A Reweighting Based Approach for Treatment Effect Estimation Under Unmeasured Confounding with Non-Representative Randomized Data

Yaxuan Li[1*], **Chuan Zhou**[1†]

[1] Peking University
yaxuanli.cn@gmail.com, zhouchuancn@pku.edu.cn

Abstract

Causal effect estimation aims to measure the true causal relationship between treatment and outcome variables, which is widely applied in areas such as medicine, commerce, and sociology. A challenge in causal effect estimation is that unmeasured variables may affect both treatment and outcome variables, which are named unmeasured confounders. Traditional methods of causal effect estimation are biased in the presence of unmeasured confounding. Previous data fusion-based methods employ observational data (OBS) combined with limited-sized randomized controlled trial (RCT) data to eliminate confounding bias. However, existing methods typically assume that the OBS and RCT data come from the same target population, a relatively strong assumption given the difficulties of randomized trials. In this paper, we consider relaxing this assumption to achieve data fusion in the case where the RCT data is a biased sample of the target population, thus eliminating selection bias and obtaining unbiased estimates of causal effects. We propose a reweighting-based approach that uses OBS and RCT data successively and debiases in the second stage via reweighting. Extensive experiments are conducted to demonstrate the effectiveness of our method.

Introduction

Estimation of causal effects, which aims at quantifying the impact of the treatment variable on the outcome variable (Imbens and Rubin 2015), is one of the most important tasks of causal science (Robins and Hernán 2016) and has a wide range of applications in medicine (Prosperi et al. 2020), sociology (Gangl 2010), and many other fields (Campbell 2007; Reich et al. 2021; Wang et al. 2023). Conditional average causal effect (CATE) is one of the most commonly considered causal effects estimands by machine learning approaches. CATE examines the average causal effect in a population given the covariates, and therefore reflects the effect of the treatment on different populations in a more fine-grained way than the average treatment effect (ATE), enabling precision medicine, fine-grained governance (Athey 2015), etc.

A classic challenge in CATE estimation is unmeasured confounding (Fewell, Davey Smith, and Sterne 2007), which refers to the presence of unobserved variables affecting both the treatment and the outcome, thus introducing an unknown bias to the estimate (VanderWeele and Arah 2011). For example, when estimating the effect of a treatment regime on a disease in a medical scenario, we are unable to accurately quantify the lifestyle of the patients, which may affect both the choice of treatment regime and the progression of the disease (Zhang et al. 2018). In sociology, parental relationship status may be a potential confounder that is difficult to measure when estimating the effect of hours spent in childcare on aggressive behaviour of children (Orri et al. 2019).

Under standard assumptions, unmeasured confounding is not identifiable (Dorn, Guo, and Kallus 2024). Thus previous classes of approaches to this problem have often relied on additional assumptions. Such methods include sensitivity analyses (Imbens 2003), instrumental variable methods (Joshua D. Angrist and Rubin 1996), and negative control methods (Shi et al. 2020). Sensitivity analysis methods have strong modelling assumptions (Borgonovo and Plischke 2016) and are difficult to apply in the real world. Instrumental variables and negative control methods, for example, rely on special variables that are very difficult to find and observe, and the assumptions about these variables are often untestable (Lousdal 2018).

One way to address unmeasured confounding is data fusion, which combines data from a small number of unbiased randomized controlled trials with a large amount of biased observational data to jointly estimate causal effects. Data fusion methods can be categorized into one-stage and two-stage approaches, where the former focuses on joint learning of causal effects (Wu et al. 2022) as well as propensity scores using the idea of entire space, and the latter focuses on the idea of correcting for the difference between the true CATE and the estimated CATE using RCT data. Some of the two-stage methods consider traditional statistical models such as linear regression (Kallus, Puli, and Shalit 2018), while others are based on machine learning (Hatt, Tschernutter, and Feuerriegel 2022). What these data fusion methods have in common is that all assume that the OBS and the RCT are derived from the same target population.

However, this assumption is difficult to justify in practice. For example, in a healthcare scenario, the target population for a new drug may be the entire population, but clinical trials rely on volunteer enrollment. Pregnant women and the elderly

[*]This work was done during the research internship of Yaxuan Li at Peking University.

