treatment groups in the training dataset as Wasserstein distance. Our method has the smallest Wasserstein distance thus it achieves more balance in covariate distribution. We also conduct more experiments in A.3 to explore why our method has superior achievement than SCP.

4.2 Experiment on ADNI dataset

ADNI Dataset Our analysis is based on data from the Alzheimer's Disease Neuroimaging Initiative, including 2129 subjects with 67% of them used as the training size. The outcome is the Mini-Mental State Examination score and confounding variables include APOE4, age, sex, marital status, years of education, and years of retirement. Treatments measure the atrophy of three brain regions. We assume that the outcomes Y follow the model (A.13) and learn the causal DAG among covariates, causes, and outcomes (Figure 6).

Results The causal graph in Figure 6 shows that shrinkage of the three regions has a significant effect on the cognitive score, which coincides with the hypothesis that the atrophy of those regions is among the most significant biomarkers of Alzheimer's disease. In Table 5, we compare the proposed and other methods in two settings: semi-synthetic data (generate outcomes Y based on (8)) and real data (real outcomes Y). In the semi-synthetic and real-world setting, our method gains the smallest RMSE on all potential outcome predictions and the lowest PEHE. We also conduct experiment on Vdem dataset to validate our approach in A.5.

References

Alaa, A. M. and van der Schaar, M. (2017). Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in Neural Information Processing Systems*, 30.

Chen, K., Yin, Q., and Long, Q. (2022). Covariate-balancing-aware interpretable deep learning models for treatment effect estimation. *arXiv preprint arXiv:2203.03185*.

Choo, I. H., Lee, D. Y., Oh, J. S., Lee, J. S., Lee, D. S., Song, I. C., Youn, J. C., Kim, S. G., Kim, K. W., Jhoo, J. H., et al. (2010). Posterior cingulate cortex atrophy and regional cingulum disruption in mild cognitive impairment and alzheimer's disease. *Neurobiology of aging*, 31(5):772–779.

Dahabreh, I. J., Hayward, R., and Kent, D. M. (2016). Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *International journal of epidemiology*, 45(6):2184–2193.

Gainotti, G., Barbier, A., and Marra, C. (2003). Slowly progressive defect in recognition of familiar people in a patient with right anterior temporal atrophy. *Brain*, 126(4):792–803.

Hassanpour, N. and Greiner, R. (2020). Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Huling, J. D. and Mak, S. (2020). Energy balancing of covariate distributions. *arXiv preprint arXiv:2004.13962*.

Iizuka, T. and Kameyama, M. (2016). Cingulate island sign on fdg-pet is associated with medial temporal lobe atrophy in dementia with lewy bodies. *Annals of nuclear medicine*, 30(6):421–429.

Johansson, F. D., Shalit, U., Kallus, N., and Sontag, D. (2020). Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*.

Kuang, K., Cui, P., Li, B., Jiang, M., Wang, Y., Wu, F., and Yang, S. (2019). Treatment effect estimation via differentiated confounder balancing and regression. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(1):1–25.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

Li, F. and Li, F. (2019). Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 13(4):2389 – 2415.

Li, L., Chen, S., Kleban, J., and Gupta, A. (2015). Counterfactual estimation and optimization of click metrics in search engines: A case study. In *Proceedings of the 24th International Conference on World Wide Web*, pages 929–934.

Li, S., Li, J., Liu, Q., and Gong, Z. (2022). Adversarial speech generation and natural speech recovery for speech content protection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7291–7297.

Li, Y. (2022). Research and application of deep learning in image recognition. In *2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)*, pages 994–999. IEEE.

Lopez, M. J. and Gutman, R. (2017). Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, pages 432–454.

Qian, Z., Curth, A., and van der Schaar, M. (2021). Estimating multi-cause treatment effects via single-cause perturbation. *Advances in Neural Information Processing Systems*, 34.

Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M., and Kaur, R. (2021). Neural machine translation for low-resource languages: A survey. *arXiv preprint arXiv:2106.15115*.

Reiman, E. M., Uecker, A., Caselli, R. J., Lewis, S., Bandy, D., De Leon, M. J., De Santi, S., Convit, A., Osborne, D., Weaver, A., et al. (1998). Hippocampal volumes in cognitively normal persons at genetic risk for alzheimer's disease. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 44(2):288–291.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155.

Rohrer, J. D., Ridgway, G. R., Modat, M., Ourselin, S., Mead, S., Fox, N. C., Rossor, M. N., and Warren, J. D. (2010). Distinct profiles of brain atrophy in frontotemporal lobar degeneration caused by progranulin and tau mutations. *Neuroimage*, 53(3):1070–1076.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Sauer, B. C., Brookhart, M. A., Roy, J., and VanderWeele, T. (2013). A review of covariate selection for non-experimental comparative effectiveness research. *Pharmacoepidemiology and drug safety*, 22(11):1139–1145.

Schröder, J. and Pantel, J. (2016). Neuroimaging of hippocampal atrophy in early recognition of alzheimer´ s disease–a critical appraisal after two decades of research. *Psychiatry Research: Neuroimaging*, 247:71–78.

Shalit, U., Johansson, F. D., and Sontag, D. (2017a). Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR.

Shalit, U., Johansson, F. D., and Sontag, D. (2017b). Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3076–3085. JMLR.org.

Spirtes, P. and Zhang, K. (2016). Causal discovery and inference: concepts and recent methodological advances. *Applied informatics*, 3(1):1–28.

VanderWeele, T. J. (2019). Principles of confounder selection. *European journal of epidemiology*, 34(3):211–219.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wang, Y. and Blei, D. M. (2019). The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596.

Yu, Y., Chen, H., Peng, C.-H., and Chau, P. Y. (2022). The causal effect of subscription video streaming on dvd sales: Evidence from a natural experiment. *Decision Support Systems*, 157:113767.

Zeng, Y., Shimizu, S., Matsui, H., and Sun, F. (2022). Causal discovery for linear mixed data. In *Conference on Causal Learning and Reasoning*, pages 994–1009. PMLR.

Zou, H., Cui, P., Li, B., Shen, Z., Ma, J., Yang, H., and He, Y. (2020). Counterfactual prediction for bundle treatment. *Advances in Neural Information Processing Systems*, 33:19705–19715.

A APPENDIX

A.1 ALGORITHM DETAILS FOR PARTIAL CAUSE INTERVENTION (PCI)

In the main article, we summarize our approach to a simplified algorithm. Here we show the details of the algorithm.

Algorithm 1 Partial Cause Intervention (PCI)

Input:

Observational dataset $\mathcal{D}_0 = \{x_i, y_i, \boldsymbol{a}_i\}_{i\in[\mathcal{N}_0]}$; Causal regression algorithm f; Tuning interval for the number of intervened causes p;

