Title
Recent Developments in Causal Inference and Machine Learning

Journal
Annual Review of Sociology, 49(1)

ISSN
0360-0572

Authors
Brand, Jennie E
Zhou, Xiang
Xie, Yu

Publication Date
2023-07-31

DOI
10.1146/annurev-soc-030420-015345

Peer reviewed

Recent Developments in Causal Inference and Machine Learning

Jennie E. Brand[1] Xiang Zhou[2] Yu Xie[3]

December 2, 2022

In preparation for *Annual Review of Sociology*

Word count = 10,587
Abstract word count = 154

[1] Professor of Sociology and Statistics, UCLA, Director, California Center for Population Research, Co-Director, Center for Social Statistics, brand@soc.ucla.edu.
[2] Associate Professor of Sociology, Harvard University, xiang_zhou@fas.harvard.edu.
[3] Bert G. Kerstetter '66 University Professor of Sociology, Princeton University, yuxie@princeton.edu.

Abstract

This paper provides a review of recent advances in causal inference relevant to sociology. We focus on a selective subset of contributions aligning with four broad topics: causal effect identification and estimation in general, causal effect heterogeneity, causal effect mediation, and temporal and spatial interference. We describe how machine learning, as an estimation strategy, can be effectively combined with causal inference, which has been traditionally concerned with identification. The incorporation of machine learning in causal inference enables researchers to better address potential biases in estimating causal effects and uncover heterogeneous causal effects. Uncovering sources of effect heterogeneity is key for generalizing to populations beyond those under study. While sociology has long emphasized the importance of causal mechanisms, historical and life-cycle variation, and social contexts involving network interactions, recent conceptual and computational advances facilitate more principled estimation of causal effects under these settings. We encourage sociologists to incorporate these insights into their empirical research.

Keywords: causal inference; counterfactuals; machine learning; treatment effect heterogeneity; mediation; extrapolation; external validity;

Recent Developments in Causal Inference and Machine Learning

1 Introduction

Many important questions in the social sciences, and everyday life, are causal questions. For example, we want to know how parental divorce affects children, how attending college affects job prospects, or how moving to a new neighborhood affects children's academic performance. We ask what would happen if individuals did or did not experience an event, like divorcing or attending college. Since reviews in sociology by Winship and Morgan (1999) and Gangl (2010), the literature on causal inference has developed several new promising directions. Some of the most exciting areas of development lie at the intersection of causal inference with machine learning (Athey & Imbens 2017, 2019; Huber 2021). This review describes several key identification strategies for causal inference and how machine learning methods can enhance our estimation of causal effects. Throughout our review, we describe some empirical applications of these methods in sociology.[4]

We emphasize four main principles in our review. First, the plausibility of the assumptions underlying different research designs and identification strategies varies by applications. Machine learning methods adapted to causal tasks facilitate estimation, but like other estimation tools, they do not assure identification of causal effects. Second, causal effect heterogeneity is the norm, and it complicates extrapolation. Researchers may exert considerable effort in establishing a model with high internal validity, or credibility of the estimator of the causal effect of interest, but with low external validity, or limited

[4] Our review differs from recent reviews in sociology (Lundberg et al. 2022; Molina & Garip 2019) and political science (Grimmer et al. 2021) on machine learning in that we focus on the intersection between causal inference and machine learning. See also Hastie et al. (2017) for a textbook treatment of statistical learning.

generalizability of the causal effect to other populations. To understand the population distribution of causal effects, we need to assess causal effect heterogeneity. Machine learning methods can help identify subpopulations most responsive to treatments. Third, when assessing social mechanisms in sociological research, we need to attend to confounding along the causal pathway, i.e., for not only the treatment-outcome relationship but also the treatment-mediator and mediator-outcome relationships. Fourth, temporal and spatial interference, typical in social settings, complicates the definition, identification, and estimation of causal effects. These complications should be addressed more routinely in sociological research. In the following sections, we discuss (1) effect identification and estimation, (2) effect heterogeneity, (3) effect mediation, and (4) temporal and spatial interference. We conclude with some general remarks.

2 Causal Effect Identification and Estimation

2.1. Notation and Estimands

Empirical work can be descriptive, such that we establish facts through associations between observables. For example, we might observe that college graduates earn higher wages than non-college graduates. But to evaluate causal effects, we draw on counterfactuals, i.e., we ask how much college-educated individuals would have earned without a college degree. The potential outcomes framework offers a conceptual apparatus for defining causal effects. The framework has roots in research on experiments by Fisher (1935) and Neyman (1923) and research in economics by Roy (1951) and Quandt (1972). Rubin formalized and extended the potential outcomes framework in a series of papers in statistics in the 1970s and 1980s (e.g., Rubin 1974, 1977, 1986).

Let us define a treatment W, e.g., an event or intervention, applied to unit i, i being a member in a population. A unit exposed to a treatment ($W_i = 1$) at a specific time could

have been exposed to an alternative treatment (i.e., control, $W_i = 0$) at the same time. For example, a person who attended college could have instead not attended college. We assume units assigned to treatment and control groups have potential outcomes in both states, the ones in which they are observed and unobserved. For a binary treatment, let Y be an outcome of interest and Y_{i1} and Y_{i0} the potential outcomes for unit i that would result from exposure to the treatment and control states, respectively. The causal effect of the treatment is thus the difference between the potential outcomes (i.e., $Y_{i1} - Y_{i0}$). The fundamental problem of causal inference is that we cannot observe both potential outcomes (Holland 1986). This framework is often applied to binary treatments, although extending to multicategory treatments is conceptually straightforward. We may also consider continuous treatments, but in this case the number of potential outcomes becomes infinite, rendering the framework more complex (Gill and Robins 2001).[5] For each unit, we assume that the treatment status and potential outcomes determine the observed outcome. Let us focus on the case of binary treatment conditions. We have $Y_i = W_i Y_{i1} + (1 - W_i) Y_{i0}$. The stable unit treatment value assumption (SUTVA) (Rubin 1986) implies that the potential outcomes for any unit do not vary with the treatment assigned to other units. In other words, there is no interference between units. However, in many social settings, SUTVA can be problematic. For example, the wages for one college graduate may be affected by the population proportion of workers completing college.

Following Heckman and Robb (1986), we assume that treatment effects are heterogeneous. Using the potential-outcomes notation, we smooth out that heterogeneity

[5] Kennedy et al. (2017) develop non-parametric methods for doubly robust estimation of continuous treatment effects.

and define different estimands for specific populations of interest.[6] The average treatment effect (ATE) is the average of individual treatment effects in the population:

$$\tau_{ATE} = E[Y_1 - Y_0], \tag{1}$$

where we omit the unit subscript i for conciseness. The average treatment effect on the treated (ATT) is the average of individual effects among the treated subpopulation:

$$\tau_{ATT} = E[Y_1 - Y_0|W = 1]. \tag{2}$$

Now consider the estimand corresponding to the difference in average outcomes between the treated and control units:

$$\tau = E[Y|W = 1] - E[Y|W = 0]. \tag{3}$$

Following Abadie and Cattaneo (2018), we note:

$$\tau = \tau_{ATE} + b_{ATE} = \tau_{ATT} + b_{ATT}, \tag{4}$$

where b_{ATE} and b_{ATT} are bias terms given by

$$b_{ATE} = (E[Y_1|W = 1] - E[Y_1|W = 0])\Pr(W = 0) + (E[Y_0|W = 1] - E[Y_0|W = 0])\Pr(W = 1), \tag{5}$$

and

$$b_{ATT} = E[Y_0|W = 1] - E[Y_0|W = 0]. \tag{6}$$

If the average potential outcomes under both states are identical between treated and control units, the bias terms b_{ATE} and b_{ATT} disappear. This condition is, however, untestable. Confounding arises when pretreatment characteristics correlated with potential outcomes also influence treatment assignment.

2.2. Experimental Studies

[6] See Lundberg, Johnson, and Stewart (2021) and Lundberg (2022) for discussions on setting theoretical estimands in precise terms, outside of any statistical model.

Randomized experiments, where we randomly assign individuals to treatment and control conditions, offer one strategy to address confounding (Fisher 1935). With successful randomization, experiments generate independence between treatment status and both potential outcomes:

$$(Y_1, Y_0) \perp W, \tag{7}$$

where $\perp$ denotes statistical independence. Consequently, the bias terms in (5) and (6) equal 0, and we can credibly attribute the difference in average outcomes between the treated ($E[Y|W = 1]$) and control groups ($E[Y|W = 0]$) to the treatment. In a traditional experiment, we assign a predetermined number of units to one of two conditions. Note, however, that unless an experiment is conducted on a population-representative sample it is not possible in general to derive the population-level ATE from experimental data. We return to the topic of extrapolating study-specific results in section 3.2.

Recent developments in randomized experiments include adaptive designs for evaluating optimal treatment assignment. For example, multi-armed bandits tailor treatments to individuals when they need to be treated. The design aims to balance the goals of "exploration" (i.e., evaluating the effects of different treatment conditions) and "exploitation" (i.e., assigning units to treatment conditions with higher payoffs) (Athey & Imbens 2019; Carranza et al. 2022; Offer-Westport et al. 2021; Scott 2010). Consider an online setting where treatment is assigned sequentially to different units, and the outcome for each unit is measured quickly after treatment assignment. A multi-armed bandit assigns treatment conditions based on information learned up to the point of the assignment, thus allowing researchers or policymakers to assign more units to conditions with higher payoffs. Sociological applications of multi-armed bandits remain scarce, but it is a promising approach for future studies.