[†]Chuan Zhou is the corresponding author.

often do not tend to enroll, and children and patients with underlying medical conditions are often excluded from trials (Ronconi, Shiner, and Watts 2014). This makes RCT data not a representative sample of the target population (Kennedy-Martin et al. 2015).

To fill this gap, we consider relaxing this assumption in this paper. Specifically, it is a reality in our clinical trials that RCTs are a biased sample of the target population. We propose a two-stage reweighting method that eliminates the confounding bias in the OBS data and selection bias in the RCT data to obtain an unbiased estimate of CATE. The contribution of this paper can be summarized as follows:

- We relax the assumption that RCT and OBS data come from the same target population in CATE estimation under unmeasured confounding.
- We propose a reweighting-based approach for CATE estimation under unmeasured confounding with non-representative randomized data.
- We conduct extensive experiments on a public semi-synthetic dataset to demonstrate the effectiveness of our method in CATE estimation.

Preliminaries

Problem Setup

We review the estimation of the conditional average treatment effect (CATE) for binary treatment with large-scale OBS and small-scale RCT data. We have m units ($i = 1, 2, \ldots, m$) from the OBS data and n units ($i = m+1, \ldots, m+n$) from the RCT data, where $m, n \in \mathbb{Z}_+$ and $m > n$. For each unit, we observe a four-element tuple (T, X, Y, G) following distribution P, where $T \in \{0, 1\}$ is an indicator for binary treatment, with $T = 1$ for the treatment group and $T = 0$ for the control group. $X \in \mathcal{X} \subseteq \mathbb{R}^d$ is a covariate vector of the unit with d dimensions, either discrete or continuous. $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is the observed outcome for the unit, in our work, it can be either binary or continuous. $G \in \{\mathcal{R}, \mathcal{O}\}$ is an indicator of data source, with $G = \mathcal{R}$ for the RCT group and $G = \mathcal{O}$ for the OBS group.

In this paper, we consider the scenario when RCT data is non-representative. In other words, we assume the RCT data is randomly sampled from the target population, while the RCT data is with selection bias. As for the covariate distributions, we have

$$P(X|G = \mathcal{R}) \neq P(X|G = \mathcal{O}). \tag{1}$$

Using the Neyman–Rubin causal model, we let $Y(1)$ be the potential outcome under treatment and $Y(0)$ be the potential outcome under control. The estimand of interest is the CATE

$$\tau(x) = \mathbb{E}(Y(1) - Y(0)|X = x), x \in \mathcal{X}. \tag{2}$$

Note that we implicitly require that the CATE in the OBS group is identical to that of the RCT group, i.e., $\mathbb{E}(Y(1) - Y(0)|X = x, G = \mathcal{O}) = \mathbb{E}(Y(1) - Y(0)|X = x, G = \mathcal{R})$, thus omitting the target population condition in $\tau(x)$. To estimate CATE with the RCT and OBS data, apart from the stable unit treatment value assumption (SUTVA), we assume $Y = Y(1)T + Y(0)(1-T)$, i.e., we can only observe one of the potential outcomes corresponding to the actual treatment received. We also require the positivity assumption, formally $0 < P(T = 1 \mid X = x) < 1$ for all $x \in \mathcal{X}$.

Since randomization is guaranteed by the randomized trial, strong ignorability holds for the RCT data, formally,

$$(Y(0), Y(1)) \perp\!\!\!\perp T|X, G = \mathcal{R}. \tag{3}$$

Meanwhile, for the OBS data, we consider the case where there is unmeasured confounding and the strong ignorability does not hold, i.e.,

$$(Y(0), Y(1)) \not\perp\!\!\!\perp T|X, G = \mathcal{O}. \tag{4}$$

Previous Work

Based on the aforementioned assumptions especially (3), one can identify the CATE with RCT data as follows.