for each $p \in 1, \ldots, M (M \leq K)$ **do**
  for each $s \in 1, \ldots, C_K^p$ **do**
    Fit $\hat{f}_{A_j(\boldsymbol{a}_{s,p})}(\cdot)$ to estimate $A_j(\boldsymbol{a}_{s,p})$ where $A_j \in \boldsymbol{A}_{-s,-p}^{\downarrow}$ upon the inputs $Pa(A_j)$ using the observed data $\mathcal{D}_0$;
    Fit $\hat{f}_{Y(\boldsymbol{a}_{s,p})}(\cdot)$ to estimate $Y(\boldsymbol{a}_{s,p})$ upon the inputs $Pa(Y)$ using the observed data $\mathcal{D}_0$;
    Initialize s^{th} interventional dataset; $\mathcal{D}_s = \emptyset$,
    for each $i \in 1, \ldots, N_0$ **do**
      Intervene the p causes: $\boldsymbol{a}'_{i,s,p} = 1 - \boldsymbol{a}_{i,s,p}$;
      Set $\boldsymbol{a}_{i,-s,-p}^{\uparrow}(\boldsymbol{a}'_{i,s,p}) = \boldsymbol{a}_{i,-s,-p}^{\uparrow}$;
      Set $a_{i,j}(\boldsymbol{a}'_{i,s,p}) = \hat{f}_{A_j(\boldsymbol{a}_{s,p})}(Pa(A_{i,j}))$,
      where $A_j \in \boldsymbol{A}_{-s,-p}^{\downarrow}$, $Pa(A_j)$ may contain some confounders and causes which are parents of a_j;
      Set $\tilde{y}_{i,s,p} := y(\boldsymbol{a}'_{i,s,p}) = \hat{f}_{Y(\boldsymbol{a}_{s,p})}(Pa(Y_i))$,
      where $Pa(Y_i)$ may contain some of confounders X_i, $\boldsymbol{a}_{i,-s,-p}^{\uparrow}$, perturbed cause $\boldsymbol{a}_{i,s,p}$ and $a_{i,j}(\boldsymbol{a}'_{i,s,p})$ which are parents of outcome Y_i;
      Combine the causes $\tilde{\boldsymbol{a}}_{i,s,p} := (\boldsymbol{a}'_{i,s,p}, \boldsymbol{a}_{i,-s,-p}(\boldsymbol{a}'_{i,s,p}))$;
      Add new data point $(\boldsymbol{x}_i, \tilde{y}_{i,s,p}, \tilde{\boldsymbol{a}}_{i,s,p})$ to $\mathcal{D}_s$;
    end for
  end for
  Obtain the augmented training data $\mathcal{D}^{Tr} = \mathcal{D}_s$, $s \in \{0, \ldots, K\}$;
  Train $f_{p,\theta}$ to estimate $\mathbb{E}[Y|\mathbf{X}, \mathbf{A}]$ upon $Pa(Y)$ using $\mathcal{D}^{Tr}$;
end for
Compare the validation error across different intervened partial-cause number p, choose the best p^*;

Output:

A trained multi-cause potential outcomes predictor $f_{p^*,\theta}$;

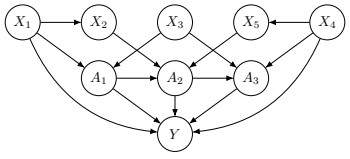


Figure 2: Example for causal graphs in multi-cause treatment effect estimation problem.

To detail our estimation procedure, we demonstrate an example with $p = 1$ shown in Figure 2. When intervening one cause A_1, we need to estimate its descendant causes A_2, A_3, and outcome Y; when intervening A_2, we need to estimate A_3 and Y; when intervening A_3, we need to estimate Y. We take the intervention on A_1 as an example to show how we estimate multi-cause CATE with causal structure knowledge: (1) train three models to fit A_2, A_3 and Y with their parent variables through causal regression and we can use any regression algorithm for these three models; (2) intervene on A_1 by setting it equal to its opposite value $1 - a_k$, then predict A_1's descendant causes (A_2 and A_3) and outcomes Y to obtain the interventional dataset $\mathcal{D}_1$; (3) Add the interventional dataset $\mathcal{D}_1$ to the observational dataset $\mathcal{D}_0$. After intervention on those three causes, we use the augmented

datasets $\{\mathcal{D}_0, \ldots, \mathcal{D}_3\}$ to fit a potential outcome model with the input $\{X_1, X_4, A_1, A_2, A_3\}$ and then estimate the multi-cause CATE.

A.2 THEORETICAL RESULTS

In this section, we detail the proofs of Proposition 1, Proposition 2 and Theorem 1.

A.2.1 PROOF OF PROPOSITION 1 AND 2

Proposition 1. *Under partial-cause assumptions (1')-(3'), we can identify the $Y(\boldsymbol{a}_{s,p})$ from observational data as: $\forall s \leq C_K^p, \forall \boldsymbol{a}_{s,p} \in \{0,1\}^p$,*

$$\begin{aligned} &\mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}\right), \boldsymbol{A}_{-s,-p}^{\downarrow}(\boldsymbol{a}_{s,p}) \mid \boldsymbol{X}, \boldsymbol{A}_{-s,-p}^{\uparrow}\right) \\ =&\mathbb{P}\left(\boldsymbol{A}_{-s,-p}^{\downarrow} \mid Pa_{\boldsymbol{X}}(\boldsymbol{A}_{-s,-p}^{\downarrow}), \boldsymbol{A}_{-s,-p}^{\uparrow}, \boldsymbol{A}_{s,p} = \boldsymbol{a}_{s,p}\right) \\ &\times \mathbb{P}\left(Y \mid Pa_{\boldsymbol{X}}(Y), \boldsymbol{A}_{-s,-p}^{\uparrow}, \boldsymbol{A}_{-s,-p}^{\downarrow}, \boldsymbol{A}_{s,p} = \boldsymbol{a}_{s,p}\right). \end{aligned} \tag{A.1}$$

Equation A.1 decomposes the joint estimation of outcomes and descendant causes into two separate estimation tasks. When we intervene the partial-cause $\boldsymbol{A}_{s,p}$, their descendant causes and outcomes would be affected accordingly. The l.h.s of equation A.1 jointly estimate outcomes and descendant causes On the r.h.s of equation A.1, we separately estimate descendant causes of the partial-cause by their parent nodes: $Pa_{\boldsymbol{X}}(\boldsymbol{A}_{-s,-p}^{\downarrow}), \boldsymbol{A}_{-s,-p}^{\uparrow}, \boldsymbol{A}_{s,p}$, and also estimate outcomes Y by parent nodes: $Pa_{\boldsymbol{X}}(Y), \boldsymbol{A}_{-s,-p}^{\uparrow}, \boldsymbol{A}_{-s,-p}^{\downarrow}, \boldsymbol{A}_{s,p}$. After the intervention, we can identify the affected variables from observational data and intervene partial-cause.

Proof of Proposition 1. Note that the partial-cause unconfoundedness assumption implies the following two equations by the properties of conditional independence,

$$\begin{aligned} \boldsymbol{A}_{-s,-p}^{\downarrow}\left(\boldsymbol{a}_{s,p}\right) &\perp\!\!\!\perp \boldsymbol{A}_{s,p} \mid Pa_{\boldsymbol{X}}(\boldsymbol{A}_{-s,-p}^{\downarrow}), \boldsymbol{A}_{-s,-p}^{\uparrow}, \\ Y(\boldsymbol{a}_{s,p}) &\perp\!\!\!\perp \boldsymbol{A}_{s,p} \mid Pa_{\boldsymbol{X}}(Y), \boldsymbol{A}_{-s,-p}^{\uparrow}, \boldsymbol{A}_{-s,-p}^{\downarrow}\left(\boldsymbol{a}_{s,p}\right). \end{aligned}$$

$\mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}\right), \boldsymbol{A}_{-s,-p}^{\downarrow}(\boldsymbol{a}_{s,p}) \mid \boldsymbol{X}, \boldsymbol{A}_{-s,-p}^{\uparrow}\right)$ can be decomposed into two terms by Bayes rule:

$$\begin{aligned} &\mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}\right), \boldsymbol{A}_{-s,-p}^{\downarrow}(\boldsymbol{a}_{s,p}) \mid \boldsymbol{X}, \boldsymbol{A}_{-s,-p}^{\uparrow}\right) \\ =&\mathbb{P}\left(\boldsymbol{A}_{-s,-p}^{\downarrow}\left(\boldsymbol{a}_{s,p}\right) \mid \boldsymbol{X}, \boldsymbol{A}_{-s,-p}^{\uparrow}\right) \cdot \mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}\right) \mid \boldsymbol{X}, \boldsymbol{A}_{-s,-p}^{\uparrow}, \boldsymbol{A}_{-s,-p}^{\downarrow}\left(\boldsymbol{a}_{s,p}\right)\right). \end{aligned} \tag{A.2}$$