2.3. Observational Studies under Unconfoundedness

For practical and ethical reasons, sociologists cannot address many interesting social questions using experiments, and some scholars have lamented the extent to which randomized experiments dominate the hierarchy of scientific evidence (Abadie & Cattaneo 2018; Deaton & Catwright 2018). But in most observational studies, the independence condition (7) may not hold. In the case of college effects, for example, the simple difference between college and non-college graduates' wages is not a credible estimate of the causal effect due to pretreatment heterogeneity, i.e., that individuals with higher skills and achievement and advantaged social backgrounds disproportionately complete college. We may observe some confounding factors in our data, while others are unobserved. Researchers may assume that after adjusting for a set of pretreatment covariates X, there are no additional confounders that affect both treatment status and the outcome. That is, they assume unconfoundedness (also called ignorability, selection-on-observables, conditional independence, or exogeneity):

$$(Y_1, Y_0) \perp W|X, \tag{8}$$

which allows for the identification of the causal effect of W on Y by adjusting for X. Figure 1 is a directed acyclic graph (DAG) representing the causal relationships between W, X, and Y under unconfoundedness.[7]

[Figure 1 about here]

To estimate the ATE, we also assume positivity, meaning that treatment assignment is probabilistic at all covariate values in the population. Positivity is a strong assumption as it rules out the possibility that treatment status has no variation (i.e., at either 0 or 1)

[7] DAGs represent assumptions about nonparametric relationships between variables (in contrast to path diagrams that reflect linear structural equations) (see Pearl [2009] and Morgan and Winship [2014] for background on the use of DAGs). Edges are directed, such that an arrow indicates the effect of one variable on another, such as the effect of W on Y. They are also acyclic, such that there are no feedback loops.

at some covariate values. The latter might happen by chance even if positivity holds in the population. For example, while there may be young adults from families in the top income decile who did not attend college, such youth may fail to appear in a particular sample. Moreover, near violations of positivity (e.g., very few treated/untreated units at some covariate values) can result in unstable estimates of causal effects for subgroups of the population. In practice, we often trim observations with very high and low estimated treatment probabilities to reduce instability in our estimated effects, or those outside the region of common support, leading to effect estimates that do not fully represent the population. This is an example where we sacrifice a degree of external validity to enhance internal validity.

Under the assumptions of unconfoundedness and positivity, researchers draw on various methods to estimate causal effects, such as regression-imputation, propensity score matching (PSM), and inverse probability weighting (IPW) (see Imbens [2004] or Gangl [2010] for a review of these methods). Using regression-imputation, the researcher fits a regression model for the conditional mean of the outcome Y given treatment status W and pretreatment covariates X, $\mu_W(X) = E[Y|W,X]$, "imputes" the potential outcomes under treatment and control for each unit, $\hat{\mu}_1(X_i) = \hat{E}[Y|W = 1, X_i]$ and $\hat{\mu}_0(X_i) = \hat{E}[Y|W = 0, X_i]$, and estimates the ATE using the average difference between these imputed outcomes:

$$\hat{\tau}_{ATE} = \frac{1}{n}\sum_{i=1}^{n}\big(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\big), \tag{9}$$

where n is the sample size. If the outcome model $\mu_W(X)$ is additive in W and X, $\hat{\tau}_{ATE}$ will reduce to the coefficient on W in the regression model.

Using PSM and IPW, the researcher fits a model for the propensity score, i.e., the conditional probability of treatment given the pretreatment covariates, $p(X) = \Pr[W = 1|X]$, and obtains the estimated propensity score for each unit, $\hat{p}(X_i)$ (Rosenbaum and Rubin 1983). With PSM, the researcher then matches treated and control units with

similar values of the propensity score and use their differences to estimate effects (Abadie & Imbens 2016; Caliendo & Kopeinig 2008; Imbens 2015). Matching algorithms differ primarily in how researchers define the distance between units (e.g., propensity scores), select the number of control units, select controls with or without replacement, and weight multiple control units (Austin & Stewart 2017; Morgan & Harding 2006; Morgan & Winship 2014). Decisions regarding how many controls to use and whether to match with or without replacement involve a bias-variance tradeoff.[8] With IPW, the researcher estimates the ATE using a weighted difference in means:

$$\hat{\tau}_{ATE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{p}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{p}(X_i)}\right). \quad (10)$$

By weighting each unit by the inverse estimated propensity (i.e., $1/\hat{p}(X_i)$ for treated units and $1/(1-\hat{p}(X_i))$ for untreated units), researchers create a weighted sample in which treatment status is expected to be independent of all pretreatment covariates. In other words, if the propensity score model is correct, we expect that treated and control units are balanced in their covariate values.

Regression-imputation, PSM, and IPW involve modeling different parts of the data distribution. While regression-imputation depends on a correctly specified outcome model, PSM and IPW depend on a correctly specified propensity score model. However, correctly specifying either model is difficult, especially when the vector of pretreatment covariates X is high-dimensional. When the outcome or propensity score model is misspecified, the corresponding regression-imputation or matching or IPW estimates can be biased. Misspecification may arise either (1) because we have many potential (observed)

[8] More control units lead to greater efficiency and greater bias, while fewer control units lead to less efficiency and less bias. Allowing replacement increases the average quality of the matches but reduces the number of unique control units used to estimate the counterfactual mean, increasing the estimator's variance (See An & Winship [2017], Imbens [2015], and Imbens & Rubin [2015] for discussion of matching procedures).

confounders of the treatment-outcome relationship in the data; or (2) because the researcher is agnostic about the functional form in which treatment status and the covariates affect the outcome. Both scenarios are common in sociological research. Given the second scenario, researchers may experiment with higher-order and interaction terms. Imbens and Rubin (2015) propose an iterative approach to produce a flexible specification of the propensity score specification. Scholars have also advocated using flexible machine learning methods to fit the outcome or propensity score models. For example, researchers have used classification and regression trees (CART), random forests, and ensemble methods to estimate propensity scores (e.g., Brand et al. 2021; Lee et al. 2010; McCaffrey et al. 2004; Westreich et al. 2010).[9] Scholars should draw on theory in the selection of covariates to include (Cinelli et al.2022; Elwert & Winship 2014; Elwert 2015; Pearl 2009).

In each scenario, however, we face complications because these methods were generally not designed for causal inference. Supervised machine learning methods are designed to minimize prediction errors rather than estimate causal effects. For example, a LASSO regression for the outcome tends to select a subset of the covariates highly predictive of the outcome. Such a subset, however, may not be the optimal subset for estimating the ATE. Furthermore, if we omit covariates highly predictive of treatment status, even if their correlations with the outcome are modest, substantial bias may arise in our treatment effect estimates (Belloni et al. 2014). Similarly, if we use an off-the-shelf machine learning method to fit the propensity score model for matching or IPW, it will seek a model that minimizes

[9] An (2010) describes Bayesian propensity score estimators that model the joint likelihood of both propensity scores and outcomes in one step to incorporate the uncertainty in propensity score estimation. Simulations show that this approach corrects for overly conservative inference based on standard propensity score estimators.

the error of predicting treatment status, which may not be the model that yields the optimal propensity score estimates for balancing covariates between the treated and control units.

Researchers have adapted existing machine learning methods to estimate causal parameters to mitigate these and other concerns central to causal inference. First, to adapt machine learning to the regression-imputation approach, Belloni et al. (2014) propose a "double selection" procedure, in which we fit two LASSO regressions, one for the outcome and one for treatment status. After that, we fit an ordinary least squares regression of the outcome on treatment status and the union of the selected covariates in the first two LASSO regressions. In doing so, researchers adjust for covariates that are important in predicting either the outcome or treatment status, avoiding the bias resulting from a single LASSO regression of the outcome. Künzel et al. (2019) propose a metalearner that can take advantage of any supervised learning algorithm to estimate average treatment effects. They show that the X-learner, using random forest and BART as base learners, performs favorably.

Second, to adapt machine learning for IPW, McCaffrey et al. (2004) proposed fitting the propensity score model using gradient boosting machines (GBM). This approach is a precursor to a literature on calibrated propensity scores (e.g., Imai and Ratcovkic 2014) and balancing weights (e.g., Hainmueller 2012; Zubizarreta 2015; Fong et al. 2018; Athey et al.2018; Zhou & Wodtke 2020). Using optimization methods, researchers choose a set of weights such that in the weighted sample, the treated and control units are either exactly or approximately balanced in pretreatment covariates (by a prespecified balancing metric). This procedure ensures that bias due to covariate imbalance is slight. Zhou (2019), for example, adapts this approach to assess the effect of college completion on intergenerational income mobility.

Finally, machine learning methods are particularly attractive when combined with the so-called "doubly robust estimators" of average treatment effects (Robins and Rotnitzky 1995; Robins, Rotnitzky, and Zhao 1994). Consider the following doubly robust estimator of the ATE:

$$\hat{\tau}_{ATE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{W_i(Y_i-\hat{\mu}_1(X_i))}{\hat{p}(X_i)} - \frac{(1-W_i)(Y_i-\hat{\mu}_0(X_i))}{1-\hat{p}(X_i)}\right). \quad (11)$$

Under the assumptions of SUTVA, unconfoundedness, and positivity, it is consistent for the ATE if either the outcome model $\mu_W(X)$ or the propensity score model $p(X)$, but not necessarily both, is correctly specified (Scharfstein et al. 1999). The double-robustness property occurs because the bias of equation (11) as an estimator of the ATE is governed by the product of two bias terms: (1) the bias of the fitted outcome model $\hat{\mu}_W(X)$ and (2) the bias of the fitted propensity score model $\hat{p}(X)$. Provided one of the two biases converges to zero, the bias of equation (11) will converge to zero. This property motivates what Chernozhukov et al. (2018) call debiased machine learning (DML) of the ATE, i.e., the use of flexible machine learning methods to construct estimates of $\mu_W(X)$ and $p(X)$ in equation (11) (see also van der Laan & Rubin [2006]). Due to the data-driven nature of machine learning methods, they generally do not provide root-n consistent estimates of the $\mu_W(X)$ and $p(X)$ functions.[10] However, because of the multiplicative structure of its bias expression, equation (11) itself remains a root-n consistent estimator of the ATE under mild conditions.[11] By contrast, the biases of the regression-imputation and IPW estimators (equations 9 and 10) do not have such a multiplicative structure, preventing root-n consistent estimation of the ATE when researchers use machine learning to estimate the

[10] Root-n consistency means that the estimator converges on the true value at a rate of $n^{-1/2}$.