Lemma 1.

$$\tau(x) = \mathbb{E}(Y \mid T = 1, X = x, G = \mathcal{R}) - \mathbb{E}(Y \mid T = 0, X = x, G = \mathcal{R}).$$

The identification result itself implies an unbiased estimator for CATE. Nevertheless, randomized controlled trials are not conducted on a large scale due to their high cost and ethical and regulatory constraints. Consequently, the amount of data is frequently less extensive than that derived from observational studies. Consequently, the utilization of RCT data alone may result in overfitting for a learning algorithm.

The estimation of CATE using a combination of large-scale OBS data and small-scale RCT data has been previously proposed as a potential approach. Due to the presence of unmeasured confounding, the average outcome of OBS data is biased. Let

$$\omega(x) = \mathbb{E}(Y \mid T = 1, X = x, G = \mathcal{O}) - \mathbb{E}(Y \mid T = 0, X = x, G = \mathcal{O})$$

be the average outcome difference in the OBS group, then the presence of unmeasured confounding and violation of strong ignorability (4) indicates that $\omega(x) \neq \tau(x)$. We denote the residuals as $\eta(x) = \tau(x) - \omega(x)$.

One approach utilizes the OBS data to derive a biased estimate $\hat{\omega}(x)$, fits the residual $\hat{\eta}(x)$, and adds $\hat{\omega}(x)$ and $\hat{\eta}(x)$ to obtain an unbiased estimate of CATE. The alternative methodology employs a two-stage pretraining-finetuning framework (TSPF), whereby the OBS data are used to obtain an initial estimator model, which is then finetuned with the RCT data to yield an unbiased estimator (Zhou et al. 2025).

Proposed Method

In the presence of unmeasured confounding, previous data fusion methods have often implicitly assumed that RCT data and OBS data come from the same target population. However, randomized trials in the real world in medicine, psychology, and other fields often have strict inclusion-exclusion criteria or have higher propensities to recruit certain groups of people. We therefore consider relaxing the assumption and consider the case where the OBS data is representative yet the RCT data is non-representative of the target population.

Note that with the biased-sampled RCT data, we can still identify the CATE using Lemma 1. However, due to the inconsistency of the distribution of covariate X on the RCT

Algorithm 1: Learning algorithm for potential outcome estimation combining large-scale OBS data and non-representative RCT data.

Input: OBS data $\mathcal{D}^{OBS} = \{(X_i, T_i, Y_i, G_i = \mathcal{O})\}_{i=1}^{m}$, RCT data $\mathcal{D}^{RCT} = \{(X_i, T_i, Y_i, G_i = \mathcal{R})\}_{i=m+1}^{m+n}$

Compute $w_i = \frac{T_i}{2u} + \frac{1-T_i}{2(1-u)}$ with $u = \frac{1}{m}\sum_{i=1}^{m} T_i$ for $i = 1, ..., m$;

Train the first-stage potential outcome estimation model $f(X, T; \theta_f)$ with $\mathcal{D}^{OBS}$ and weights w_i;

With data $\mathcal{D}^{RCT} \cup \mathcal{D}^{OBS}$, train a model $q(X; \theta_q)$ to predict the probability that the data comes from $\mathcal{D}^{RCT}$ with only covariate X as the input;

Initialize the second-stage potential outcome estimation model $g(X, T; \theta_g)$, where part of the parameters are initialized with that of the first-stage model $f(X, T; \theta_f)$ to make sure $g(x, t; \theta_g) = f(x, t; \theta_f), \forall x \in \mathcal{X}, t \in \{0, 1\}$;

Compute $v_i = (\frac{T_i}{2u} + \frac{1-T_i}{2(1-u)}) \cdot \frac{1-q(X_i;\theta_q)}{q(X_i;\theta_q)} \cdot \frac{n}{m}$ with $u = \frac{1}{n}\sum_{i=m+1}^{m+n} T_i$ for $i = m+1, ..., m+n$;

Finetune the second-stage potential outcome estimation model $g(X, T; \theta_g)$ with $\mathcal{D}^{RCT}$ and weights v_i;

Output: The potential outcome estimation model $g(X, T; \theta_g)$.

with that on the OBS, direct use of the raw RCT data on a learning approach may result in slow convergence, especially with limited data size (Hatt et al. 2022). Take a binary covariate such as biological sex for example. If males are much more likely to participate in a randomized controlled trial, this may result in a very small sample of females in the RCT data and thus the algorithm may be hard to converge and results in large error.