Because partial-cause positivity $\mathbb{P}\left(\boldsymbol{A}_{s,p} \mid Pa_{\boldsymbol{X}}(\boldsymbol{A}_{-s,-p}^{\downarrow}), \boldsymbol{A}_{-s,-p}^{\uparrow}\right) > 0$, we can seperately treat $\boldsymbol{A}_{-s,-p}^{\downarrow}$ and Y as outcome. Invoking the standard identification theory, we obtain

$$\begin{aligned} \mathbb{P}\left(\boldsymbol{A}_{-s,-p}^{\downarrow}\left(\boldsymbol{a}_{s,p}\right) \mid \boldsymbol{X}, \boldsymbol{A}_{-s,-p}^{\uparrow}\right) &= \mathbb{P}\left(\boldsymbol{A}_{-s,-p}^{\downarrow}\left(\boldsymbol{a}_{s,p}\right) \mid Pa_{\boldsymbol{X}}(\boldsymbol{A}_{-s,-p}^{\downarrow}), \boldsymbol{A}_{-s,-p}^{\uparrow}, \boldsymbol{A}_{s,p}\right), \\ \mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}\right) \mid \boldsymbol{X}, \boldsymbol{A}_{-s,-p}^{\uparrow}, \boldsymbol{A}_{-s,-p}^{\downarrow}\left(\boldsymbol{a}_{s,p}\right)\right) &= \mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}\right) \mid Pa_{\boldsymbol{X}}(Y), \boldsymbol{A}_{-s,-p}^{\uparrow}, \boldsymbol{A}_{-s,-p}^{\downarrow}\left(\boldsymbol{a}_{s,p}\right), \boldsymbol{A}_{s,p}\right). \end{aligned}$$

The consistency assumption states that when $\boldsymbol{A}_{s,p} = \boldsymbol{a}_{s,p}, \mathbf{A}_{-s,-p}^{\downarrow}\left(\boldsymbol{a}_{s,p}\right) = \mathbf{A}_{-s,-p}^{\downarrow}\left(\boldsymbol{A}_{s,p}\right) = \mathbf{A}_{-s,-p}^{\downarrow}$. Hence, we can replace $\mathbf{A}_{-s,-p}^{\downarrow}\left(\boldsymbol{a}_{s,p}\right)$ on the right hand side with $\mathbf{A}_{-s,-p}^{\downarrow}$ which concludes the proof. □

Proposition 2. *Under the sequential ignorability assumption (Robins and Greenland, 1992), $\forall s \leq C_K^p$,*

$$\begin{aligned} &\mathbb{E}\left(Y\left(\boldsymbol{a}\right) \mid \mathbf{X}\right) \\ =&\mathbb{E}\left(Y\left(\boldsymbol{a}_{s,p}\right) \mid Pa_{\boldsymbol{X}}(Y), \boldsymbol{A}_{-s,-p}\left(\boldsymbol{a}_{s,p}\right) = \boldsymbol{a}_{-s,-p}\right). \end{aligned} \tag{A.3}$$

Proof of Proposition 2. We start by recognizing the right hand side of the Equation (4) follows

$$\begin{aligned}
&\mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}\right) \mid Pa_{\boldsymbol{X}}(Y), \boldsymbol{A}_{-s,-p}\left(\boldsymbol{a}_{s,p}\right)=\boldsymbol{a}_{-s,-p}\right) \\
=&\mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}\right) \mid Pa_{\boldsymbol{X}}(Y) \cap Pa_{\boldsymbol{X}}(\boldsymbol{A}), \boldsymbol{A}_{-s,-p}\left(\boldsymbol{a}_{s,p}\right)=\boldsymbol{a}_{-s,-p}\right) \\
=&\mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}, \mathbf{A}_{-s,-p}\left(\boldsymbol{a}_{s,p}\right)\right) \mid Pa_{\boldsymbol{X}}(Y) \cap Pa_{\boldsymbol{X}}(\boldsymbol{A}), \boldsymbol{A}_{-s,-p}\left(\boldsymbol{a}_{s,p}\right)=\boldsymbol{a}_{-s,-p}\right) \\
=&\mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}, \mathbf{a}_{-s,-p}\right) \mid Pa_{\boldsymbol{X}}(Y) \cap Pa_{\boldsymbol{X}}(\boldsymbol{A}), \boldsymbol{A}_{-s,-p}\left(\boldsymbol{a}_{s,p}\right)=\boldsymbol{a}_{-s,-p}\right),
\end{aligned} \tag{A.4}$$

$\forall \boldsymbol{a}_{s,p} \in \{0,1\}^p$ and $\boldsymbol{a}_{-s,-p} \in \{0,1\}^{K-p}$. Furthermore, we have for all $\boldsymbol{a}_{-s,-p} \in \{0,1\}^{K-p}$

$$\mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}, \mathbf{a}_{-s,-p}\right) \mid Pa_{\boldsymbol{X}}(Y) \cap Pa_{\boldsymbol{X}}(\boldsymbol{A}), \boldsymbol{A}_{-s,-p}\left(\boldsymbol{a}_{s,p}\right)=\boldsymbol{a}_{-s,-p}\right)=\mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}, \boldsymbol{a}_{-s,-p}\right) \mid \boldsymbol{X}\right).$$

Combining the previous two equations, we immediately see that

$$\mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}, \boldsymbol{a}_{-s,-p}\right) \mid \mathbf{X}\right)=\mathbb{P}\left(Y\left(\boldsymbol{a}_{s,p}\right) \mid Pa_{\boldsymbol{X}}(Y), \boldsymbol{A}_{-s,-p}\left(\boldsymbol{a}_{s,p}\right)=\boldsymbol{a}_{-s,-p}\right),$$

$\forall \boldsymbol{a}_{s,p} \in \{0,1\}^p$ and $\boldsymbol{a}_{-s,-p} \in \{0,1\}^{K-p}$ which concludes the proof. □

A.2.2 Details and proof of Theorem 1

In this section, we give the complete formulation of Theorem 1. Under multi-cause setting that $\boldsymbol{a}, \boldsymbol{a}' \in \{0,1\}^K$, the multi-cause population risk R would be decomposed similarly:

$$R\left(f_{\boldsymbol{a}}\right)=\pi_{\boldsymbol{a}} \underbrace{R_{\boldsymbol{a}}(f_{\boldsymbol{a}})}_{\text{Observable}}+\sum_{\substack{\boldsymbol{a}' \in\{0,1\}^K \\ \boldsymbol{a}' \neq \boldsymbol{a}}} \pi_{\boldsymbol{a}'} \underbrace{R_{\boldsymbol{a}'}\left(f_{\boldsymbol{a}}\right)}_{\text{Unobserved}}, \tag{A.5}$$

where the factual risk $R_{\boldsymbol{a}}(f_{\boldsymbol{a}})$ means the expected error for the population if all the individuals were assigned to the treatment $\boldsymbol{a}$, and the counterfactual risk $R_{\boldsymbol{a}'}(f_{\boldsymbol{a}})$ means the risk of the individuals with treatment $\boldsymbol{a}$ which have not received the treatment $\boldsymbol{a}$ and $\pi_{\boldsymbol{a}}=p(\boldsymbol{A}=\boldsymbol{a})$.

To bound the risk of $f_{\boldsymbol{a}}$ on the whole population, it is sufficient for us to bound the counterfactual risk $R_{\boldsymbol{a}'}\left(f_{\boldsymbol{a}}\right)$ and estimate $R_{\boldsymbol{a}}(f_{\boldsymbol{a}})$ empirically. $p_{\boldsymbol{a}}(x)=p(\boldsymbol{X}=\boldsymbol{x} \mid \boldsymbol{A}=\boldsymbol{a})$ is covariates distribution with treatment $\boldsymbol{a}$, also denoted as $p_{\boldsymbol{a}}$. The empirical weighted factual risk is defined as $\hat{R}_{\boldsymbol{a}}\left(f_{\boldsymbol{a}}\right):=\frac{1}{n_{\boldsymbol{a}}} \sum_{i:\boldsymbol{a}_i=\boldsymbol{a}} L\left(f_{\boldsymbol{a}}\left(x_i\right), y_i\right)$ where $n_{\boldsymbol{a}}$ is the number of individuals with treatment $\boldsymbol{a}$. We give Theorem 1 to bound the population risk.