[11] This is true if the product of the convergence rates of the machine learning estimators of $\mu_W(X)$ and $p(X)$ is faster than $n^{-1/2}$. We can achieve this property when, for example, both converge to the truth at a faster-than $n^{-1/4}$ rate, which is attainable for many machine learning methods.

outcome or the propensity score model. In a recent application of this approach, Zhou and Pan (Forthcoming) employed a DML approach to assess the heterogeneous effects of college attendance and BA completion on earnings for Black and White Americans.

When DML is used, it is advisable to use sample splitting, whereby, for example, a portion of the data is used as a training sample to estimate the outcome and propensity score models, and another portion is used to evaluate equation (11). This procedure removes the "overfitting bias" of machine learning estimators of the outcome and propensity score models.[12] However, a conventional sample splitting procedure would involve a waste of data. To retain efficiency, researchers may draw on cross-fitting, which includes the following steps (Chernozhukov et al. 2018): (1) randomly partition the sample into J folds, $S_1, S_2, \ldots S_J$, where J is a small number such as five; (2) for each j, obtain a fold-specific estimate of the ATE using only data from S_j, but with the outcome and propensity score models estimated from the remainder of the sample ($S \backslash S_j$); and (3) average these fold-specific estimates to form a final estimate of the target parameter.

Finally, researchers should routinely consider how the results obtained under the unconfoundedness assumption would change if we relaxed that assumption. One common approach is to conduct sensitivity analyses by subtracting a bias term from the point estimate and confidence interval of the estimated treatment effects (VanderWeele & Arah 2011; Gangl 2015). The bias term is equal to the product of two parameters:

$$B = \gamma\lambda, \tag{12}$$

where

[12] Machine learning methods attend to the issue of overfitting more than conventional statistical models (Athey & Imbens 2019). The goal is to select flexible models that fit well, but not so well that out-of-sample prediction is compromised. Regularization techniques calibrate machine learning methods to minimize a loss function and avoid overfitting.

$$\gamma = E(Y|U = 1, W, X) - E(Y|U = 0, W, X) \tag{13}$$

and

$$\lambda = Pr(U = 1|W = 1, X) - Pr(U = 1|W = 0, X). \tag{14}$$

That is, γ is the mean difference in the outcome associated with a unit change in an unobserved binary confounder, U, and λ is the mean difference in the unobserved confounder between treated and control units. See Cinelli and Hazlett (2020) for additional measures and graphical tools for assessing sensitivity to unobserved confounding.

2.4. Quasi-Experimental Designs

In settings where researchers deem the unconfoundedness assumption (8) implausible, they may seek to identify causal effects using quasi-experimental designs, such as instrumental variables or regression discontinuity. Instrumental variables (IV) are widely used in randomized experiments with imperfect compliance and in "natural experiments" using observational data (Angrist, Imbens, and Rubin 1996; Imbens and Angrist 1994). As an example of the latter, several studies have used proximity to a local college as an IV for college attendance to assess the effects of attendance on wages (Card 2001; Deaton 2010). Figure 2 is a DAG representation of the IV design, where an unobserved confounder U may affect both the treatment W and the outcome Y. The instrumental variable Z affects W and can affect Y only indirectly through its effect on W. Exogenous variation in Z, which induces changes in W, is used to identify the causal effect of W on Y. An IV analysis is typically implemented using two-stage least squares (2SLS). In the first stage, a linear model is used to predict treatment status given the IV and a set of pretreatment covariates X. In the

second stage, the outcome Y is regressed on X and the fitted values of W from the first stage, whose coefficient represents the causal effect of W on Y.[13]

[Figure 2 about here]

The IV approach allows for unobserved confounding of the W-Y relationship but relies on other stringent assumptions. First, conditional on the pretreatment covariates X, the instrument must be exogenous. That is, no unobserved confounding exists for the Z-W and Z-Y relationships (i.e., the independence assumption). Second, we assume that the IV affects the likelihood of treatment, even if it does so within a small range (i.e., the relevance assumption). Third, we assume that the IV affects the outcome only indirectly through the treatment (i.e., the "exclusion restriction"). Finally, allowing for heterogeneous treatment effects, we assume that although the instrument may not affect some people, all those affected are affected in the same direction (i.e., the "monotonicity" assumption). With these assumptions in place, researchers have suggested that the 2SLS identifies the local average treatment effect (LATE) for a binary treatment W (Angrist & Pischke 2009):

$$\tau_{LATE} = E[Y_1 - Y_0 | W_1 > W_0], \tag{15}$$

where W_0 and W_1 denote the potential value of treatment when the instrumental variable Z takes the value of 0 and 1, respectively. In actual social settings, the inducement effect of an IV is often small. Low inducement can be a major limitation in IV analysis because it can subject the causal effect estimate to large variance, substantial finite-sample bias, and high sensitivity to violations of the exclusion restriction (Bound et al. 1995). Felton and Stewart (2022) contend that while sociologists have increasingly adopted IV as a strategy, assumptions underlying the model often go unstated and robust uncertainty measures are rarely used. Moreover, the 2SLS approach relies on correct specification of the treatment

[13] See Steiner et al. (2017) for a discussion of graphical models for quasi-experimental designs.

and outcome models, which can be difficult to justify when the pretreatment covariates X are high-dimensional. Blandhol et al. (2022) show that a saturated specification for 2SLS that correctly specifies the relationship between the instruments and the covariates (including interactions) is necessary for the estimator to be interpreted as an average of covariate-specific LATEs. Chernozhukov et al. (2018) outline a DML approach for estimating the LATE, which involves fitting three models: (1) a model for $E[Z|X]$; (2) a model for $E[W|Z,X]$; and (3) a model for $E[Y|Z,X]$. In contrast to 2SLS, the DML approach allows all these models to be fit using flexible machine learning methods, thus reducing model dependency. This approach offers a more principled method for estimating the LATE.

In a regression discontinuity (RD) design, access to treatment is determined by a cutoff value c on a continuous running variable X (see Cattaneo et al. [2019] and Cattaneo &Titiunik [2022]). The RD design assumes that the average response of units just below the cutoff provides a good approximation to the average response that we would have observed for units just above the cutoff had they not been assigned to treatment. Under this assumption, a comparison between units just below and above the cutoff mimics a randomized experiment and reveals a local treatment effect, i.e.:

$$\tau_{RD} = E[Y_1 - Y_0|X = c]. \tag{16}$$

To estimate this quantity, we fit two local linear regressions, one for units below the cutoff and one for units above the cutoff and use their difference in the predicted outcome at $X = c$ as an estimate of τ_{RD} (Imbens & Lemieux 2008). For example, suppose students were admitted to college based on a minimum score on an admission test. Students just above the minimum score are arguably comparable to those just below the minimum score in terms of other characteristics that predict college-going. Around the test score cutoff, we can compare the outcomes of those who are and are not admitted. Yet, we can imagine situations where not everyone admitted to college would choose to attend, in which case we

would have a fuzzy rather than a sharp RD design. This situation allows the cutoff to change treatment status for some yet not all units, i.e., compliers. A fuzzy RD design allows the researcher to identify a local treatment effect among compliers (Hahn et al. 2001), i.e.:

$$\tau_{RD(F)} = E[Y_1 - Y_0 | X = c, W_1 > W_0], \tag{17}$$

where W_0 and W_1 denote the potential value of treatment when the running variable X is just below and above the cutoff. We can estimate equation (17) using a combination of 2SLS and local linear regressions (Imbens & Lemieux 2008).[14] Researchers should assess the validity of the underlying assumptions using supplementary analyses to test for evidence of the manipulation of the cutoff variable and for discontinuities in average covariate values at the threshold. RD methods can have high internal validity for an observational study, but low external validity. Several approaches have been proposed to enable valid extrapolation (Cattaneo & Titiunik 2022).

3 Causal Effect Heterogeneity

As we note above, individuals differ not only in pretreatment characteristics (i.e., pretreatment heterogeneity) but also in how they respond to a common treatment (i.e., treatment effect heterogeneity). Analyses that estimate heterogeneous treatment effects can yield insights into how scarce social resources are distributed in an unequal society and how events differentially impact populations with different expectations of their occurrence (e.g., Brand 2022; Heckman et al. 2018). In some cases, we may hypothesize that an event has significant consequences for some subgroups but less or no effect among others (e.g., Brand et al. 2019b). Scholars may aim to identify the most responsive subgroups to determine

[14] Researchers need also to consider bandwidth selection in RD designs (see Imbens & Lemieux [2008] or Lee & Lemieux [2010] for a discussion).

which individuals benefit most from treatment so that policymakers can better assign different treatments to balance competing objectives, such as reducing costs and maximizing outcomes for targeted groups (Athey & Imbens 2019; Manski & Garfinkel 1992; Zhou and Xie 2019, 2020). An important feature of the potential outcome framework is that it allows for general heterogeneity in treatment effects from the outset. Attending to treatment effect heterogeneity can also help extrapolate findings to diverse populations and contexts.

