6 Appendix

6.1 Proofs of theorems

Lemma 3.1. *Suppose that* $\mathbb{E}[Y \mid \mathrm{Pa}(Y) = a] \neq \mathbb{E}[Y \mid \mathrm{Pa}(Y) = a']$ *whenever* $a \neq a'$. *Then a representation* Φ *is invariant across all valid environments if and only if* $\mathbb{E}[Y^e|\Phi(T^e, X^e)] = \mathbb{E}[Y|\mathrm{Pa}(Y)]$ *for all valid environments.*

Proof. The if direction is immediate.

To establish the only if direction, we first show that Φ must contain at least $\mathrm{Pa}(Y)$, in the sense $\mathbb{E}[Y \mid \Phi(X)] = \mathbb{E}[Y \mid \mathrm{Pa}(Y) \cup Z]$ for some set Z. We proceed by contradiction. Suppose that conditioning on Φ is equivalent to conditioning on only $\mathrm{Pa}(Y) \setminus \{P\} \cup Z$, where P is a parent of Y. We now create two environments by setting $P = p$ and $P = p'$. Since P is a parent of Y this follows from the second rule of do calculus [Pea00],

$$\mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z; do(P = p)] = \mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z, P = p]$$

and

$$\mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z; do(P = p')] = \mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z, P = p'].$$

The equality $\mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z, P = p] = \mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z, P = p']$ holds only if P is conditionally independent of Y given $\mathrm{Pa}(Y) \setminus \{P\} \cup Z$. Since P is a parent of Y, by the first assumption of the lemma, the equality does not hold. It follows that $\mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z\,;\, do(P = p)] \neq \mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z\,;\, do(P = p')]$. That is, if conditioning on Φ was equivalent to conditioning on less information than $\mathrm{Pa}(Y) \cup Z$, then Φ would not be invariant across all valid environments.

It remains to show that Φ does not contain any more information than $\mathrm{Pa}(Y)$.

Φ cannot contain any descendants of the outcome. Suppose that Φ depends on some descendant D of Y in the sense that there is at least one environment and $d \neq d'$ where $\mathbb{E}[Y \mid \Phi(X \setminus D, D = d)] \neq \mathbb{E}[Y \mid \Phi(X \setminus D, D = d')]$. Then, construct a new environment e by randomly intervening and setting $\mathrm{do}(D = d)$ or $\mathrm{do}(D = d')$, each with probability 0.5. In this new environment, there is no relationship between Y and D. Accordingly, $\mathbb{E}[Y^e \mid \Phi(X^e \setminus D^e, D^e = d)] = \mathbb{E}[Y^e \mid \Phi(X^e \setminus D^e, D^e = d')]$. Thus, the conditional expectations are not equal (as functions of d) in the two environments—a contradiction.

Next, we show that, Φ need not to contain the non-parent ancestors A of the outcome, because $\mathbb{E}[Y \mid \{A\} \cup \mathrm{Pa}(Y)] = \mathbb{E}[Y \mid \mathrm{Pa}(Y)]$ by the Markov property of the causal graph, where A is any non-ancestor variables. Since Φ contains $\mathrm{Pa}(Y)$, it follows that Φ does not depend on any non-parent ancestor A.

Theorem 3.2. *Let* L *be a loss function such that the minimizer of the associated risk is a conditional expectation, and let* Φ *be a representation that elicits a predictor* Q^{inv} *that is invariant for all valid distributions. Assuming there is no mediators between the treatment and the outcome, then* $\psi = \mathbb{E}\left[Q^{\mathrm{inv}}(1, X) - Q^{\mathrm{inv}}(0, X)|T = 1\right]$.

Proof. We assume the technical condition of lemma 3.1, that $\mathbb{E}[Y \mid \mathrm{Pa}(Y) = a] \neq \mathbb{E}[Y \mid \mathrm{Pa}(Y) = a']$ whenever $a \neq a'$. This is without loss of generality because violations of this condition will not lead to different causal effects.