Theorem 1. *Assume that weak unconfoundedness (Assumption (2') holds w.r.t $\boldsymbol{X}$. Given samples $(\boldsymbol{x}_1, \boldsymbol{a}_1, y_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{a}_n, y_n) \overset{i.i.d.}{\sim} p(\boldsymbol{X}, \boldsymbol{A}, Y)$ with empirical measure $\hat{p}^n$, and $n_{\boldsymbol{a}}:=\sum_{i=1}^n \mathbb{1}(\boldsymbol{a}_i=\boldsymbol{a})$ for $\boldsymbol{a} \in \{0,1\}^K$. Let $f_{\boldsymbol{a}}(\boldsymbol{x}) \in \mathcal{H}$ be a hypothesis of $\mathbb{E}_{Y(\boldsymbol{a})|\boldsymbol{X}}[Y(\boldsymbol{a}) \mid \boldsymbol{X}=\boldsymbol{x}]$ and $\ell_{f_{\boldsymbol{a}}}(\boldsymbol{x}):=\mathbb{E}_{Y|\boldsymbol{X}}\left[L\left(f_{\boldsymbol{a}}(\boldsymbol{x}), Y(\boldsymbol{a})\right) \mid \boldsymbol{X}=\boldsymbol{x}\right]$ where $L\left(y, y'\right)=\left(y-y'\right)^2$. With probability at least $1-2\delta$,*

$$R\left(f_{\boldsymbol{a}}\right) \leq \mathcal{A}(f_{\boldsymbol{a}})+\mathcal{B}_{\boldsymbol{a}}+\sigma^2_{Y(\boldsymbol{a})}, \tag{A.6}$$

where $\mathcal{A}(f_{\boldsymbol{a}})$ contains the empirical factual risk and the gap between population and empirical factual risk, $\mathcal{B}_{\boldsymbol{a}}$ contains the empirical distribution distance and the gap between its population and empirical form, $\sigma^2_{Y(\boldsymbol{a})}$ is the expected variance in $Y(\boldsymbol{a})$.

In equation (A.6) of Theorem 1, $\mathcal{A}(f_{\boldsymbol{a}}) = \frac{1}{n_{\boldsymbol{a}_0}+n_{\boldsymbol{a}_+}}\Big(\sum_{i:\boldsymbol{a}_i=\boldsymbol{a}}^{n_{\boldsymbol{a}_0}} L\left(f_{\boldsymbol{a}}\left(x_i\right), y_i\right) + \sum_{\substack{i>n_{\boldsymbol{a}_0} \\ \boldsymbol{a}_i=\boldsymbol{a}}}^{n_{\boldsymbol{a}_0}+n_{\boldsymbol{a}_+}} L\left(f_{\boldsymbol{a}}\left(x_i\right), \tilde{y}_i\right) + \sum_{i>n_{\boldsymbol{a}_0} \boldsymbol{a}_i=\boldsymbol{a}}^{n_{\boldsymbol{a}_0}+n_{\boldsymbol{a}_+}} L\left(\tilde{y}_i, y_i\right)\Big) + \frac{1}{\left(n_{\boldsymbol{a}_0}+n_{\boldsymbol{a}_+}\right)^{3/8}} V_{p_{\boldsymbol{a}}}\left(\ell_{f_{\boldsymbol{a}}}\right) C^{\mathcal{H}}_{n_{\boldsymbol{a}_0}+n_{\boldsymbol{a}_+},\delta}$, where $V_{p_{\boldsymbol{a}}}\left(\ell_{f_{\boldsymbol{a}}}\right)=\max\left(\sqrt{\mathbb{E}_{p_{\boldsymbol{a}}}\left[\ell^2_{f_{\boldsymbol{a}}}\right]}, \sqrt{\mathbb{E}_{\hat{p}_{\boldsymbol{a}}}\left[\ell^2_{f_{\boldsymbol{a}}}\right]}\right)$, $C^{\mathcal{H}}_{n,\delta}$ is a function of the pseudo-dimension of $\mathcal{H}$ and $\tilde{y}_i$ denotes the augmented outcomes. On the r.h.s of $\mathcal{A}(f_{\boldsymbol{a}})$, (1) the first term is prediction error under observervational data $\mathcal{D}_0$ where $n_{\boldsymbol{a}_0}$ is the number of individuals with treatment $\boldsymbol{a}$ which have received the treatment $\boldsymbol{a}$ in dataset $\mathcal{D}_0$; (2) prediction error under augmented data where $n_{\boldsymbol{a}_+}$ is the augmented number of individuals with treatments $\boldsymbol{a}$ which have not received the treatment $\boldsymbol{a}$ in $\mathcal{D}_0$; (3) the third term the loss between y_i and $\tilde{y}_i$ is the augmented error which measures the distance between true potential outcomes and augmented potential outcomes.

We denote $\mathcal{B}_{\boldsymbol{a}} = B \sum_{\substack{\boldsymbol{a}' \in \{0,1\}^K \\ \boldsymbol{a}' \neq \boldsymbol{a}}} \pi_{\boldsymbol{a}'} \operatorname{IPM}_{\mathcal{L}}(\hat{p}_{\boldsymbol{a}}, \hat{p}_{\boldsymbol{a}'}) + \sum_{\substack{\boldsymbol{a}' \in \{0,1\}^K \\ \boldsymbol{a}' \neq \boldsymbol{a}}} \pi_{\boldsymbol{a}'} \mathcal{D}_{\delta}^{\mathcal{L}} \left(\frac{1}{\sqrt{n_{\boldsymbol{a}_0} + n_{\boldsymbol{a}_+}}} + \frac{1}{\sqrt{n_{\boldsymbol{a}'_0} + n_{\boldsymbol{a}'_+}}} \right)$, where $\pi_{\boldsymbol{a}} = \mathbb{P}(\boldsymbol{A} = \boldsymbol{a})$, $\mathcal{D}_{\delta}^{\mathcal{L}}$ is a function of the kernel norm of $\mathcal{L}$ (see lemma 3). Assume that there exists a constant $B > 0$ such that $\ell_{f_{\boldsymbol{a}}}(\boldsymbol{x})/B \in \mathcal{L}$, where $\mathcal{L}$ is a reproducing kernel Hilbert space of a kernel k such that $k(\boldsymbol{x}, \boldsymbol{x}) < \infty$. IPM is the integral probability metric that is used to measure the distance between two distributions. Empirical IPM estimation is a measure for empirical covariates distributions under different treatment groups $\hat{p}_{\boldsymbol{a}}$, and its population form is $p_{\boldsymbol{a}}(x) = \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{A} = \boldsymbol{a})$. When $\mathcal{L}$ is the family of functions with Lipschitz constant at most 1 and $\operatorname{IPM}_{\mathcal{L}}$ the Wasserstein distance. The term $\operatorname{IPM}_{\mathcal{L}}(\hat{p}_{\boldsymbol{a}}, \hat{p}_{\boldsymbol{a}'})$ is used to bound the counterfactual risk jointly with the factual risk. When the conditional probability $p_{\boldsymbol{a}}(x)$ is modeled by the causal graph, $\mathbb{P}(\boldsymbol{X}|\boldsymbol{A}) = \mathbb{P}(\boldsymbol{X}, \boldsymbol{A}) / \int_{\mathcal{X}} \mathbb{P}(\boldsymbol{X}, \boldsymbol{A}) d\boldsymbol{X}$ where $\mathbb{P}(\boldsymbol{X}, \boldsymbol{A}) = \mathbb{P}(\boldsymbol{X}) \prod_{k=1}^{K} \mathbb{P}(A_k \mid parents(A_k))$ based on Local Markov assumption (A variable is independent of its nondescendants given its parents (only the parents)). The last term at r.h.s of $\mathcal{B}_{\boldsymbol{a}}$ demonstrates the gap between population and empirical distribution distance. The gap will decline as the sample size increases. And the larger the augmented samples, the lower the gap.