3.1 Estimating Heterogeneous Causal Effects

Social scientists employ a variety of approaches to estimate heterogeneous effects. Most commonly, researchers partition their samples into subgroups defined by individual characteristics, like gender, race, or social class, to explore variation in treatment effects. Yet, for questions of causal inference, the association between the treatment effect and treatment propensity constitutes a key axis of heterogeneity (Heckman et al. 2006; Brand et al.2013; Xie 2013). One way to identify heterogeneity by selection into treatment is to compare different population parameters. For example, the ATE and ATT may differ. If ATE $>$ ATT, those with a lower propensity of treatment have larger estimated treatment effects, and if ATT $>$ ATE, those with a higher propensity of treatment have larger estimated treatment effects. Or we might directly assess how treatment effects vary by the estimated propensity score (Brand et al. 2013; Xie et al. 2012). For example, Cheng et al. (2021) use growth curve models to assess how the effects of college on long-term wages vary across strata of the estimated likelihood that individuals complete a degree. Alternatively, we can obtain matched differences between treated and control units, plot them along a continuous propensity score axis, and then use local polynomial smoothing to observe variation in effects by the likelihood of treatment. For example, Brand and Simon Thomas (2014) use this approach to explore how the effects of job displacement on children's

educational attainment vary by the likelihood of losing a job. Economists also often compare IV (LATE) estimates with OLS estimates to assess differential response patterns. As we indicate above, with treatment effect heterogeneity, the LATE can differ from the ATE and the ATT.[15]

Researchers tend to base decisions as to which subgroups to explore in analyses of effect heterogeneity on theoretical priors. For example, researchers may stratify by gender or race because they are interested in sociodemographic variation. In contrast to this approach, emerging machine learning methods allow researchers to explore sources of variation that they may not have previously considered (Lundberg et al. 2022; Shu & Ye 2022). For example, we can search for effect heterogeneity by adapting a variable selection algorithm such as LASSO, which automatically selects the more predictive interactions between the treatment and covariates (Imai and Ratkovic 2013). Social scientists have also employed tree-based methods to uncover differential responses to treatment. Decision trees, a widely used machine learning approach, recursively split data into increasingly smaller subsets where data bear greater similarity (Brand et al. 2020).[16] Decision trees are attractive for social research because they are easily interpretable. Causal trees, i.e., decision trees adapted for causal inference, partition the data to minimize heterogeneity in within-leaf treatment effects (Athey & Imbens 2016; Brand et al. 2021).[17] We split the data and

[15] Bloome and Schrage (2021) describe an approach for estimating heterogeneous treatment effects using covariance regression models. They demonstrate the approach by analyzing the effects of sharing information about income inequality on redistributive preferences.

[16] At each decision, splits are chosen by selecting a covariate and threshold that minimize an in-sample loss function. This partitioning process is repeated until a regularization penalty selected through cross-validation limits the depth of the tree.

[17] Causal trees bear similarity to kernel regression or matching methods. We can think of the leaf as defining the set of nearest neighbors for a given target observation in a leaf, and the estimator from a single tree as a

construct a tree using a training sample and estimate leaf-specific treatment effects using an estimation sample. This approach allows researchers to uncover subpopulations of interest that they had not prespecified with greater flexibility by searching over high dimensional functions of covariates.[18] We can then use several methods described above, such as weighting, matching, or machine learning, to estimate leaf-specific effects in the presence of observed confounding. For example, Brand et al. (2021) estimate the effects of college completion on reducing low-wage work with a causal tree. They find that individuals who had the largest effects of college on reducing low-wage work are those with disadvantaged backgrounds and low psychosocial skills.

Single decision trees benefit from interpretability but can be unstable and do not allow causal effects to change more smoothly across covariates. A causal forest builds on the causal tree algorithm by averaging over many trees (Breiman 2001; Wager & Athey 2018; Atheyet al. 2019).[19] In principle, every individual has a distinct estimate. Using this strategy, researchers may consider effect heterogeneity by ranking estimated individual treatment effects and then considering the characteristics of groups in the highest and lowest ranked categories. Recent approaches also combine supervised learning of the response variable with supervised learning of the propensity score to estimate treatment effect heterogeneity. For example, Nie and Wager (2021) describe a general class of two-step algorithms for heterogeneous treatment effect estimation in observational studies.

---

matching estimator with alternative ways of selecting the nearest neighbor to a treated unit (Athey and Imbens 2019).

[18] We also may include the propensity score as an input variable (e.g., Hahn et al. 2020).

[19] Building on the comparison to kernel regression or matching, we can think of a causal forest as an average of matching estimators.

Complementing these forest approaches is a DML approach proposed by Semenova and Chernozhukov (2021). Instead of detecting effect heterogeneity from many covariates, this approach allows researchers to directly estimate conditional average treatment effects (CATEs) given a few prespecified covariates. The method is helpful in applications where the researcher wants to see how the treatment effect differs by selected characteristics such as gender, race, or social class categories. For example, Zhou (2022a) adapts the DML approach to study group-based heterogeneity in the total effect of college attendance and its direct and indirect effects via degree completion (see also Zhou & Pan [Forthcoming]). Closely related to treatment effect heterogeneity is an emerging literature on "policy learning" (Athey & Wager 2021). In this case, researchers learn in a data-driven way the optimal assignment of treatment to specific subgroups defined in terms of observed characteristics (e.g., parental income categories). Accordingly, policymakers can target those for whom the treatment effects are largest. Policy learning is especially useful in settings where we aim to optimize an outcome for a costly treatment (e.g., a social intervention with limited funds to cover treatment costs). Yadlowsky et al. (2021) propose a rank-weighted average treatment effect to determine treatment prioritization rules based on responsiveness to treatment.

In studies considering treatment effect heterogeneity, researchers should consider how unobserved selection may contribute to heterogeneous response patterns. Localized sensitivity analyses should be routinely performed for analyses that involve effects stratified by unit characteristics, propensity scores, or machine learning generated categories. Notably, Zhou and Xie (2019, 2020) also describe the relationship between heterogeneity by observed and unobserved selection into treatment and consider the policy implications under different scenarios as the treated population composition shifts.

3.2 Implications of Heterogeneous Causal Effects for Extrapolation

If effects were the same for everyone, it would be easy to generalize an effect estimate from a sample to a population. Effect heterogeneity complicates the generalizability of average treatment effect estimates. Researchers should consider the population of interest when interpreting treatment effect estimates from heterogeneous subgroups. Social scientists who aim to minimize confounding may draw on experimental or quasi-experimental methods. Yet as a researcher attempts to extrapolate or generalize from a specific group of subjects under study to a target population, the average effects may differ due to compositional differences (Hartman et al. 2015; Kern et al. 2016; Manski & Garfinkel 1992; Westreich et al. 2019). In other words, researchers often face a tradeoff between internal and external validity. Internal validity is our degree of confidence that a causal relationship exists between the treatment and the outcome. External validity is our ability to generalize findings to other populations (Manski 1995; Manski & Garfinkel 1992). The literature in causal inference is primarily concerned with internal validity of a causal relationship, which must be complemented by a focus on generalizable knowledge. Social science and public policy demand greater attention to external validity (Egami & Hartman 2020; Findley et al. 2021).

As described in Section 2, researchers use a variety of strategies to claim the internal validity of their effect estimates. For example, a randomized controlled trial may give us sample average treatment effects free from pretreatment heterogeneity bias. However, we may be limited in our ability to extrapolate and provide estimates of population average treatment effects (Hartman et al. 2015; Stuart et al. 2015; Xie 2013). Indeed, the population of units for which we credibly assess causal effects might be quite small. For example, some experimental effect estimates may only apply to treated units in the specific geographical

setting in which the study was conducted, such as the 1962–67 HighScope Perry Preschool Program conducted in Ypilanti, Michigan (Xie et al. 2020).

Compositional differences may also arise in a dynamic setting where treatment gradually expands over successive segments (Xie 2013). In this case, units with higher treatment propensity are likely overrepresented when the population treated is small. As the treated population expands, the overrepresentation of high propensity treated units declines. This compositional shift among newly recruited units, typically from high to lower propensity units, will impact treatment effect estimates from an experimental study that targets those at the "margin" of being treated. In a social intervention on a graduated schedule where participation is need-based, the poorest individuals may be chosen first and benefit most from the intervention. Researchers may calculate an ATE for the subpopulation subject to the experiment. Yet under these conditions, individuals selected at later stages (i.e., becoming eligible only after the eligibility cut-point is moved up the income distribution) would exhibit lower average treatment effects. Low external validity is problematic for policy purposes, as policymakers require evidence of the effectiveness of interventions for target populations that may differ from those represented by experimental participants.

Similarly, if we adopt an IV design, and have a valid IV, we may have a stronger basis for asserting internal validity than a standard regression approach without an IV, but only for a small segment of the subpopulation induced by the IV. Thus, we may not be able to extrapolate from those induced into treatment to a broader population. For example, we may use college proximity as an instrument for college attendance to assess the effects of college attendance on wages. If those induced into college have different effects than average college-goers, we cannot extrapolate the findings to the broader population (Mogstad & Torgovitsky 2018). With regression discontinuity designs, researchers also need to consider

under what conditions we can extrapolate estimated effects to populations further away from the threshold (Angrist & Rokkanen 2015; Bertanha & Imbens 2020; Dong & Lewbel 2015).[20]

Several approaches may help us move from the sample to the population average treatment effect, such as bias-corrected matching (Hotz et al.2005), propensity score weighting (Cole & Stuart 2010; Stuart et al. 2011), propensity score subclassification (Tipton 2013; Tipton et al. 2014), entropy weighting (Hartman et al. 2015), machine-learning-based estimation of heterogeneous treatment effects (e.g., Kern et al. 2016), and calibration methods to generate balancing weights (Josey et al. 2022). A propensity score approach, for example, models membership in the population versus the experimental sample, and then the propensity scores are used to make the sample subjects resemble the target population (e.g., Xie et al. 2020). This approach can work well when the covariates strongly predict membership in the target population and treatment effect heterogeneity (Pearl & Bareinboim 2014). Machine learning methods can automate the detection of treatment-by-covariate interactions. Kern et al. (2016) show that BART performs reasonably well for extrapolating from the sample to a target population when observed covariates are sufficient for accounting for treatment effect heterogeneity.