By the assumption on the loss function, the elicited invariant predictor is $\mathbb{E}[Y \mid \Phi(T, X)]$. Lemma 3.1 shows that $\mathbb{E}[Y \mid \Phi(T, X)] = \mathbb{E}[Y \mid \mathrm{Pa}(Y)]$. We further observe that the non-treatment parents of Y are sufficient to block backdoor paths. It follows the ATT can be expressed as the following.

$$\begin{aligned}\Psi &= \mathbb{E}\left[\mathbb{E}[Y \mid T = 1, \mathrm{Pa}(Y) \setminus \{T\}] - \mathbb{E}[Y \mid T = 0, \mathrm{Pa}(Y) \setminus \{T\})] \mid T = 1\right] \\ &= \mathbb{E}\left[\mathbb{E}[Y \mid \Phi(1, X)] - \mathbb{E}[Y \mid \Phi(0, X)] \mid T = 1\right]\end{aligned}$$

Theorem 3.3. *Suppose* $0 < P(T^e = 1|X^e) < 1$ *with probability 1, then* $0 < P(T^e = 1|\Phi(X^e)) < 1$ *with probability 1.*

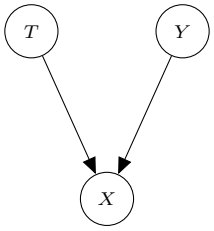

Figure 6: V-structure graph. We denote the bias induced by conditioning on X as V-bias.

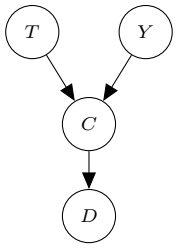

Figure 7: Y-structure graph. We denote the bias induced by conditioning on D as Y-bias.

Proof. The proof follows directly from Theorem 1 in [D'A+20]. The intuition is that the richer the covariate set is, the more likely it is to predict the treatment assignment accurately [D'A+20]. The covariate representation $\Phi(X^e)$ by definition contains less information than X^e, therefore $\Phi(X^e)$ satisfies overlap if X^e satisfies overlap.

Consider the DGP with binary variables $\{X, Y, T\}$ illustrated in figure 6, where X is causally influence by Y and T.

Theorem 6.1. *Let cov denote the covariance between two variables, we define collider bias at $X = c$ as $\Delta(X = c) = cov(T, Y|X = c) - cov(T, Y)$, and collider bias of X as $\Delta(X) = |P(X = 1)\Delta(X = 1) + P(X = 0)\Delta(X = 0)|$. Let $\Phi(T, X)$ be a random variable, where $P(\Phi(T, X) = X) \geq 0.5$. Suppose $P(X = 1) = 0.5$, and $\Delta(X = 1)$ has the same sign as $\Delta(X = 0)$, conditioning on X induce more collider bias than conditioning its coarsening $\Phi(T, X)$:*

$$\Delta(\Phi(T, X)) \leq \Delta(X)$$

Proof. The proof follows corollary 2.1 in [NDO19].

Corollary 2.1. *We refer to collider bias in the V substructure embedded in the Y structure as 'embedded V-bias' and denote it as $\Delta(C = c)$. For the covariance effect scale, Y-bias $\Delta(D = d)$ relates to embedded V-bias through the following formula:*

$$\Delta(D = d) = \frac{p(D = d \,|\, C = 1) - p(D = d \,|\, C = 0)}{\{\mathrm{P}(D = d)\}^2} \cdot \begin{bmatrix} p(D = d \,|\, C = 1)\{\mathrm{P}(C = 1)\}^2 \cdot \Delta(C = 1) - \\ p(D = d \,|\, C = 0)\{\mathrm{P}(C = 0)\}^2 \cdot \Delta(C = 0) \end{bmatrix}.$$

With the corollary above, let D denote $\Phi(T, X)$, let C denote the collider X in figure 6. The bias induced by conditioning on D is less than the bias induced by conditioning on C.