Before the proof of Theorem 1, we illustrate some definitions used in Theorem 1. Formally, the expected *pointwise* loss of a hypothesis $f_{\boldsymbol{a}}$ at a point x is defined as $\ell_{f_{\boldsymbol{a}}}(\boldsymbol{x}) := \mathbb{E}_{Y(\boldsymbol{a})|\boldsymbol{X}}[L(Y(\boldsymbol{a}), f_{\boldsymbol{a}}(\boldsymbol{x})) \mid \boldsymbol{X} = \boldsymbol{x}]$, where $L(\cdot, \cdot)$ is the error function (e.g. square error).

The marginal risk of a hypothsis $f_{\boldsymbol{a}}$ w.r.t. a population $p(\boldsymbol{X})$ is defined as $R(f_{\boldsymbol{a}}) := \mathbb{E}_{\boldsymbol{X}}[\ell_{f_{\boldsymbol{a}}}(\boldsymbol{X})]$. $R(f_{\boldsymbol{a}})$ means the expected error for the population if all the individuals were assigned to the treatment $\boldsymbol{a}$. $R(f_{\boldsymbol{a}})$ is the combination of the factual risk and counterfactual risk because not all the individuals received the same treatment $\boldsymbol{a}$. The factual risk of $f_{\boldsymbol{a}}$ w.r.t. treatment group $p(\boldsymbol{X} \mid \boldsymbol{A} = \boldsymbol{a})$ is $R_{\boldsymbol{a}}(f_{\boldsymbol{a}}) := \mathbb{E}_{\boldsymbol{X}|\boldsymbol{A}}[\ell_{f_{\boldsymbol{a}}}(\boldsymbol{X}) \mid \boldsymbol{A} = \boldsymbol{a}]$, means the risk of the individuals with treatment $\boldsymbol{a}$ which have received the treatment $\boldsymbol{a}$. The counterfactual risk is $R_{\boldsymbol{a}'}(f_{\boldsymbol{a}}) := \mathbb{E}_{\boldsymbol{X}|\boldsymbol{A}}[\ell_{f_{\boldsymbol{a}}}(\boldsymbol{X}) \mid \boldsymbol{A} = \boldsymbol{a}']$, $\boldsymbol{a}' \neq \boldsymbol{a}$ and $\boldsymbol{a}' \in \{0,1\}^K$, means the risk of the individuals with treatment $\boldsymbol{a}$ which have not received the treatment $\boldsymbol{a}$. Therefore, our target is minimizing $R(f_{\boldsymbol{a}})$.

The factual risk $R_{\boldsymbol{a}}(f_{\boldsymbol{a}})$ is identifiable under consistency, as

$$\ell_{f_{\boldsymbol{a}}}(\boldsymbol{X}) = \mathbb{E}_{Y(\boldsymbol{a})|\boldsymbol{X}}\left[L\left(f_{\boldsymbol{a}}(\boldsymbol{X}), Y(\boldsymbol{a})\right) \mid \boldsymbol{X}\right] = \mathbb{E}_{Y|\boldsymbol{X},\boldsymbol{A}}\left[L\left(f_{\boldsymbol{a}}(\boldsymbol{X}), Y\right) \mid \boldsymbol{X}, \boldsymbol{A} = \boldsymbol{a}\right].$$

In multi-cause problem, multi-cause conditional average treatment effect $\tau(\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{x}) = \mathbb{E}\left[Y(\boldsymbol{a}) - Y(\boldsymbol{a}') \mid \boldsymbol{X} = \boldsymbol{x}\right]$ was focused on. Using Theorem 1, we can bound the population risk of treatment effect $R(\hat{\tau}_{\boldsymbol{a},\boldsymbol{a}'})$ that $R(\hat{\tau}_{\boldsymbol{a},\boldsymbol{a}'}) \leq 2(R(f_{\boldsymbol{a}}) + R(f_{\boldsymbol{a}'})) - 4\sigma^2_{Y_{\boldsymbol{a},\boldsymbol{a}'}}$ where $\hat{\tau}_{\boldsymbol{a},\boldsymbol{a}'} = f_{\boldsymbol{a}} - f_{\boldsymbol{a}'}$ and $\sigma^2_{Y_{\boldsymbol{a},\boldsymbol{a}'}} := \max(\sigma^2_{Y(\boldsymbol{a})}, \sigma^2_{Y(\boldsymbol{a}')})$. Based on Theorem 1, the population risk of $f_{\boldsymbol{a}}$ is influenced by both the empirical risks and the IPM (integral probability metric distance), which quantifies the shift in distribution between different treatment groups. Moreover, increasing the sample size can alleviate the upper bound, thereby making data augmentation a viable approach to mitigating the population risk. Additionally, a smaller validation error for the training model (referred to as augmented error) leads to a lower bound, and the number of intervened partial-cause p impacts the validation error as a larger p may increase the augmented error. Consequently, the intervened cause number p serves as a tuning parameter, necessitating the selection of an appropriate value to enhance performance.

Proof of Theorem 1. For a hypothesis f with expected point-wise loss $\ell_f(x)$ such that $\ell_f / \|\ell_f\|_{\mathcal{L}} \in \mathcal{L}$ with $\boldsymbol{a} \in \{0,1\}^K$,

$$R_{\boldsymbol{a}'}(f) - R_{\boldsymbol{a}}(f) \leq \|\ell_f\|_{\mathcal{L}} \operatorname{IPM}_{\mathcal{L}}\left(p_{\boldsymbol{a}'}, p_{\boldsymbol{a}}\right), \tag{A.7}$$

where $p_{\boldsymbol{a}}(x) = p(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{A} = \boldsymbol{a})$, IPM is the integral probability metric distance between $p_{\boldsymbol{a}}$ and $p_{\boldsymbol{a}'}$ w.r.t. $\mathcal{L}$ defined as follows: $\operatorname{IPM}_{\mathcal{L}}(p, q) := \sup_{\ell \in \mathcal{L}} |\mathbb{E}_p[\ell(x)] - \mathbb{E}_q[\ell(x)]|$. $\mathcal{L} \subset \{\mathcal{X} \to \mathbb{R}_+\}$ is a space of pointwise loss functions with respect to the covariates $\boldsymbol{X}$ endowed with a norm $\|\cdot\|_{\mathcal{L}}$. Here we assume that the expected conditional loss $\ell_{f_{\boldsymbol{a}}}$ for each potential outcome belongs to such a family, i.e. that $\ell_{f_{\boldsymbol{a}}} \in \mathcal{L}$.

Based on equation A.7, we can derive the bound of population risk $R(f_{\boldsymbol{a}})$ as:

$$R\left(f_{\boldsymbol{a}}\right) \leq R_{\boldsymbol{a}}\left(f_{\boldsymbol{a}}\right) + \sum_{\substack{\boldsymbol{a}' \in \{0,1\}^K \\ \boldsymbol{a}' \neq \boldsymbol{a}}} \pi_{\boldsymbol{a}'} \|\ell_{f_{\boldsymbol{a}}}\|_{\mathcal{L}} \operatorname{IPM}_{\mathcal{L}}\left(p_{\boldsymbol{a}'}, p_{\boldsymbol{a}}\right). \tag{A.8}$$

Particular choices of $\mathcal{L}$ make the IPM equivalent to different well-know distances on distributions. With $\mathcal{L}$ the family of functions in the norm-1 ball in a reproducing kernel Hilbert space (RKHS), $\mathrm{IPM}_{\mathcal{L}}$ is the Maximum Mean Discrepancy (MMD); When $\mathcal{L}$ is the family of functions with Lipschitz constant at most 1, we obtain the Wasserstein distance.