4 Causal Effect Mediation

While traditional sociological approaches to mediation analysis relied on parametric structural equation models to define and estimate direct and indirect effects (e.g., Duncan

[20] Regression model estimates from representative samples of the population also face external validity problems, as the units in the sample contribute to the causal effects to differing extents (see Aronow & Samii [2016]).

1966; Alwin & Hauser 1975; Baron & Kenny, 1986; Bollen 2014), a large body of research has emerged within the causal inference literature that disentangles the tasks of causal definition, identification and estimation. Causal mediation analysis seeks to uncover whether and how a treatment affects an outcome by quantifying the pathways through which a causal effect operates. Building upon the potential-outcomes framework and graphical causal models (Pearl 2009), a new body of research has provided model-free definitions of direct and indirect effects (Robins & Greenland 1992; Pearl 2001), established the assumptions needed for identifying these effects (Pearl 2001; Robins 2003), and developed an array of estimation strategies (e.g., VanderWeele 2015, 2016). These tools can help researchers discover mechanistic explanations, build theory, and design policy interventions. Sociologists would do well to consider these conceptual and computational tools in any study involving mechanisms. This section provides a brief review of the causal approach to mediation analysis and its recent developments.

4.1. Estimating Direct and Indirect Effects in a Causal Mediation Analysis

Let M denote a mediator hypothesized to transmit the effect of the treatment W on the outcome Y. For example, Wodtke and Parbst (2017) investigated how school poverty mediates the effect of living in a disadvantaged neighborhood on a student's academic achievement. In this context, W is neighborhood disadvantage, Y academic achievement, M school poverty, and X a set of background characteristics. Figure 3 is a DAG representing the causal relationships involving these variables. Note that the pretreatment covariates X may confound not only the treatment-outcome relationship but also the treatment-mediator and mediator-outcome relationships.

[Figure 3 about here]

The most common approach to assessing causal mediation involves decomposing the total effect of W on Y into two components: an indirect effect operating through the mediator M and a direct effect operating through alternative pathways not explicitly considered in the analysis. In Figure 3, we capture the indirect and direct effects by the causal paths $W \rightarrow M \rightarrow Y$ and $W \rightarrow Y$, respectively. These effects can be defined more formally using the potential-outcomes notation. Specifically, if we use Y_{wm} to denote the potential outcome under treatment status w and mediator value m, and M_w to denote the potential value of the mediator M under treatment status w, we can write the ATE of W on Y as:

$$\tau_{ATE} = E[Y_{1M_1} - Y_{0M_0}], \tag{18}$$

which we can then decompose into the natural indirect effect (NIE) and natural direct effect (NDE):

$$\tau_{ATE} = \underbrace{E\left[Y_{1M_1} - Y_{1M_0}\right]}_{\tau_{NIE}} + \underbrace{E\left[Y_{1M_0} - Y_{0M_0}\right]}_{\tau_{NDE}}. \tag{19}$$

The NIE is the expected difference in the outcome if each unit were treated ($W = 1$) and subsequently exposed to the mediator value they experienced as a result of being treated (M_1) rather than the mediator value they would have experienced had they not been treated (M_0). For example, in Wodtke and Parbst's (2017) study, the NIE gauges the effect of neighborhood disadvantage operating through school poverty by fixing the level of neighborhood disadvantage ($W = 1$) for each student and then comparing students' academic achievements under the levels of school poverty that they would have "naturally" experienced with neighborhood disadvantage (M_1) versus without neighborhood disadvantage (M_0). The NDE, by contrast, reflects the average effect of treatment if the mediator for each unit were fixed at its "natural" level under the reference treatment level ($W = 0$).

Both the NIE and NDE depend on Y_{1M_0}, a variable in which two different levels of W (0 and 1) are nested within the counterfactual for Y. Consequently, this counterfactual does not correspond to any experimental intervention on W and M. That is, to know the value of M_1 for a unit, it is necessary to set W to 1, but then this precludes setting W to 0 for M. The counterfactual Y_{1M_0} is thus called a "cross-world" counterfactual (Robins et al. 2022).

To identify the ATE from observational data, we invoke the unconfoundedness assumption, which states that after adjusting for a set of pretreatment covariates X, no additional confounders exist that affect both treatment status and the outcome. To identify the NIE and NDE, such an unconfoundedness assumption is needed for not only the treatment-outcome relationship but also the treatment-mediator and mediator-outcome relationships. Specifically, the NDE and NIE are nonparametrically identified if, after adjusting for pretreatment covariates X, there is (a) no unobserved treatment-outcome confounding, (b) no unobserved treatment-mediator confounding, and (c) no unobserved mediator-outcome confounding (Imai et al. 2010; VanderWeele & Vansteelandt 2009). Under these assumptions, the mean of the counterfactual $Y_{wM_{w^*}}$ for any $w, w^* \in \{0,1\}$ can be identified using Pearl's (2001) mediation formula:

$$E\left[Y_{wM_{w^*}}\right] = \int E[Y|x,w,m]dP(m|x,w^*)dP(x), \tag{20}$$

where $P(\cdot)$ represents the cumulative distribution function of a random variable. We can use equation (20) to identify the NIE and NDE by setting w and w^* at different values.

Given identification assumptions (a)-(c) and the mediation formula (20), we can use a variety of strategies to estimate the NIE and NDE. Imai et al. (2010) propose a regression-simulation estimator that involves first modeling the conditional mean of the outcome ($E[Y|x,w,m]$) and the conditional distribution of the mediator ($P(m|x,w)$) and then evaluating equation (20) through Monte-Carlo draws from the estimated conditional

distribution of the mediator. This estimator can be viewed as a plug-in estimator of equation (20). Alternatively, one can rewrite equation (20) as $E[E[Y|x,w,M]|x,w^*]$, which leads to a regression-imputation estimator that involves modeling only the conditional means of the outcome (Vansteelandt, Bekaert, and Lange 2012), or as $E[\frac{I(W=w)}{p(W|X)}\frac{p(M|X,w^*)}{p(M|X,w)}Y]$, which leads to a weighting estimator that models the treatment's and the mediator's conditional distributions (VanderWeele 2009). Finally, drawing on semiparametric theory, Tchetgen Tchetgen and Shpitser (2012) develop a "triply robust" estimator of equation (20) that involves fitting three models: (1) a model for the conditional distribution of the treatment given pretreatment covariates (i.e., a propensity score model), (2) a model for the conditional distribution of the mediator given the treatment and pretreatment covariates, and (3) a model for the conditional mean of the outcome given the treatment, mediator, and pretreatment covariates. The resulting estimator is triply robust in that it is consistent if any two of the three models are correctly specified. Moreover, like the doubly robust estimator for the ATE, this triply robust estimator is particularly suitable for using flexible machine learning methods to estimate its nuisance functions (e.g., the treatment, mediator, and outcome models). This fact makes it highly attractive in high-dimensional settings.

The identification assumptions (a)-(c) are strong and unverifiable, and the estimated NIE and NDE can be biased whenever unobserved confounding exists for any of the causal relationships involved. In practice, to assess the robustness of mediation analysis results to different forms of unobserved confounding, one can employ a set of general-purpose bias formulas developed in VanderWeele (2010) and VanderWeele and Arah (2011) (see Brand et al. [2019a] for a recent sociological application).

4.2. Treatment-Induced Confounding

Among identification assumptions (a)-(c), (c) is especially restrictive because it requires that there must not be any observed or unobserved confounders of the mediator-outcome relationship that are affected by the treatment. This assumption is plausible if the treatment and mediator are temporally and mechanistically proximate to each other but likely violated in other settings. For example, Klein and Kühhirt (2021) investigated the role of parental cognitive ability in mediating the effect of grandparents' education on grandchildren's cognitive ability. In this case, it is likely that some posttreatment variables, such as grandparents' income and occupational status, are affected by the treatment (grandparents' education) and affect both the mediator (parental cognitive ability) and the outcome (children's cognitive ability). Figure 4 depicts a DAG where $\boldsymbol{L}$ is a treatment-induced confounder of the mediator-outcome relationship.

[Figure 4 about here]

Treatment-induced confounders pose a dilemma for causal mediation analysis. If they were omitted, our estimated effects of the mediator on the outcome, and by extension, the estimated NIE and NDE, would be biased. However, controlling for treatment-induced confounders is also problematic because it blocks causal pathways and potentially unblocks noncausal pathways from the treatment to the outcome, leading to biased estimates of the NIE and NDE (Elwert & Winship 2014). In fact, the NIE and NDE are not nonparametrically identified in the presence of treatment-induced confounding. Scholars have proposed several strategies to address this challenge. First, the NIE and NDE can be identified in the presence of treatment-induced confounding if we impose an additional assumption positing that the treatment and mediator have no interaction effect on the outcome for each unit (Imai & Yamamoto 2013; Robins 2003). This assumption, however, is implausible in most applications because the no-interaction assumption must hold for

every unit. To overcome this limitation, these scholars have developed sensitivity analysis methods for assessing the robustness of findings to potential violations of the no-interaction assumption.