Inspired by the TSPF (Zhou et al. 2025), in this paper we adopt a two-stage approach named W-TSPF, where the OBS data is used to train a CATE estimation model containing a representation module and two predictor heads in the first stage, while the non-representativeness of the RCT data is addressed in the second stage using a re-weighting method.

Two-Stage Framework

We adopt the TSPF framework (Zhou et al. 2025) to build our method. In this framework, the models in the two stages are different. In the first stage, we follow the methodology of previous work to train a neural network $f(X, T; \theta_f)$ consisting of a representation module $\phi(X, \theta_\phi)$ to balance the covariate, a reconstruction module $\psi(\phi; \theta_\psi)$ to ensure not much information of covariate is lost, and two prediction heads $h_0(\phi; \theta_{h_0}), h_1(\phi; \theta_{h_1})$ to predict the potential outcomes. Note that in the first stage, we only use the OBS data and adopt a weighting approach to address the problem of uneven sample sizes in the treatment and control groups. The sample weights are given by $w_i = \frac{T_i}{2u} + \frac{1-T_i}{2(1-u)}$, where $u = \frac{1}{n}\sum_{i=1}^{m} T_i$. The weights are added on the empirical risk

$$\mathcal{L}_f = \frac{1}{m}\sum_{i=1}^{m} w_i \cdot l(Y_i, f(X_i, T_i; \theta_f)),$$

where l is a loss function like mean squared error (MSE).

In the second stage, we use another neural network $g(X, T; \theta_g)$ consisting of a representation module ϕ directly adopted from f, an augmented representation module $\tilde{\phi}(X, \theta_{\tilde{\phi}})$, and two prediction heads $\tilde{h}_0(\phi, \tilde{\phi}; \theta_{\tilde{h}_0}), \tilde{h}_1(\phi, \tilde{\phi}; \theta_{\tilde{h}_1})$. Note that the reconstruction module ψ is removed. $\tilde{h}_0$ and $\tilde{h}_1$ are an neural network augmented from h_0 and h_1 respectively. Specifically, $\tilde{h}_0$ and $\tilde{h}_1$ have the same layer numbers as h_0 and h_1, yet the layers in $\tilde{h}_0$ and $\tilde{h}_1$ are wider. During the second-stage training, we fix the representation module ϕ, leaving only $\tilde{\phi}, \tilde{h}_0$ and $\tilde{h}_1$ to be trainable. In the original framework, the training data of the second stage is from the RCT group. However, we aim to address the non-representativeness of the RCT data, thus motivated to combine both OBS and RCT data in this stage.

Re-weighting in the Second Stage

We aim to re-weight the samples in RCT data to adapt the covariate distribution $P(X|G = \mathcal{R})$ to that of the OBS data. For simplicity, we denote the covariate distribution of the OBS group as $P_{\mathcal{O}}(X)$ and the covariate distribution of the RCT group as $P_{\mathcal{R}}(X)$. We can get the following weights inspired by Colnet et al. (2022).

Proposition 1. *Given the assumption $P_{\mathcal{R}}(x) > 0, \forall x \in \mathcal{X}$, the weight $\frac{P_{\mathcal{O}}(X)}{P_{\mathcal{R}}(X)}$ can adjust the covariate distribution of the RCT group to that of the OBS group.*

The Proposition 1 can be validated as follows. Consider an arbitrary integrable function $\zeta(\cdot)$, we have $\mathbb{E}(\zeta(X) \mid G = \mathcal{R}) = \int_{\mathcal{X}} \zeta(x) P(X = x \mid G = \mathcal{R}) \mathrm{d}x$. The re-weighting lead us to $\int_{\mathcal{X}} \zeta(x) P(X = x \mid G = \mathcal{R}) \cdot \frac{P_{\mathcal{O}}(x)}{P_{\mathcal{R}}(x)} \mathrm{d}x = \int_{\mathcal{X}} \zeta(x) P(X = x \mid G = \mathcal{O}) \mathrm{d}x = \mathbb{E}(\zeta(X) \mid G = \mathcal{O})$. In other words, the re-weighting shifts the expectation of the RCT population to that of the targeted OBS population.