$$\begin{aligned}
\Delta(D = 1) &= \frac{2\alpha - 1}{0.25}(0.25\alpha \cdot \Delta(C = 1) - 0.25(1 - \alpha) \cdot \Delta(C = 0)) \\
&= (2\alpha - 1)(\alpha \cdot \Delta(C = 1) - (1 - \alpha) \cdot \Delta(C = 0)) \\
\Delta(D = 0) &= \frac{1 - 2\alpha}{0.25}(0.25(1 - \alpha) \cdot \Delta(C = 1) - 0.25\alpha \cdot \Delta(C = 0)) \\
&= (1 - 2\alpha)((1 - \alpha) \cdot \Delta(C = 1) - \alpha \cdot \Delta(C = 0)) \\
\Delta(C) &= |0.5 \cdot \Delta(C = 0) + 0.5 \cdot \Delta(C = 1)| \\
\Delta(D) &= |0.5 \cdot \Delta(D = 0) + 0.5 \cdot \Delta(D = 1)| \\
\Delta(D) &= |0.5((1 - 2\alpha)((1 - \alpha) \cdot \Delta(C = 1) - \alpha \cdot \Delta(C = 0)) \\
&\quad + 0.5(2\alpha - 1)((\alpha \cdot \Delta(C = 1) - (1 - \alpha) \cdot \Delta(C = 0)))| \\
&= |0.5(2\alpha - 1)^2 \cdot \Delta(C = 1) + 0.5(2\alpha - 1)^2 \cdot \Delta(C = 0)| \\
&\leq \Delta(C)
\end{aligned}$$

6.2 The Case of Mediators

In the most part of the paper, we assumed no mediators between treatment and outcome. What happens to the interpretation of the learned parameter if the adjustment set contains mediators? Intuitively, NICE retains the direct link between the treatment and the outcome. Specifically, if there are no mediators, the parameter reduces to ATT. If there are mediators but no confounders, the

parameter reduces to the Natural Direct Effect [Pea00]. If there are mediators and confounders, the NICE estimand is a non-standard causal target that we call the natural direct effect on the treated (NDET).

Conceptually, NDET describes the expected change in outcome Y for the treated population, induced by changing the value of T, while keeping all mediating factors M, constant at whatever value they would have obtained under $\mathrm{do}(t)$. The main point is that NDET provides answers to questions such as, "does this treatment have a substantial direct effect on this outcome?". Substantively, NDET is the natural direct effect, adjusted for confounders.

Formally, NDET is

$$\psi = \mathbb{E}_{M \mid T=1} \left[\mathbb{E}\left[Y \mid M\,;\, \mathrm{do}(T=1)\right] - \mathbb{E}\left[Y \mid M\,;\, \mathrm{do}(T=0)\right] | T=1 \right]. \tag{6.1}$$

With adjustment set W, the causal effect can be expressed through a parameter of the observational distribution:

$$\psi = \mathbb{E}_{M,W} \left[\mathbb{E}\left[Y \mid T=1, M, W\right] - \mathbb{E}\left[Y \mid T=0, M, W\right] \mid T=1 \right]. \tag{6.2}$$

Importantly, the mediators M and the confounders W show up in the same way in (6.2). Accordingly, we don't need to know which observed variables are mediators and which are confounders to compute the parameter. Under the NICE procedure, we condition on all parents of Y, including possible mediators. Thus, the NICE estimand is the NDET.