Second, scholars have proposed an alternative class of estimands known as interventional direct and indirect effects (see Nguyen et al. [2021] for a review). Unlike the NIE and NDE, interventional effects can still be nonparametrically identified in the presence of treatment-induced confounding. Among interventional effects, a special case is the controlled direct effect (CDE), which measures the strength of the treatment-outcome relationship when the mediator is fixed at a given value for all units (Acharya et al. 2018; Pearl 2001; Robins 2003). A nonzero CDE thus implies that the effect of the treatment on the outcome does not operate exclusively through the mediator of interest. For example, in Klein and Kühhirt's (2021) study, a nonzero CDE would imply the effect of grandparent education on grandchildren's cognitive ability does not operate solely through parental cognitive ability.

Apart from the CDE, another set of interventional effects are the so-called randomized interventional analogs to the NDE (rNDE) and the NIE (rNIE) (Didelez et al. 2006; Geneletti 2007; VanderWeele et al. 2014). The rNDE and rNIE are like the NDE and NIE except that, instead of setting the mediator to the level it would have naturally been for each unit under a particular treatment status, these estimands involve setting the mediator to a value randomly drawn from its population distribution under a given treatment status. The rNDE and rNIE thus evaluate the effects of a hypothetical intervention on the distribution of a putative mediator. For example, Wodtke et al. (2020) used the rNDE and rNIE to assess the extent to which school quality mediates the effect of neighborhood disadvantage on children's academic achievement.

Researchers can estimate interventional effects such as the CDE, rNDE, and rNIE via several alternative methods, such as sequential g-estimation (Vansteelandt 2009) and IPW (VanderWeele et al. 2014). More recently, Zhou and Wodtke (2019) proposed the regression-with-residuals (RWR) method, which is algebraically equivalent to sequential g-estimation in special cases, but, unlike the latter, can accommodate several types of effect moderation (see also Wodtke and Zhou [2020]). RWR has been applied in several sociological studies (e.g., Levy et al. 2019; Wodtke et al. 2020; Klein & Kühhirt 2021; Wodtke et al. 2022). Nonetheless, as with sequential g-estimation and IPW, RWR is premised on a set of strong modeling assumptions, which, when violated, can lead to biased estimates. Scholars have recently leveraged semiparametric theory to reduce model dependence and develop more robust estimators of interventional direct and indirect effects (Díaz et al. 2021; Xia et al. 2021). Like the doubly robust estimator for the ATE and the triply robust estimator for the NDE and NIE, researchers can combine these estimators with machine learning to yield optimal performance.

4.3. Causal Mediation Analysis with Multiple Mediators

Researchers often aim to test several "competing hypotheses" of underlying processes when analyzing causal mechanisms, leading to multiple mediators of interest. In the presence of multiple mediators, the prevailing practice is to treat different mediators as causally independent (i.e., assuming they do not affect each other) and then estimate the NIE for each mediator separately. In many applications, however, the mediators are likely causally dependent. In general, if two mediators are present and one mediator affects both the other mediator and the outcome, treating these mediators as causally independent may lead to biased estimates of the NIE for the second mediator. This is true because it fails to account for the first mediator as a potential confounder of the relationship between the second

mediator and the outcome. However, to the extent that the first mediator is affected by the treatment, it is a treatment-induced confounder, which renders the NIE for the second mediator non-identifiable without functional form assumptions. In such cases, we could attempt to evaluate the NIE via additional assumptions and sensitivity analysis (Imai & Yamamoto 2013) or consider interventional effects such as the rNIE.[21]

Apart from interventional effects, other mediation estimands that can still be identified in the presence of multiple causally dependent mediators are path-specific effects (PSEs; Avin et al. 2005), in what Duncan (1966) called "a simple causal chain" in path analysis. Specifically, suppose we have K causally ordered mediators M_1, M_2, ... M_K that lie on the causal paths from W to Y. Then, under the assumption that no unobserved confounding exists for any of the treatment-mediator, treatment-outcome, and mediator-outcome relationships, the ATE can be decomposed into $K + 1$ PSEs: one "direct effect" and K mutually exclusive indirect effects that each reflect the contribution of a specific mediator beyond the contributions of its preceding mediators (Daniel et al. 2015; Zhou & Yamamoto 2022). Like the NIE and NDE, researchers can estimate these PSEs via regression-simulation (Miles et al. 2017), regression-imputation (Zhou & Yamamoto 2022), IPW (VanderWeele et al. 2014), or multiply robust methods that are amenable to machine learning estimation of its nuisance functions (Miles et al. 2020; Zhou 2022b). In a recent study, Ahearn et al. (2022) investigated the pathways through which college attendance increases voting, focusing on three sets of causally ordered mediators: degree completion, family formation and stability, and socioeconomic status (SES). Using the regression-imputation approach, they estimated the corresponding PSEs.

[21] See Reardon and Raudenbush (2013) for discussion of multisite, multiple-mediator IV models.

5 Temporal and Spatial Interference

Many sociological questions involve the study of effects over time or interactions within networks. Indeed, historical or life cycle variation and network interactions lie at the center of sociological inquiry. But these settings complicate the definition and identification of causal effects. Just as sociologists studying temporal variation or network settings should consider causal processes, causal inference scholars should consider the complications involved in allowing treatments and effects to vary over time and interference between units under study. The stable unit treatment value assumption (SUTVA) posits that one unit's outcome is not affected by the treatment status of other units in the population. However, we are often faced with temporal or spatial interference that renders SUTVA untenable. In this section, we briefly review causal inference methods developed to study temporal and spatial interference.

5.1. Estimating Treatment Effects in the Presence of Temporal Interference

Temporal interference may arise in settings with time-varying treatments in which treatment status at a given time has not only contemporaneous effects, i.e., effects on outcomes measured immediately thereafter, but also carry-over effects, i.e., effects on outcomes at later time points. For example, exposure to family instability in early childhood may differ from exposure in adolescence, and exposure may have both short-term and long-term effects on a child's cognitive and socioemotional development (Lee & McLanahan 2015). A common strategy to incorporate temporal interference in causal analysis is through Robins' (1986, 1997) extension of the potential-outcomes framework to time-varying treatments. Consider a study with $T \geq 2$ time points where we are interested in the effect of a time-varying treatment W_t on an end-of-study outcome Y. Apart from a set of baseline or time-invariant confounders X, there is also a vector of observed time-varying confounders,

L_t, that may be affected by prior treatments. Note that L_t may also include current or past measures of the outcome (Brand & Xie 2007). See Figure 5 for a DAG representation of this setting when $T = 2$. In Lee and McLanahan's (2015) study of the relationship between family instability and child development, W_t denotes a family transition at time t, Y denotes a child's developmental outcome at time T, X includes a set of time-invariant covariates (e.g., mother's education), and L_t includes a set of time-varying covariates (e.g., poverty status). Following Robins et al. (2000), we use overbars to denote treatment histories such that $\overline{W} = (W_1, \ldots W_T)$ represents the observed treatment history until the end of the study and $Y_{\overline{w}}$ represents the potential outcome under a given treatment history $\overline{w}$. This notation allows us to consider various treatment effects based on contrasts between different potential outcomes. For instance, with two time points ($T = 2$), we could consider the distal treatment effect (DTE), defined as

$$\tau_{DTE} = E[Y_{1.0} - Y_{0.0}], \tag{21}$$

which captures the average effect of receiving treatment only at time 1 rather than never. Alternatively, we could consider the following treatment effects (Wodtke et al. 2020):

$$\tau_{PTE} = E[Y_{w_1,1} - Y_{w_1,0}], \tag{22}$$

$$\tau_{CTE} = E[Y_{1,1} - Y_{0,0}], \text{ and} \tag{23}$$

$$\tau_{INE} = E[(Y_{1,1} - Y_{1,0}) - (Y_{0,1} - Y_{0,0})]. \tag{24}$$

where τ_{PTE} is the proximal treatment effect (PTE), τ_{CTE} is the cumulative treatment effect (CTE), and τ_{INE} is the interaction effect (INE). Note that if there is no temporal interference, i.e., if the potential outcome Y_{w_1,w_2} depends only on treatment status at time 2, the DTE and the INE will both be zero, and the PTE will equal the CTE.

To identify the various causal contrasts considered above, it suffices to identify the expected potential outcome $E[Y_{\overline{w}}]$ for every treatment sequence $\overline{w}$. A key identification assumption for this quantity is sequential ignorability, which states that treatment at each

time point is unconfounded conditional on past treatments and observed confounders. Although it does not allow for unobserved confounding, the assumption of sequential ignitability allows for both carryover effects (past treatments affect current outcomes) and feedback effects (past outcomes affect current treatments). These are typically assumed away in fixed-effects models (Imai & Kim 2019).[22] Under sequential ignorability, we can estimate the expected potential outcome $E[Y_{\overline{w}}]$ via a range of parametric and semiparametric methods. A common method is inverse-probability-weighted (IPW) estimation of marginal structural models (MSMs) (Robins et al. 2000; see Wodtke et al. [2011] for a sociological application).[23] Apart from MSMs, time-varying treatment effects can also be assessed through structural nested mean models (SNMMs) and their associated estimators, such as the g-estimator (e.g., Naimi et al. 2017; Vansteelandt 2009; Vansteelandt & Sjolander 2016) and the RWR estimator (Wodtke 2020; Wodtke et al. 2020). These estimators involve modeling the conditional mean of the outcome as well as the conditional means and distributions of time-varying confounders. As with IPW, these methods are based on a set of strong modeling assumptions, which, when violated, can lead to biased estimates. To reduce model dependence, Bang and Robins (2005) propose a semiparametric estimator

[22] Elwert and Pfeffer (2019) incorporate future treatments as a proxy for an unmeasured confounder to address selection bias and discuss the conditions under which future values of the treatment can reduce or fully remove bias.