Next we aim to bound the population risk by the difference between empirical estimates of $R_{\boldsymbol{a}}(f_{\boldsymbol{a}})$ and $\mathrm{IPM}_{\mathcal{L}}(p_{\boldsymbol{a}}, p_{\boldsymbol{a}'})$ and their expected counterparts. We modify the results of Johansson et al. (2020) to give bounds on multi-cause setting. Let $\ell_f = \mathbb{E}_{Y|\boldsymbol{X}}[L(f(\boldsymbol{x}), Y) \mid \boldsymbol{X} = \boldsymbol{x}]$ be the expectation of the squared loss $L(y, y') = (y - y')^2$ of a hypothesis $f \in \mathcal{H} \subset \{f' : \mathcal{X} \to \mathbb{R}\}$, let $d_P = \mathrm{Pdim}(\{\ell_f : f \in \mathcal{H}\})$ where Pdim is the pseudo-dimension of $\mathcal{H}$ and let $\sigma_Y^2 = \mathbb{E}_{\boldsymbol{X},Y}\left[L\left(Y, \mathbb{E}_{Y|\boldsymbol{X}}[Y \mid \boldsymbol{X}]\right)\right]$. Given samples $((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n))$ with empirical distribution $\hat{p}$, with probability at least $1 - \delta$,

$$R_{\boldsymbol{a}}(f_{\boldsymbol{a}}) \leq \hat{R}_{\boldsymbol{a}}(f_{\boldsymbol{a}}) + V_{p_{\boldsymbol{a}}, \hat{p}_{\boldsymbol{a}}}\left[l_{f_{\boldsymbol{a}}}(\boldsymbol{x})\right] \frac{\mathcal{C}_{n_{\boldsymbol{a}}}^{\mathcal{H}}}{n_{\boldsymbol{a}}^{3/8}} + \sigma_{Y_{\boldsymbol{a}}}^2, \tag{A.9}$$

where $\mathcal{C}_{n_{\boldsymbol{a}}}^{\mathcal{H}} = 2^{5/4}\left(d_P \log \frac{2n_{\boldsymbol{a}}e}{d_P} + \log \frac{4}{\delta}\right)^{3/8}$ and $V_{p_{\boldsymbol{a}}, \hat{p}_{\boldsymbol{a}}}\left[l_{f_{\boldsymbol{a}}}(\boldsymbol{x})\right] = \max\left(\sqrt{\mathbb{E}_{\boldsymbol{X}}\left[\ell_{f_{\boldsymbol{a}}}^2(\boldsymbol{X})\right]}\sqrt{\mathbb{E}_{\boldsymbol{X} \sim \hat{p}_{\boldsymbol{a}}}\left[\ell_{f_{\boldsymbol{a}}}^2(\boldsymbol{X})\right]}\right)$.

Equation A.9 allows us to separate the bias (the IPM-term) and variance. The efficiency with which samples may be used to estimate $\mathrm{IPM}_{\mathcal{L}}$ depends on the chosen function family $\mathcal{L}$.

The distribution shift between population distribution and empirical distribution is proved as follows. Suppose k is a universal, measurable kernel such that $\sup_{\boldsymbol{x} \in \mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x}) \leq C \leq \infty$ and $\mathcal{L}$ the reproducing kernel Hilbert space induced by k, with $\nu := \sup_{\boldsymbol{x} \in \mathcal{X}, f \in \mathcal{L}} f(\boldsymbol{x}) \leq \infty$. Then with $\hat{p}, \hat{q}$ the empirical distributions of p, q from m and n samples, and with probability at least $1 - \delta$,

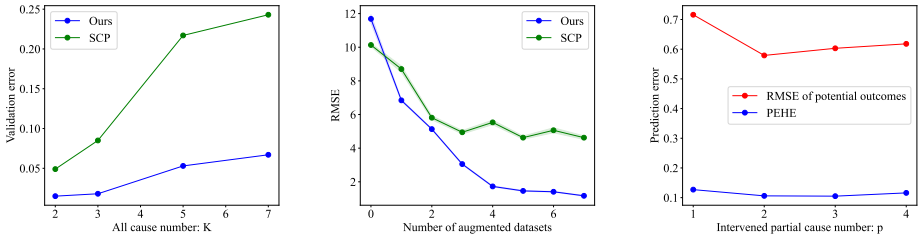
$$\left|\mathrm{IPM}_{\mathcal{L}}(p, q) - \mathrm{IPM}_{\mathcal{L}}(\hat{p}, \hat{q})\right| \leq \sqrt{18\nu^2 \log \frac{4}{\delta} C}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right). \tag{A.10}$$

Also the augmented outcomes would bring bias that $\hat{R}_{\boldsymbol{a}}(f_{\boldsymbol{a}}) = \frac{1}{n_{\boldsymbol{a}}} \sum_{\boldsymbol{a}_i = \boldsymbol{a}} L(f_{\boldsymbol{a}}(\boldsymbol{x}), \tilde{y}_i) + \frac{1}{n_{\boldsymbol{a}}} \sum_{\boldsymbol{a}_i = \boldsymbol{a}} L(y_i, \tilde{y}_i)$ where $\tilde{y}_i$ means the augmented outcome.

$$\begin{aligned}
R\left(f_{\boldsymbol{a}}\right) \leq & R_{\boldsymbol{a}}\left(f_{\boldsymbol{a}}\right) + \sum_{\substack{\boldsymbol{a}' \in \{0,1\}^K \\ \boldsymbol{a}' \neq \boldsymbol{a}}} \pi_{\boldsymbol{a}'} \left\|\ell_{f_{\boldsymbol{a}}}\right\|_{\mathcal{L}} \mathrm{IPM}_{\mathcal{L}}\left(p_{\boldsymbol{a}'}, p_{\boldsymbol{a}}\right) \\
\leq & \hat{R}_{\boldsymbol{a}}(f_{\boldsymbol{a}}) + V_{p_{\boldsymbol{a}}, \hat{p}_{\boldsymbol{a}}}\left[l_{f_{\boldsymbol{a}}}(\boldsymbol{x})\right] \frac{\mathcal{C}_{n_{\boldsymbol{a}}}^{\mathcal{H}}}{n_{\boldsymbol{a}}^{3/8}} + \sigma_{Y_{\boldsymbol{a}}}^2 \\
& + B \sum_{\substack{\boldsymbol{a}' \in \{0,1\}^K \\ \boldsymbol{a}' \neq \boldsymbol{a}}} \pi_{\boldsymbol{a}'} \mathrm{IPM}_{\mathcal{L}}\left(\hat{p}_{\boldsymbol{a}}, \hat{p}_{\boldsymbol{a}'}\right) + \sum_{\substack{\boldsymbol{a}' \in \{0,1\}^K \\ \boldsymbol{a}' \neq \boldsymbol{a}}} \pi_{\boldsymbol{a}'} \mathcal{D}_\delta^{\mathcal{L}}\left(\frac{1}{\sqrt{n_{\boldsymbol{a}_0} + n_{\boldsymbol{a}_+}}} + \frac{1}{\sqrt{n_{\boldsymbol{a}_0'} + n_{\boldsymbol{a}_+'}}}\right) \\
\leq & \frac{1}{n_{\boldsymbol{a}_0} + n_{\boldsymbol{a}_+}}\left(\sum_{i:\boldsymbol{a}_i = \boldsymbol{a}}^{n_{\boldsymbol{a}_0}} L\left(f_{\boldsymbol{a}}\left(x_i\right), y_i\right) + \sum_{\substack{i > n_{\boldsymbol{a}_0} \\ \boldsymbol{a}_i = \boldsymbol{a}}}^{n_{\boldsymbol{a}_0} + n_{\boldsymbol{a}_+}} L\left(f_{\boldsymbol{a}}\left(x_i\right), \tilde{y}_i\right) + \sum_{\substack{i > n_{\boldsymbol{a}_0} \\ \boldsymbol{a}_i = \boldsymbol{a}}}^{n_{\boldsymbol{a}_0} + n_{\boldsymbol{a}_+}} L\left(\tilde{y}_i, y_i\right)\right) \\
& + \frac{1}{\left(n_{\boldsymbol{a}_0} + n_{\boldsymbol{a}_+}\right)^{3/8}} V_{p_{\boldsymbol{a}}}\left(\ell_{f_{\boldsymbol{a}}}\right) \mathcal{C}_{n_{\boldsymbol{a}_0} + n_{\boldsymbol{a}_+}, \delta}^{\mathcal{H}} \\
& + B \sum_{\substack{\boldsymbol{a}' \in \{0,1\}^K \\ \boldsymbol{a}' \neq \boldsymbol{a}}} \pi_{\boldsymbol{a}'} \mathrm{IPM}_{\mathcal{L}}\left(\hat{p}_{\boldsymbol{a}}, \hat{p}_{\boldsymbol{a}'}\right) + \sum_{\substack{\boldsymbol{a}' \in \{0,1\}^K \\ \boldsymbol{a}' \neq \boldsymbol{a}}} \pi_{\boldsymbol{a}'} \mathcal{D}_\delta^{\mathcal{L}}\left(\frac{1}{\sqrt{n_{\boldsymbol{a}_0} + n_{\boldsymbol{a}_+}}} + \frac{1}{\sqrt{n_{\boldsymbol{a}_0'} + n_{\boldsymbol{a}_+'}}}\right) \\
& + \sigma_{Y(\boldsymbol{a})}^2.
\end{aligned} \tag{A.11}$$

Then with all the equations in this proof, we derive the bound in Theorem 1.