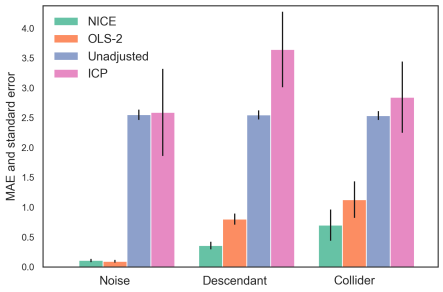

Figure 8: Models performance under the scrambled and heteroskedastic setting

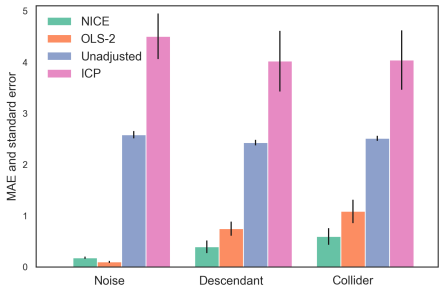

Figure 9: Models performance under the scrambled and homoscedastic setting

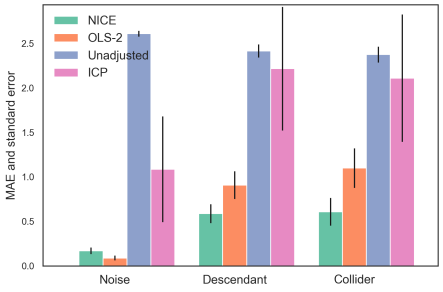

Figure 10: Models performance under the unscrambled and heteroskedastic setting

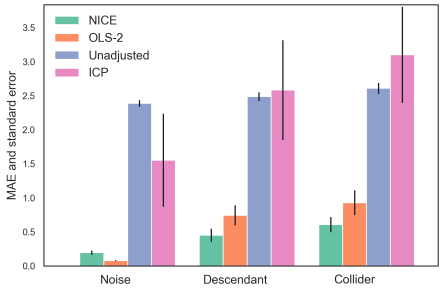

Figure 11: Models performance under the ubscrambled and homoscedastic setting

6.3 Details of the experiments

Experiment 1

We simulate data with three causal graphs in Figure 3. With a slight abuse of notation, each intervention e generates a new environment e with interventional distribution $P(X^e, T^e, Y^e)$. T^e is the binary treatment and Y^e is the outcome. X^e is a 10-dimensional covariate set that differs across DGPs. $X^e = (X_1^e, X_2^e)$, where X_1^e is a five-dimensional confounder. X_2^e is either noise, a descendant, or a collider in each DGP. We examine the models' performance under two types of variations: 1) whether the observed covariates are scrambled versions of the true covariates. 2) whether the treatment effects are heteroskedastic across environments. The data generating process is illustrated below.

$$
\begin{aligned}
X_1^e &\leftarrow \mathcal{N}(0, e^2) \\
P^e &\leftarrow \text{sigmoid}(X_1^e \cdot w_{xt^e} + \mathcal{N}(0,1)) \\
T^e &\leftarrow Bern(P^e) \\
\tau &\leftarrow 5 + \mathcal{N}(0, \sigma^2) \\
Y^e &\leftarrow X_1^e \cdot w_{xy^e} + T^e \cdot \tau + \mathcal{N}(0, e^2)
\end{aligned}
$$

X_2^e equals $\mathcal{N}(0, e^2)$ in setting a), X_2^e equals $e * Y^e + \mathcal{N}(0, q)$ in setting b), and X_2^e equals $e * Y^e + T^e + \mathcal{N}(0, q)$ in setting c). For the four variants, in the scrambled setting $\mathcal{N}(0, \sigma^2) = \mathcal{N}(0, e^2)$, in the un-scrambled setting $\mathcal{N}(0, \sigma^2) = \mathcal{N}(0, 1)$. In the environment-level heteroskedastic setting $\tau \leftarrow 5 + \mathcal{N}(0, e^2)$. In the environment-level homoscedastic setting $\tau \leftarrow 5 + \mathcal{N}(0, 1)$. The performance under the four variants are illustrated in Figure 8, Figure 9, Figure 10, and Figure 11.

Experiment 2

We validate NICE for the non-linear case on the benchmark dataset SpeedDating. SpeedDating was collected to study the gender difference in mate selection [Fis+06]. The study recruited university students to participate in speed dating, and collected objective and subjective information such as 'undergraduate institution' and 'perceived attractiveness'. It has 8378 entries and 185 covariates. ACIC 2019's simulation samples subsets of the covariates to simulate the treatment T and outcome

Y. Specifically, it provides four modified DGPs: Mod1: parametric models; Mod2: complex models; Mod3: parametric models with poor overlap; Mod4: complex models with treatment heterogeneity. Each modification includes three versions: low, med, high, indicating an increasing number of covariates included in the models for T and Y.

We compare the estimation quality of the within-sample SATT and out-of-sample CATE over 10 bootstraps in Table 3 and Table 4. The main paper reports the models performance under the low setting. We now report results for the med and high setting.