[23] The method of IPW involves modeling the conditional distribution of treatment at each time point t given past treatments and observed confounders. It is thus difficult to use when the treatment is continuous, in which case estimates of the conditional density functions tend to be unstable and highly sensitive to model misspecification. Moreover, even if models for these conditional distributions are correctly specified, IPW is often inefficient and susceptible to large finite sample biases. To overcome these limitations, Zhou and Wodtke (2020) propose an alternative method of constructing weights for MSMs called "residual balancing," which requires modeling the conditional means of the post-treatment confounders rather than the conditional distributions of treatment and is therefore easier to use with continuous treatments. It can be viewed as an extension of balancing weights (see Section 2) to longitudinal settings with temporal interference.

for the expected potential outcome $E[Y_{\overline{w}}]$. This estimator involves fitting $2 \cdot T$ models: a propensity score model at each time point and a model for an iteratively imputed outcome at each time point. This estimator is multiply robust in that it is consistent whenever the first k propensity score models and the last $K - k$ "outcome models" are correctly specified, where k can be any integer from 0 to K (Rotnitzky et al. 2017). The estimating equations are amenable to using DML.[24] Given its reduced dependence on model specification and complementarity with machine learning, we expect this semiparametric estimator (Bang & Robins 2005) and its variants (e.g., van der Laan & Rose 2018) to be more widely used in future research.[25]

5.2 Estimating Treatment Effects in the Presence of Spatial Interference

Spatial interference may arise in settings where units under consideration are not isolated but connected by a common physical or social space, such as schools, neighborhoods, and friendship networks, leading to "spillover effects." In such settings, one unit's potential outcome is a function of not only its own treatment status but also the treatment status of other related units (Aronow & Samii 2017; Athey et al. 2018; Tchetgen Tchetgen & VanderWeele 2012; VanderWeele 2015). Such interferences, or interactions, are prevalent in social settings (An 2018; An & VanderWeele 2022; Egami 2021). For example, encouraging an individual to vote by some intervention can increase the turnout for

---

[24] Specifically, if cross-fitting is used and the estimators of the propensity score and outcome models all converge to the truth at a faster-than-$n^{-1/4}$ rate, the Bang-Robins estimator will be root-n consistent, asymptotically normal, and semi-parametrically efficient.

[25] We may also be interested in situations in which both the treatment and an effect moderator vary over time. Wodtke and Almirall (2017) describe moderated intermediate causal effects and structural nested mean models for analyzing effect moderation in a longitudinal setting. Using this approach, they examine whether the effects of time-varying exposure to poor neighborhoods on the risk of adolescent childbearing are moderated by time-varying family income.

members of the household (Imai and Jiang 2020). In some cases, such interactions are the focus of analysis; in other cases, they are considered a nuisance to estimating treatment effects (given the assumption of no interference) (Hong & Raudenbush 2015; Ogburn et al. 2022). Yet ignoring interference can lead to biased estimates of causal effects and incorrect statistical inferences (An 2018; Basse & Airoldi 2018; Lee & Ogburn 2021).

When the pattern of interference is unconstrained, spillover effects are hard to study because (a) the number of counterfactuals for each unit increases exponentially as the number of units increases, leading to many causal contrasts that can be hard to estimate nonparametrically; and (b) the outcome of different units will be dependent, complicating statistical inference. Given these challenges, researchers often study spillover effects under two simplifying assumptions. First, the partial interference assumption posits that individuals are clustered in groups so that interference is limited to individuals within the same group (Sobel 2006). Second, the stratified inference assumption posits that within the same group, the effect of other units' treatment status on a focal unit's outcome operates through a known summary function g (e.g., the mean treatment status among other units in the same group) (Hudgens & Halloran 2008). The stratified inference assumption is quite strong, but it helps simplify the analysis, especially when there are more than a few units within each group. For example, when studying the spillover effect of grade retention on a child's test scores, Hong and Raudenbush (2015) invoke both assumptions by specifying a student's test score to be a function of their retention status and the retention rate of their peers in the same school.

Under the assumptions of partial interference and stratified interference, we can denote a unit's potential outcome as $Y_{w,g}$, where w denotes the unit's treatment status and g denotes a summary value of peer treatment status. The average individual effect can be defined as

$$E[Y_{1,g} - Y_{0,g}] \tag{25}$$

and the spillover effect can be defined as

$$E[Y_{w,g*} - Y_{w,g}], \tag{26}$$

where g^* denotes an alternative value of peer treatment status (VanderWeele 2015). As shown in Hudgens and Halloran (2008), these effects can be identified and unbiasedly estimated using an experimental design with a two-stage randomization procedure (i.e., first at the group level and then at the individual level within groups). To identify these effects in observational studies, one needs to invoke a group-level unconfoundedness assumption, i.e., conditional on a set of group-level covariates (which may include their individual-level components), the treatment assignments of all units within a group are independent of their potential outcomes (Tchetgen Tchetgen & VanderWeele 2012). Under this assumption, researchers can estimate the average individual and spillover effects through various strategies such as IPW, regression-imputation, and doubly robust methods (Liu et al. 2019). They can combine the doubly robust approach with DML to yield optimal performance (Park & Kang 2022).

In many social settings, people interact with each other through multiple channels and networks, such as friends, family, neighbors, and others. It is important to estimate the spillover effects that arise through each network; however, oftentimes those network interactions are unobserved, rendering unbiased estimation of spillover effects difficult. Egami (2021) develops sensitivity analysis methods for assessing the potential influence of unobserved networks on causal findings. Relatedly, An (2018) emphasizes the importance of collecting data on treatment diffusion to properly measure treatment interference, and then to estimate the direct treatment effect, treatment interference effect, and the treatment effect on interference.

6 Conclusion

Over the past three decades, causal inference has been an active research area in sociology and related disciplines such as economics, statistics, computer science, and political science. While earlier developments in causal analysis, in the form of path analysis and structural equations, were developed primarily in sociology (e.g., Duncan 1966) and then exported to other fields, much of what constitutes today's sociological methodology on causal inference has heavily borrowed knowledge from other disciplines.

Our review updates the latest advances in causal inference methodology. Given the large size of this literature, we chose to focus on four topics: causal effect identification and estimation in general, causal effect heterogeneity, causal effect mediation, and temporal and spatial interference. Our choice reflects long-standing sociological interests in these topics: population heterogeneity (e.g., Brand & Xie 2010; Xie 2013; Xie et al. 2012), causal mechanisms (e.g., Duncan 1966), and the importance of historical or life cycle variation and social context (e.g., Mason et al. 1983). As we reviewed, identifying and estimating causal components – a perennial objective in sociology – is no easy task with a counterfactual framework. There is no simple, one-size-fits-all solution. Causal inference with observational data, including quasi-experimental data, is a proposition specific to each research context. Often, researchers must invoke unverifiable assumptions and make consequential research decisions to draw causal conclusions. What makes sense for one research setting may not make sense for another. In applying new methods, we recommend that researchers thoroughly understand their underlying assumptions and tradeoffs to apply them judiciously. Researchers should also assess how effects vary across the population and whether the results of their study and sample generalize to a broader population. Sociological inquiry also invites the careful analysis of mechanism linking treatments to outcomes.

The past literature on causal inference has primarily been concerned with identification issues, while machine learning is often tasked with executing heavy computations with large data sets. The merge of the two strands of literature is facilitated by a long-recognized insight we discussed in the article: causal effects can be highly heterogeneous across different units. We review the latest developments in causal inference that utilize machine-learning methods to learn about heterogeneous treatment effects. We also describe how researchers can use machine learning to minimize biases in estimating population-level quantities of interest, including direct and indirect effects and effects with temporal and spatial interference.

Moving forward, we expect the continuation of fruitful cross-disciplinary fertilization in this area. We also anticipate increases in the use of machine-learning methods in causal inference to reduce estimation biases and detect causal effect heterogeneity. Machine-learning methods are particularly attractive and feasible considering future improvements in computational power as well as increasing availability of large administrative, commercial, and digital trace data (often called "big data") for social science research. However, we caution the reader that no computational method, machine-learning methods included, can solve what Holland (1986) called "the fundamental problem in causal inference," i.e., we never observe counterfactual outcomes. Good research design is primary. Computation is useful but only secondary. Hence, the bridge between machine-learning methods and causal inference can be productive only with innovative and appropriate research designs to address social-scientifically sensible research questions.

Acknowledgements

National Institutes of Health Grant R01 HD07460301A1 provided financial support for this research. The first author benefited from facilities and resources provided by the California Center for Population Research at UCLA (CCPR), which receives core support (P2C-HD041022) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). We thank affiliates of the Social Inequality Data Science Lab (https://www.sidatasciencelab.org), Ian Lundberg, Nathan Hoffmann, and an anonymous reviewer from *Annual Review of Sociology* for helpful comments and suggestions. The ideas expressed herein are those of the authors.

References

Abadie A, Cattaaneo MD. 2018. "Econometric methods for program evaluation." *Annual Review of Economics* 10:465-503.

Abadie A, Imbens GW. 2016. "Matching on the estimated propensity score." *Econometrica* 84: 781-807.

Acharya A, Blackwell M, Sen M. 2018. "Analyzing causal mechanisms in survey experiments." *Political Analysis* 26: 357-378.

Ahearn C, Brand JE, Zhou X. 2022. "How, and for whom, does higher education increase voting?" *Research in Higher Education* 1-24.