□



(a) (b) (c)

Figure 3: (a)Validation error for train causal models before data augmentation on simulation data. (b) Augmented data points reduces prediction error. RMSE declines as more datasets $\mathcal{D}_k$ are added for training. There are $K = 7$ causes in this simulation. (c) Prediction error of potential outcomes (red line) and M-CATE (blue line) varies as intervened partial cause number p increases when all cause number $K = 5$.

A.3 Supplement for Simulation Study

Data setting Each dataset contains 500 samples for training, 200 samples for validation, and 500 samples for testing. The training sets and validation sets contain observations $(\boldsymbol{x}_i, y_i, \boldsymbol{a}_i)$ whereas the testing set contains $(\boldsymbol{x}_i, y_i(\boldsymbol{a}), \forall \boldsymbol{a} \in \Omega$. To generate an observation, we first sample d_c covariates independently or partly dependently based on the known simulated graphs: for some $d_c \leq d, x_{id_c} \sim N(0,1)$ or $x_{id_c} = \sum_{m=1}^{d_c} b_m^{(x,d_c)} x_{im} + e_{id_c}$. Then we obtain the causes $a_{ik}, \forall k \leq K$ and the outcome y_i :

$$a_{ik} \sim \mathrm{B}\left[\sigma\left(\sum_{m=1}^{d} b_m^{(a)} x_{im} + \sum_{n=1}^{k-1} c_n^{(a)} a_{in} + \epsilon_{ik}\right)\right], \tag{A.12}$$

$$\begin{aligned} y_i = & \phi\Big(\sum_{m=1}^{d} b_m^{(y)} x_{im} + \sum_{n=1}^{K} c_n^{(y)} a_{in} + \sum_{l=1}^{d}\sum_{j=l}^{d} e_{lj} x_{il} x_{ij} \\ & + \sum_{r=1}^{K}\sum_{s=l}^{K} q_{rs} a_{ir} a_{is} + \sum_{u=1}^{d}\sum_{v=l}^{K} w_{uv} x_{iu} a_{iv} + \varepsilon_i\Big). \end{aligned} \tag{A.13}$$

b, c, e, q, w are weights(only a fraction of them are non-zero) and superscripts (a) mean that the variables work on the cause a. $\mathrm{B}[\cdot]$ denotes a Bernoulli random variable, σ denotes the sigmoid function, $\phi(\cdot)$ is either identity or sigmoid function.

Other results We also compare the validation error before data augmentation to explore why our method has superior achievement than SCP in Figure 3(a). As the number of causes increases, the validation error is also growing. Our method's validation error is smaller than SCP with a different number of causes. In Figure 3(b), prediction error declines as the size of the augmented dataset increases and our results reduce more significantly. To study how the intervened partial-cause number p influences the prediction error, we plot the changes in Figure 3(c) when all causes number $K = 5$. We find the prediction error first decreases as p increases and then becomes higher under bigger p. Too big p is not better. Thus in experiments, p needs to be chosen based on the validation set. Through Table 2 and Figure 3, we can verify that causal graph could strengthen the prediction performance: (1) Under causal graph model, our method would have a more concrete model for outcomes; (2) Obeying causal graph would attain more similar potential causes with initial dataset and it will benefit for the outcome prediction because step one models are learned from initial data; (3) Outcome distribution conditional on covariates and causes in our augmented data will be closer to real data, thus final outcome model with enlarged data performs better. Furthermore, Figure 3(b) demonstrates that the prediction error would reduce as the sample size enlarges.

Table 2: Prediction error on potential outcomes (RMSE) and treatment effect (PEHE), and data balancing (Wasserstein distance) on simulation data

		RMSE (std)		
Method	$K = 2$	3	5	7
NN	0.352 (.012)	1.004 (.026)	1.943 (.089)	7.105 (.088)
NN_IPW	0.324 (.010)	0.984 (.034)	2.381 (.140)	13.06 (.277)
OP	0.546 (.021)	1.423 (.041)	2.789 (.099)	8.811 (.237)
VSR	0.369 (.009)	0.902 (.026)	1.893 (.090)	9.243 (.137)
DEC	0.393 (.009)	0.703 (.017)	1.819 (.079)	6.505 (.050)
CFR	0.374 (.026)	2.162 (.090)	3.069 (.098)	176.9 (4.903)
DR-CFR	0.546 (.024)	2.892 (.089)	4.297 (.100)	160.1 (4.740)
SCP	0.223 (.010)	0.663 (.048)	1.033 (.070)	6.012 (.102)
PCI (Ours)	**0.169 (.006)**	**0.559 (.044)**	**0.631 (.068)**	**1.056 (.024)**
		PEHE (std)		
Method	$K = 2$	3	5	7
NN	0.176 (.006)	0.277 (.015)	0.293 (.016)	6.477 (.056)
NN_IPW	0.204 (.004)	0.215 (.008)	0.366 (.021)	7.878 (.210)
OP	0.199 (.005)	0.298 (.006)	0.430 (.020)	15.11 (.500)
VSR	0.245 (.037)	0.211 (.011)	0.283 (.016)	8.150 (.068)
DEC	0.278 (.026)	0.174 (.009)	0.250 (.012)	4.986 (.033)
CFR	0.154 (.006)	0.657 (.032)	0.509 (.023)	14.48 (.523)
DR-CFR	0.184 (.008)	0.580 (.019)	0.769 (.034)	14.31 (.516)
SCP	0.115 (.008)	0.156 (.010)	0.151 (.008)	4.774 (.062)
PCI (Ours)	**0.063 (.002)**	**0.133 (.008)**	**0.102 (.015)**	**1.056 (.024)**
		Wasserstein distance ($\mathrm{IPM}_{\mathcal{L}}$)		
Dataset	$K = 2$	3	5	7
Dataset0	13.5	28.1	177	1947
SCP	6.26	17.6	145	1871
PCI (Ours)	**5.96**	**17.5**	**137**	**1810**

A.4 Simulation on document recommendation

Data setting We construct a simulation environment about document recommendation to mimic the recommendation systems in the real world following the setting in VSR Zou et al. (2020). A document D_i is characterized by the topic c_i and quality q_i. Let $\boldsymbol{X} \in \mathbb{R}^d$ be the preference variable of users for different document topics, where d is the number of document topics. The recommending score of each document D_i is calculated as the document quality plus the preference value of the document topic, i.e. $Score_i = x_{c_i} + q_i$. The s documents with the highest score are selected as recommended documents forming the treatment. The outcome is generated from a pre-defined function, determined by both confounders x and treatments a:

$$\mathbf{y} = \sum_{j=1}^{d} \sum_{k=1}^{K} x_j d_{j,k} a_k + \varepsilon_y, \tag{A.14}$$

where $\mathbf{D} = \{d_{j,k}\}_{1 \le j \le d, 1 \le k \le n}$ is a pre-defined matrix, and ε_y is a normal noise.