Within-sample

	ϵ_{att}			
	MOD1	MOD2	MOD3	MOD4
med				
TARNet	$.04 \pm .02$	$.09 \pm .1$	$.23 \pm .16$	$.12 \pm .09$
+NICE	$.02 \pm .02$	$.06 \pm .03$	$.07 \pm .04$	$.02 \pm .02$
Dragon	$.12 \pm .12$	$.09 \pm .1$	$.23 \pm .16$	$.13 \pm .07$
+NICE	$.04 \pm .02$	$.06 \pm .03$	$.07 \pm .04$	$.07 \pm .08$
high				
TARNet	$.06 \pm .05$	$.18 \pm .12$	$.10 \pm .08$	$.03 \pm .03$
+NICE	$.01 \pm .01$	$.06 \pm .03$	$.05 \pm .02$	$.05 \pm .06$
Dragon	$.18 \pm .12$	$.18 \pm .12$	$.10 \pm .08$	$.05 \pm .03$
+NICE	$.02 \pm .01$	$.06 \pm .03$	$.05 \pm .02$	$.09 \pm .09$

Table 3: NICE perform well relative to the baselines if the adjustment set does not contain bad controls. The table reports MAE and bootstrap standard deviation of the SATT estimation. The model is trained and evaluated on all three environments.

Out-of-sample

	$\sqrt{\epsilon_{\text{PEHE}}}$			
	MOD1	MOD2	MOD3	MOD4
med				
TARNet	$.14 \pm .04$	$.13 \pm .03$	$.11 \pm .05$	$.09 \pm .06$
+NICE	$.06 \pm .02$	$.06 \pm .01$	$.08 \pm .05$	$.07 \pm .02$
Dragon	$.15 \pm .03$	$.13 \pm .03$	$.22 \pm .15$	$.07 \pm .02$
+NICE	$.04 \pm .01$	$.04 \pm .01$	$.08 \pm .02$	$.07 \pm .04$
high				
TARNet	$.14 \pm .06$	$.14 \pm .08$	$.13 \pm .03$	$.08 \pm .10$
+NICE	$.05 \pm .01$	$.07 \pm .01$	$.06 \pm .01$	$.07 \pm .08$
Dragon	$.14 \pm .04$	$.14 \pm .05$	$.15 \pm .04$	$.09 \pm .08$
+NICE	$.06 \pm .01$	$.06 \pm .02$	$.05 \pm .01$	$.05 \pm .06$

Table 4: NICE perform well relative to the baselines if the adjustment set does not contain bad controls. The table reports PEHE and bootstrap standard deviation of the out-of-distribution CATE estimation. The model is trained on two environments and evaluated on the third.

To examine whether NICE helps reduce collider bias, we simulated 20 copies of a collider: $X_{co}^e = T^e + Y^e + \mathcal{N}(0, e^2)$, where $e \in \{0.01, 0.2, 1\}$ and include it in the covariate set. table 5 compares Dragonnet trained under standard empirical risk minimization framework and trained under NICE. NICE reduces collider bias across simulation setups.

Experiment 3

In the third experiment, we consider data generated according figure 2. Notably, in the setup, we observe $P(A, X, T, Y, Z)$, where $A = \{A_1, A_t, A_y\}$ and $X = \{X_t, X_y\}$. Here A is a 50 dimensional covariate, X a 30 dimensional covariate, and Z a 50 dimensional covariate. Z is causally affected by A and Y.

We compare NICE against a neural network model similar to the structure of TARNet. The model architecture is the same as the models in the SpeedDating experiment, except the hidden layer

Within-sample

	ϵ_{att}			
	MOD1	MOD2	MOD3	MOD4
low				
Dragon	$.32 \pm .16$	$.39 \pm .14$	$.32 \pm .02$	$.50 \pm .04$
+NICE	$.11 \pm .05$	$.08 \pm .07$	$.15 \pm .04$	$.08 \pm .05$
med				
Dragon	$.39 \pm .15$	$.29 \pm .13$	$.37 \pm .13$	$.27 \pm .10$
+NICE	$.08 \pm .04$	$.17 \pm .10$	$.09 \pm .03$	$.06 \pm .04$
high				
Dragon	$.36 \pm .11$	$.35 \pm .09$	$.49 \pm .16$	$.28 \pm .06$
+NICE	$.09 \pm .06$	$.15 \pm .08$	$.10 \pm .08$	$.14 \pm .09$

Table 5: NICE reduces estimation bias in the presence of colliders. The table reports the MAE and bootstrap standard deviation of SATT. The model is trained and evaluated on three environments.