From the discussion above we can see that adding the weight $\frac{P_{\mathcal{O}}(X)}{P_{\mathcal{R}}(X)}$ can address the non-representativeness of RCT data theoretically. However, when the covariate X is high-dimensional, which is very likely in real-world data, to estimate the covariate distributions $P_{\mathcal{O}}(X)$ and $P_{\mathcal{R}}(X)$ directly is extremely challenging.

We propose to transform the weight using the Bayesian formula to obtain an equivalent form that is easier to estimate.

Lemma 2. $\frac{P_{\mathcal{O}}(x)}{P_{\mathcal{R}}(x)} = \frac{P(G=\mathcal{O}|X=x)}{P(G=\mathcal{R}|X=x)} \cdot \frac{P(G=\mathcal{R})}{P(G=\mathcal{O})}$.

Notice that the right-hand side of Lemma 2 consists of four terms$P(G = \mathcal{O} \mid X = x), P(G = \mathcal{R} \mid X = x), P(G = \mathcal{R})$ and $P(G = \mathcal{O})$. The first two terms, i.e., the probabilities of being in the OBS group or RCT group given covariate $X = x$, can be directly estimated via logistic regression. The last two terms, i.e., the marginal distribution of G, can be estimated with the sample size of RCT and OBS data. After fitting a model $q(X; \theta_q)$ to estimate the probability

$P(G = \mathcal{R} \mid X)$, then another probability $P(G = \mathcal{O} \mid X)$ would be $1 - q(X; \theta_q)$. The maximum likelihood estimator will be $\frac{n}{m+n}$ and $\frac{m}{m+n}$ for $P(G = \mathcal{R})$ and $P(G = \mathcal{O})$.

We use the plug-in estimator combined with the weights balancing treatment and control groups to obtain the ultimate weights for RCT data as follows.

$$v_i = (\frac{T_i}{2u} + \frac{1 - T_i}{2(1-u)}) \cdot \frac{1 - q(X_i; \theta_q)}{q(X_i; \theta_q)} \cdot \frac{n}{m},$$

with $u = \frac{1}{n} \sum_{i=m+1}^{m+n} T_i$ for $i = m+1, ..., m+n$. As in the first stage, we only add weights on the empirical risk

$$\mathcal{L}_g = \frac{1}{n} \sum_{i=1}^{n} v_i \cdot l(Y_i, g(X_i, T_i; \theta_g)).$$

The overall algorithm of our method is presented in Algorithm 1. Note that the algorithm outputs a potential outcome estimation model with two heads estimating $\mathbb{E}(Y(0) \mid X)$ and $\mathbb{E}(Y(1) \mid X)$ respectively. For CATE estimation, we need to compute the difference between the outputs of two heads $\tilde{h}_1$ and $\tilde{h}_0$.

Experiments

Dataset and Preprocessing

Following previous work, we validate our approach on a publicly available semi-synthetic dataset, **IHDP** (Hill 2011). The original dataset includes 747 units with 25 covariates. 19% of the units are treated while 81% of them are in the control group. We divide the dataset into training, validation and test sets with the ratio 80/10/10. For sample i in the index set of the training samples $\mathcal{T}$, we randomly assign them into RCT group with probability $\frac{\beta \cdot \sigma(W \cdot X_i)}{\sum_{j \in \mathcal{T}} \sigma(W \cdot X_j)}$, where $\beta = 0.1$ is a hyperparameter controlling the ratio of RCT samples in the whole training set, $\sigma(\cdot)$ is the sigmoid function, and $W \sim \mathcal{N}(a\mathbf{1}_d, \mathbf{I}_d)$. We set $a = 5, 10, 20$ to generate RCT samples with different covariate distribution patterns. For the training samples assigned into the RCT group and all the validation samples, we randomly re-assign treatment to simulate randomized trial.

$$T_{new} = \mathrm{Bern}(0.5), \quad Y_{new} = \mathbb{I}\{T_{new} = T\}(Y_f - Y_{cf}) + Y_{cf},$$

where the $\mathrm{Bern}(\cdot)$ is the Bernoulli distribution, Y_f is the factual outcome, and Y_{cf} is the counterfactual outcome.