In this simulation, we set the sample size $n = 2000$, the number of document topics $d = 8$, selected documents $s = 6$, and the noise variable $\varepsilon_y \sim \mathcal{N}\left(0, 0.01^2\right)$.

Results Based on the validation error, we choose the intervened partial-cause number to be $p = 2$ when cause number $K = 8, 9, 10$. In Table 3, Our PCI method gains the smallest PEHE and consistently outperforms the benchmarks across the different number of causes K. The performance gain becomes more pronounced as the number of causes increases. Our method also derives the smallest Wasserstein distance. In this simulation, we verified that the intervened number p influences the treatment effect and outcome estimation, thus choosing the appropriate p is crucial.

Table 3: Prediction error on treatment effect estimation (PEHE), and data balancing (Wasserstein distance) on recommendation simulation

	PEHE (std)		
Method	$K = 8$	9	10
NN	0.92 (.001)	1.01 (.001)	0.99 (.001)
NN_IPW	0.58 (.001)	1.07 (.002)	1.05 (5e-4)
OP	0.51 (.001)	0.93 (.001)	0.40 (4e-4)
VSR	1.06 (.002)	1.24 (.002)	1.10 (.001)
DEC	1.47 (.002)	1.05 (.001)	1.33 (.001)
CFR	0.79 (.003)	1.05 (.002)	0.87 (.001)
DR-CFR	1.68 (.003)	2.01 (.003)	2.12 (.002)
SCP	0.56 (.001)	1.85 (.002)	1.91 (.004)
PCI (Ours)	**0.48 (.001)**	**0.88 (.001)**	**0.27 (2e-4)**
	Wasserstein distance ($\text{IPM}_{\mathcal{L}}$)		
Dataset	$K = 8$	9	10
Dataset0	12.6	42.6	61.6
SCP	12.5	40.9	59.7
PCI (Ours)	**6.42**	**22.6**	**55.7**

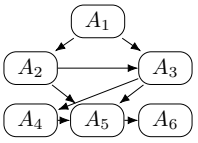


Figure 4: Causal graph of 6 causes in V-Dem dataset.

A.5 Experiment on V-Dem dataset

V-Dem Dataset Our second application focuses on the role of democratic political institutions in reducing the likelihood of civil war onset. Democracy is a fundamental concept when modeling the quality of governance, but drawing inferences about its effect represents a straightforward example of the multi-cause setting. In particular, democracy cannot be measured as a single unambiguous feature – instead, it is a confluence of many conceptually related by empirically distinct features describing different aspects of a system of governance. The causal effect of democracy on outcomes like conflict initiation is typically measured using a dimension reduction of the features representing the individual institutions. We refine features describing the system of governance present in a country in the Varieties of Democracy Dataset (V-Dem) to measure the democratic political institutions and finally concentrate on 6 causes and 27 covariates to quantify the effect of these political institutions on civil war onset (binary outcome generated in semi-synthetic setting). These 6 causes represent: (1) does the electoral principle of democracy achieve? (2) do the elected local and regional governments operate without interference from unelected bodies? (3) are citizens able to openly discuss political issues in private homes and in public spaces? (4) are laws transparently, independently, predictably, impartially, and equally enforced? (5) does government respect press and media freedom as well as the freedom of academic and cultural expression? (6) is civil society robust? Covariates in the V-Dem dataset are relevant to rights and equality, democratic and clean politics. The causal relationships among these causes are shown in a causal graph. In the semi-synthetic experiment, we model the outcome to be linear with causes and covariates as equation A.13.

Results Based on the validation error, we choose $p = 1$ in the linear outcomes setting and $p = 2$ in the non-linear setting. In Table 4, we compare the proposed and other methods under generated outcomes Y. In semi-synthetic data, our method gains the smallest RMSE on all potential outcomes prediction. Because we only have observational test data, we only compare the performance on the observational outcome prediction and our method achieves the best accuracy in terms of RMSE.

Table 4: Prediction error on treatment effect (PEHE) and potential outcomes (RMSE) under V-Dem dataset

Method	Linear outcomes		Non-linear outcomes	
	PEHE (std)	RMSE (std)	PEHE (std)	RMSE (std)
NN	0.692 (.006)	0.906 (.011)	0.024 (4e-5)	0.101 (.001)
NN_IPW	0.841 (.006)	1.287 (.009)	0.042 (5e-5)	0.092 (2e-4)
OP	0.885 (.006)	0.930 (.008)	0.042 (7e-5)	0.155 (.001)
VSR	0.738 (.006)	0.816 (.006)	0.020 (3e-5)	0.083 (.001)
DEC	0.696 (.006)	0.793 (.007)	0.020 (5e-5)	0.086 (.001)
SCP	0.729 (.003)	0.782 (.003)	0.136 (.001)	0.233 (4e-4)
PCI (Ours)	**0.649 (.004)**	**0.697 (.006)**	**0.013 (4e-5)**	**0.071 (.001)**
	Wasserstein Distance ($\text{IPM}_{\mathcal{L}}$)			
Method	Linear outcomes		Non-linear outcomes	
dataset0	357.8		357.8	
SCP	336.8		336.8	
PCI (Ours)	**326.2**		**330.7**	

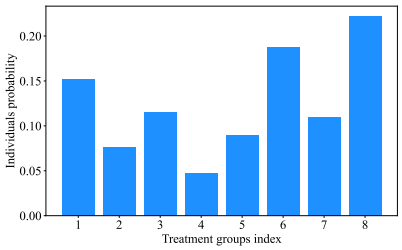


Figure 5: ADNI dataset in multi-cause setting: Histogram of individuals distribution

A.6 ADNI ANALYSIS

In AD fields, when choosing patients' treatments, researchers need to estimate the influence of brain atrophy on cognition based on the causal relationships among variables (Choo et al., 2010; Gainotti et al., 2003; Schröder and Pantel, 2016). Cognition decline is one of the main symptoms of Alzheimer's Disease which is strongly associated with the atrophy of three brain regions (temporal (A_1), cingulate cortex (A_2), and hippocampal regions (A_3)). These $K = 3$ causes (three regions: hippocampus, temporal and cingulate cortex) are valued as $\{0, 1\}$ and form 8 treatments in 2129 individuals. The individuals with different treatments is unevenly distributed shown in Figure 2's histogram.

When learning the causal graph of ADNI data, we obey some rules based on medical knowledge and common sense: (1) genes may have effects on individuals' intelligence and Alzheimer's Disease (Reiman et al., 1998; Rohrer et al., 2010), (2) cingulate is associated with the atrophy of temporal region (Iizuka and Kameyama, 2016) and (3) sex and age cannot be affected by the covariates in ADNI data. Figure 6 is the learned causal graph of ADNI data.

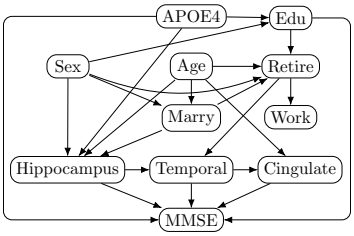


Figure 6: ADNI dataset in multi-cause setting: the effect of brain atrophy on cognition.

Table 5: Prediction error on treatment effect (PEHE) and outcomes (RMSE) under ADNI dataset

Method	Generated outcomes		Real outcomes
	PEHE (std)	RMSE (std)	RMSE (std)
NN	0.106 (.001)	1.759 (.017)	4.542 (.113)
NN_IPW	0.087 (4e-4)	1.784 (.017)	4.424 (.111)
OP	0.086 (4e-4)	1.800 (.017)	4.442 (.115)
VSR	0.086 (4e-4)	1.794 (.018)	4.507 (.105)
DEC	0.085 (4e-4)	1.787 (.016)	4.535 (.110)
SCP	0.086 (3e-4)	1.785 (.016)	4.316 (.114)
PCI	**0.070 (.001)**	**0.379 (.003)**	**3.910 (.110)**
	Wasserstein Distance ($\mathrm{IPM}_{\mathcal{L}}$)		
Method	generate outcome		real MMSE
dataset0	47.18		47.18
SCP	33.89		33.89
PCI	**33.55**		**33.45**