Out-of-sample

	ϵ_{pehe}			
	MOD1	MOD2	MOD3	MOD4
low				
TARNet	$.18 \pm .05$	$.42 \pm .03$	$.25 \pm .04$	$.36 \pm .12$
+NICE	$.08 \pm .02$	$.07 \pm .01$	$.08 \pm .02$	$.08 \pm .03$
Dragon	$.25 \pm .06$	$.49 \pm .05$	$.29 \pm .06$	$.45 \pm .06$
+NICE	$.09 \pm .01$	$.09 \pm .03$	$.09 \pm .03$	$.09 \pm .04$
med				
TARNet	$.41 \pm .08$	$.28 \pm .08$	$.35 \pm .06$	$.21 \pm .03$
+NICE	$.09 \pm .02$	$.08 \pm .02$	$.07 \pm .02$	$.09 \pm .01$
Dragon	$.40 \pm .08$	$.32 \pm .04$	$.47 \pm .11$	$.25 \pm .06$
+NICE	$.07 \pm .01$	$.10 \pm .05$	$.08 \pm .03$	$.07 \pm .01$
high				
TARNet	$.24 \pm .06$	$.26 \pm .10$	$.28 \pm .08$	$.25 \pm .10$
+NICE	$.06 \pm .02$	$.09 \pm .02$	$.07 \pm .02$	$.10 \pm .05$
Dragon	$.34 \pm .14$	$.36 \pm .03$	$.37 \pm .08$	$.28 \pm .07$
+NICE	$.09 \pm .03$	$.13 \pm .04$	$.07 \pm .01$	$.12 \pm .05$

Table 6: NICE reduces estimation bias in the presence of colliders. The table reports the CATE and standard deviation of CATE. The model is trained on two environments and evaluated on a third environment.

size is 200 for the shared representation. For the exact data generating process and the detailed implementation of the models, see the associated codebase.

References

[Arj+19] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. "Invariant risk minimization". In: *arXiv preprint arXiv:1907.02893* (2019).

[BP14] E. Bareinboim and J. Pearl. "Transportability from multiple environments with limited experiments: completeness results". In: *Advances in neural information processing systems*. 2014.

[BV07] J. Bhattacharya and W. B. Vogt. *Do instrumental variables belong in propensity scores?* Tech. rep. National Bureau of Economic Research, 2007.

[Büh18] P. Bühlmann. "Invariance, causality and robustness". In: *arXiv preprint arXiv:1812.08233* (2018).

[CEP16] K. Chalupka, F. Eberhardt, and P. Perona. "Multi-level cause-effect systems". In: *Artificial Intelligence and Statistics*. 2016.

[CPE14] K. Chalupka, P. Perona, and F. Eberhardt. "Visual causal feature learning". In: *arXiv preprint arXiv:1412.2309* (2014).

[CH20] C. Cinelli and C. Hazlett. "Making sense of sensitivity: extending omitted variable bias". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1 (2020).

[D'A+20] A. D'Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon. "Overlap in observational studies with high-dimensional covariates". In: *Journal of Econometrics* (2020).

[Fis+06] R. Fisman, S. S. Iyengar, E. Kamenica, and I. Simonson. "Gender differences in mate selection: evidence from a speed dating experiment". In: *The Quarterly Journal of Economics* 2 (2006).

[GZS19] C. Glymour, K. Zhang, and P. Spirtes. "Review of causal discovery methods based on graphical models". In: *Frontiers in genetics* (2019).

[Gru+] S. Gruber, G. Lefebvre, A. PichÃľ, and T. Schuster. *Data Challenge*. URL: `https://sites.google.com/view/acic2019datachallenge`.

[Haa43] T. Haavelmo. "The statistical implications of a system of simultaneous equations". In: *Econometrica, Journal of the Econometric Society* (1943).

[HDPM18] C. Heinze-Deml, J. Peters, and N. Meinshausen. "Invariant causal prediction for nonlinear models". In: *Journal of Causal Inference* 2 (2018).

[Hil11] J. Hill. "Bayesian nonparametric modeling for causal inference". In: *Journal of Computational and Graphical Statistics* (2011).