Baselines and Evaluation Metrics

We choose two important and commonly used baselines in CATE estimation performance comparison, **T-learner** (Künzel et al. 2019) which has a simple model architecture and **DragonNet** (Shi, Blei, and Veitch 2019) which integrates propensity score in the model. We also include two-stage methods **CorNet** (Hatt, Tschernutter, and Feuerriegel 2022) and **TSPF** (Zhou et al. 2025) to further validate the effectiveness of the re-weighting strategy.

Following previous work in CATE estimation (Shalit, Johansson, and Sontag 2017), we evaluate the performance with the square root of Precision in Estimation of Heterogeneous Effects (PEHE), as well as the Average Treatment Effect (ATE). The definitions of the metrics are as follows.

$$\sqrt{\epsilon_{\text{PEHE}}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} ((\hat{Y}_{i,1} - \hat{Y}_{i,0}) - (Y_{i,1} - Y_{i,0}))^2},$$

$$\epsilon_{\text{ATE}} = \frac{1}{N} | \sum_{i=1}^{N} ((\hat{Y}_{i,1} - \hat{Y}_{i,0} - (Y_{i,1} - Y_{i,0})) |.$$

Performance Analysis

The performance of the baselines and our method in the case of $a = 5, 10, 20$ is shown in Tables 1. Our method outperforms the baselines in all scenarios on both metrics. Particularly, the significant superiority against the other two-stage methods CorNet and TSPF demonstrats the effectiveness of the reweighting strategy.

Table 1: Performance on IHDP dataset with $a = 5, 10, 20$.

$a = 5$	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
T-learner	3.72 ± 0.46	2.19 ± 0.95
DragonNet	3.85 ± 0.39	2.27 ± 0.51
CorNet	3.16 ± 0.21	1.92 ± 0.63
TSPF	2.86 ± 0.47	1.65 ± 0.80
W-TSPF	$\mathbf{1.97 \pm 0.31}$	$\mathbf{0.72 \pm 0.53}$
$a = 10$	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
T-learner	3.81 ± 0.45	2.34 ± 1.01
DragonNet	3.99 ± 0.43	2.54 ± 0.55
CorNet	3.25 ± 0.28	2.06 ± 0.74
TSPF	2.97 ± 0.53	1.78 ± 0.85
W-TSPF	$\mathbf{2.02 \pm 0.34}$	$\mathbf{0.77 \pm 0.52}$
$a = 20$	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
T-learner	4.08 ± 0.52	2.49 ± 1.06
DragonNet	4.17 ± 0.50	2.68 ± 0.61
CorNet	3.33 ± 0.39	2.13 ± 0.77
TSPF	3.14 ± 0.59	1.90 ± 0.83
W-TSPF	$\mathbf{2.08 \pm 0.30}$	$\mathbf{0.84 \pm 0.62}$

Conclusion

In the presence of unmeasured confounding, previous methods of estimating treatment effects combining OBS and RCT data have tended to rely on an implicit assumption that both the OBS and the RCT are unbiased samples of the target population. This paper aims to relax this assumption by considering a scenario where OBS data consists of representative samples while RCT data does not. We propose a re-weighting strategy based on a two-stage pre-training fine-tuning framework to adapt the samples in the RCT group to represent the target population. Experiments conducted on a semi-synthetic dataset IHDP demonstrate the effectiveness of our method. One limitation of this paper is the assumption that the support set of the covariate distribution function of the RCT group is the full set, since inclusion criteria may exclude parts of the population from randomized controlled trials in real studies.

References

Athey, S. 2015. Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 5–6.

Borgonovo, E.; and Plischke, E. 2016. Sensitivity analysis: A review of recent advances. *European Journal of Operational Research*, 248(3): 869–887.

Campbell, J. 2007. An interventionist approach to causation in psychology. *Causal learning: Psychology, philosophy, and computation*, 58–66.

Colnet, B.; Josse, J.; Varoquaux, G.; and Scornet, E. 2022. Reweighting the RCT for generalization: finite sample analysis and variable selection. *arXiv:2208.07614*.

Dorn, J.; Guo, K.; and Kallus, N. 2024. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *Journal of the American Statistical Association*, 1–12.

Fewell, Z.; Davey Smith, G.; and Sterne, J. A. 2007. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *American Journal of Epidemiology*, 166(6): 646–655.

Gangl, M. 2010. Causal inference in sociological research. *Annual Review of Sociology*, 36(1): 21–47.

Hatt, T.; Berrevoets, J.; Curth, A.; Feuerriegel, S.; and van der Schaar, M. 2022. Combining observational and randomized data for estimating heterogeneous treatment effects. *arXiv:2202.12891*.

Hatt, T.; Tschernutter, D.; and Feuerriegel, S. 2022. Generalizing off-policy learning under sample selection bias. In *Uncertainty in Artificial Intelligence*, 769–779. PMLR.

Hill, J. L. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1): 217–240.

Imbens, G. W. 2003. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2): 126–132.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Joshua D. Angrist, G. W. I.; and Rubin, D. B. 1996. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434): 444–455.

Kallus, N.; Puli, A. M.; and Shalit, U. 2018. Removing hidden confounding by experimental grounding. *Advances in Neural Information Processing Systems*, 31.

Kennedy-Martin, T.; Curtis, S.; Faries, D.; Robinson, S.; and Johnston, J. 2015. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials*, 16: 1–14.

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.

Lousdal, M. L. 2018. An introduction to instrumental variable assumptions, validation and estimation. *Emerging Themes in Epidemiology*, 15(1): 1.

Orri, M.; Tremblay, R. E.; Japel, C.; Boivin, M.; Vitaro, F.; Losier, T.; Brendgen, M. R.; Falissard, B.; Melchior, M.; and Côté, S. M. 2019. Early childhood child care and disruptive behavior problems during adolescence: A 17-year population-based propensity score study. *Journal of Child Psychology and Psychiatry*, 60(11): 1174–1182.

Prosperi, M.; Guo, Y.; Sperrin, M.; Koopman, J. S.; Min, J. S.; He, X.; Rich, S.; Wang, M.; Buchan, I. E.; and Bian, J. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7): 369–375.

Reich, B. J.; Yang, S.; Guan, Y.; Giffin, A. B.; Miller, M. J.; and Rappold, A. 2021. A review of spatial causal inference methods for environmental and epidemiological applications. *International Statistical Review*, 89(3): 605–634.

Robins, J. M.; and Hernán, M. 2016. Causal inference.

Ronconi, J. M.; Shiner, B.; and Watts, B. V. 2014. Inclusion and exclusion criteria in randomized controlled trials of psychotherapy for PTSD. *Journal of Psychiatric Practice®*, 20(1): 25–37.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, 3076–3085. PMLR.

Shi, C.; Blei, D.; and Veitch, V. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32.

Shi, X.; Miao, W.; Nelson, J. C.; and Tchetgen Tchetgen, E. J. 2020. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(2): 521–540.

VanderWeele, T. J.; and Arah, O. A. 2011. Unmeasured confounding for general outcomes, treatments, and confounders: bias formulas for sensitivity analysis. *Epidemiology*, 22(1): 42.

Wang, W.; Zhang, Y.; Li, H.; Wu, P.; Feng, F.; and He, X. 2023. Causal recommendation: Progresses and future directions. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3432–3435.

Wu, A.; Kuang, K.; Li, B.; and Wu, F. 2022. Instrumental variable regression with confounder balancing. In *International Conference on Machine Learning*, 24056–24075. PMLR.

Zhang, X.; Faries, D. E.; Li, H.; Stamey, J. D.; and Imbens, G. W. 2018. Addressing unmeasured confounding in comparative observational research. *Pharmacoepidemiology and Drug Safety*, 27(4): 373–382.

Zhou, C.; Li, Y.; Zheng, C.; Zhang, H.; Zhang, M.; Li, H.; and Gong, M. 2025. Your Neighbor Matters: Towards Fair Decisions Under Networked Interference. In *Proceedings of the 31th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '25.