[Imb04] G. W. Imbens. "Nonparametric estimation of average treatment effects under exogeneity: a review". In: *Review of Economics and statistics* 1 (2004).

[Kao+12] C.-H. Kao, L.-M. Sun, J.-A. Liang, S.-N. Chang, F.-C. Sung, and C.-H. Muo. "Relationship of zolpidem and cancer risk: a taiwanese population-based cohort study". In: *Mayo clinic proceedings*. 5. Elsevier. 2012.

[KB14] D. P. Kingma and J. Ba. "Adam: a method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[KLK12] D. F. Kripke, R. D. Langer, and L. E. Kline. "Hypnotics' association with mortality or cancer: a matched cohort study". In: *BMJ open* 1 (2012).

[Mag+18] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. "Domain adaptation by using causal inference to predict invariant conditional distributions". In: *Advances in Neural Information Processing Systems*. 2018.

[ML09] K. L. Moore and M. J. van der Laan. "Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation". In: *Statistics in medicine* 1 (2009).

[MM+99] K. Murphy, S. Mian, et al. *Modelling gene expression data using dynamic Bayesian networks*. Tech. rep. Citeseer, 1999.

[NDO19] T. Q. Nguyen, A. Dafoe, and E. L. Ogburn. "The magnitude and direction of collider bias for binary variables". In: *Epidemiologic Methods* 1 (2019).

[Pat+17] E. Patorno, R. J. Glynn, R. Levin, M. P. Lee, and K. F. Huybrechts. "Benzodiazepines and risk of all cause mortality in adults: cohort study". In: *bmj* (2017).

[Pea00] J. Pearl. *Causality: models, reasoning and inference*. 2000.

[Pea09] J. Pearl. *Causality*. 2009.

[PP14] J. Pearl and A. Paz. "Confounding equivalence in causal inference". In: *Journal of Causal Inference J. Causal Infer.* 1 (2014).

[PBM16] J. Peters, P. Bühlmann, and N. Meinshausen. "Causal inference by using invariant prediction: identification and confidence intervals". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2016).

[RJ91] L. D. Robinson and N. P. Jewell. "Some surprising results about covariate adjustment in logistic regression models". In: *International Statistical Review/Revue Internationale de Statistique* (1991).

[RC+18] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. "Invariant models for causal transfer learning". In: *The Journal of Machine Learning Research* 1 (2018).

[Ros02] P. R. Rosenbaum. "Overt bias in observational studies". In: *Observational studies*. 2002.

[RR83] P. R. Rosenbaum and D. B. Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* (1983).

[Rub09] D. B. Rubin. "Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups?" In: *Statistics in Medicine* 9 (2009).

[SJS16] U. Shalit, F. D. Johansson, and D. Sontag. "Estimating individual treatment effect: generalization bounds and algorithms". In: *arXiv e-prints arXiv:1606.03976* (2016).

[SBV19] C. Shi, D. M. Blei, and V. Veitch. "Adapting neural networks for the estimation of treatment effects". In: *Advances in neural information processing systems* (2019).

[Shi+06] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. "A linear non-gaussian acyclic model for causal discovery". In: *Journal of Machine Learning Research* Oct (2006).

[SE17] S. M. Shortreed and A. Ertefaie. "Outcome-adaptive lasso: variable selection for causal inference". In: *Biometrics* 4 (2017).

[Spi+00] P. Spirtes, C. Glymour, R. Scheines, S. Kauffman, V. Aimale, and F. Wimberly. "Constructing bayesian network models of gene expression networks from microarray data". In: (2000).

[Wei+14] S. Weich, H. L. Pearce, P. Croft, S. Singh, I. Crome, J. Bashford, and M. Frisher. "Effect of anxiolytic and hypnotic drug prescriptions on mortality hazards: retrospective cohort study". In: *Bmj* (2014).

[Zha+20] K. Zhang, M. Gong, P. Stojanov, B. Huang, and C. Glymour. "Domain adaptation as a problem of inference on graphical models". In: *arXiv preprint arXiv:2002.03278* (2020).

[Zha+16] Q. Zhao, C. Zheng, T. Hastie, and R. Tibshirani. "Comment on causal inference using invariant prediction". In: *arXiv preprint arXiv: 1501.01332* (2016).