Alwin DF, Hauser RM. 1975. The decomposition of effects in path analysis. *American Sociological Review* 40: 37-47.

An W. 2010. Bayesian propensity score estimators: incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology* 40: 151-189.

An W. 2018. Causal inference with networked treatment diffusion. *Sociological Methodology* 48: 152-181.

An W, VanderWeele TJ. 2022. Opening the blackbox of treatment interference: tracing treatment diffusion through network analysis. *Sociological Methods & Research* 51: 141-164.

An W, Winship C. 2017. Causal inference in panel data with application to estimating race-of-interviewer effects in the general social survey. *Sociological Methods & Research* 46: 68-102.

Angrist JD, Imbens GW, Rubin DB. 1996. "Identification of causal effects using instrumental variables." *Journal of the American Statistical Association* 91: 444-455.

Angrist JD, Pischke J-S. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Angrist JD, Rokkanen M. 2015. Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association* 110:1331–1344.

Aronow PM, Samii C. 2016. Does regression produce representative estimates of causal effects? *American Journal of Political Science* 60: 250-267.

Aronow PM, Samii C. 2017. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics* 11:1912-47.

Athey S, Eckles D., Imbens GW. 2018. Exact P-values for network interference. *Journal of the American Statistical Association* 113:230–240.

Athey S, Imbens G. 2015. A measure of robustness to misspecification. *American Economic Review* 105: 476–80.

Athey S, Imbens G. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113:7353-7360.

Athey S, Imbens G. 2017. The state of applied econometrics: causality and policy evaluation. *Journal of Economic Perspectives* 31:3-32.

Athey S, Imbens G. 2019. Machine learning methods that economists should know about. *Annual Review of Economics* 11:685-725.

Athey S, Imbens G, Wager S. 2018. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80:597-623.

Athey S, Wager S. 2021. Policy learning with observational data. *Econometrics* 89(1); 133-161.

Athey S, Tibshirani J, Wager S. 2019. Generalized random forests. *The Annals of Statistics* 47(2):1148-1178.

Austin PC, Stuart EA. 2017. Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical methods in medical research* 26: 2505-2525.

Avin C, Shpitser I, Pearl J. 2005. Identifiability of path-specific effects. *UCLA: Department of Statistics, UCLA*. Retrieved from https://escholarship.org/uc/item/45x689gq

Baron RM, Kenny DA. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* 51: 1173-1182.

Bang H, Robins JM. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61: 962-973.

Basse GW, Airoldi EM. 2018. Limitations of design-based causal inference and A/B testing under arbitrary and network interference. *Sociological Methodology* 48: 136-151.

Belloni A, Chernozhukov V, Hansen C. 2014. High-dimensional methods and inference on structural and treatment effects. *J. Econ. Perspect.* 28:29–50

Bertanha M,Imbens GW. 2020. External validity in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics* 38:593-612.

Blandhol C, Bonney J, Mogstad M, Torgovitsky A. 2022. *When is TSLS actually LATE?*. No. w29709. National Bureau of Economic Research.

Bloome D, Schrage D. 2021. Covariance regression models for studying treatment effect heterogeneity across one or more outcomes: understanding how treatments shape inequality. *Sociological Methods & Research* 50: 1034-1072.

Bollen KA. 2014. *Structural Equations with Latent Variables*. John Wiley & Sons.

Bound J, Jaeger DA, and Baker RM. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90: 443-450.

Brand JE. 2022. *Overcoming the Odds: The Far-Reaching Benefits for Unlikely College Graduates*. Unpublished book manuscript.

Brand, JE, Xu J, Koch B. 2020. Machine learning. In *Research Methods in the Social Sciences Foundation*. ed. P Atkinson, et al. Thousand Oaks: Sage.

Brand JE, Moore R, Song X, Xie Y. 2019a. Why does parental divorce lower children's educational attainment? A causal mediation analysis. *Sociological Science* 6:264-292.

Brand JE, Moore R, Song X, Xie Y. 2019b. Parental divorce is not uniformly disruptive to children's educational attainment. *Proceedings of the National Academy of Sciences* 116:7266-7271.

Brand JE, Simon Thomas J. 2013. Causal effect heterogeneity. Pp. 189-214 in *Handbook of Causal Analysis for Social Research*, Stephen L. Morgan ed., Springer Series.

Brand JE, Simon Thomas J. 2014. Job displacement among single mothers: effects on children's outcomes in young adulthood. *American Journal of Sociology* 119:955-1001.

Brand JE, Xie Y. 2007. Identification and estimation of causal effects with time-varying treatments and time-varying outcomes. *Sociological Methodology* 37:393-434.

Brand JE, Xie Y. 2010. Who benefits most from college? evidence for negative selection in heterogeneous economic returns to higher education. *American Sociological Review* 75: 273-302.

Brand JE, Xu J, Koch B, Geraldo P. 2021. Uncovering sociological effect heterogeneity using machine-learning. *Sociological Methodology* 51:189-223.

Breiman L. 2001. Random forests. *Machine Learning* 45:5–32.

Caliendo M, Kopeinig S. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* 22: 31-72.

Card D. 2001. Estimating the return to schooling: progress on some persistent econometric problems. *Econometrica* 69:1127-60.

Carranza AGael, Krishnamurth SK, Athey S. 2022. Flexible and efficient contextual bandits with heterogeneous treatment effect oracle.arXiv:2203.16668v1.

Cattaneo MD, Idrobo N, Titiunik R. 2019. *A Practical Introduction to Regression Discontinuity Designs: Foundations.* Cambridge University Press.

Cattaneo MD Titiunik R. 2022. Regression discontinuity designs. *Annual Review of Economics* 14: 821-51.

Cheng S, Brand JE, Zhou X, Xie Y, Hout M. 2021. "Heterogeneous returns to college over the life course." *Science Advances* 7: eabg7641.

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, C1–C68.

Cinelli C, Forney A, Pearl J. 2022. A crash course in good and bad controls. *Sociological Methods and Research* 1-34.

Cinelli C, Hazlett C. 2020. Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society, Series B* 82: 39-67.

Cole SR, Stuart EA. 2010. Generalizing evidence from randomized clinical trials to target populations: The ACTG-320 trial. *American Journal of Epidemiology* 172, 107–115.

Daniel RM, De Stavola BL, Cousens SN, Vansteelandt S. 2015. Causal mediation analysis with multiple mediators. *Biometrics*, *71*, pp.1-14.

Deaton A. 2010. Instruments, randomization, and learning about development. *Journal of Economic Literature* 48:424–55.

Deaton A, Cartwright N. 2018. "Understanding and misunderstanding randomized controlled trials." *Social Science & Medicine* 210, 2– 21.

Díaz I, Hejazi NS, Rudolph KE, van Der Laan MJ. 2021. "Nonparametric efficient causal mediation with intermediate confounders." *Biometrika* 108: 627-641.

Didelez V, Dawid AP, Geneletti S. 2006. Direct and indirect effects of sequential treatments. In *23rd Annual Conference on Uncertainty in Artifical Intelligence*.

Dong Y, Lewbel A. 2015. Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics* 97:1081–1092.

Duncan OD. 1966. Path analysis: sociological examples. *The American Journal of Sociology*, 72, 1-16.

Egami N. 2021. Spillover effects in the presence of unobserved networks. *Political Analysis* 29(3): 287-316.

Egami N, and Hartman E. 2020. "Elements of external validity: framework, design, and analysis." *SSRN*. Available at SSRN: https://ssrn.com/abstract=3775158 or http://dx.doi.org/10.2139/ssrn.3775158

Elwert F. 2015. Graphical causal models. In *Handbook of Causal Analysis for Social Research*, ed. SL Morgan, pp. 245-273. Springer.

Elwert F, Pfeffer FT. 2019. The future strikes back: using future treatments to detect and reduce hidden bias. *Sociological Methods and Research* 51: 1014-1051.

Elwert F, Winship C. 2014. Endogenous selection bias: the problem of conditioning on a collider variable. *Annual Review of Sociology* 40: 31-53.

Felton C, Stewart B. 2022. Handle with care: a sociologist's guide to causal inference with instrumental variables. Open Science Framework.

Findley MG,Kikuta K, Denley M. 2021. External validity. *Annual Review of Political Science* 24: 365-393.

Fisher RA. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Fong C, Hazlett C, Imai K. 2018. Covariate balancing propensity score for a continuous treatment: application to the efficacy of political advertisements. *The Annals of Applied Statistics* 12: 156-177.

Gangl M. 2010. Causal inference in sociological research. *Annual Review of Sociology* 36:21-47.

Gangl M. 2015. Partial identification and sensitivity analysis. In *Handbook of Causal Analysis for Social Research, ed. SL Morgan,* pp. 377-402. Springer.

Geneletti S. 2007. Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society: Series B (Statistical Methodology* 69:199-215.

Gill RD, Robins JM. 2001. Causal Inference for complex longitudinal data: the continuous case. *The Annals of Statistics* 29: 1785–1811.

Grimmer J, Roberts ME, Stewart BE. 2021. Machine learning for social science: an agnostic approach. *Annual Review of Political Science* 24:395–419.

Hahn PR, Murray J, Carvalho C. 2020. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis* 15: 965-1056.

Hahn J, Todd P, Van der Klaauw W. 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69: 201-209.

Hainmueller J. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 20: 25-46.

Hartman E, Grieve R, Ramsahai R, Sekhon JS. 2015. From sample average treatment effect to population average treatment effect on the treated: Combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society, Series A* 178:757-778.

Hastie T, Tibshirani R, Friedman JH. 2017. *The Elements of Statistical Learning, 2nd Edition.* Berlin: Springer.

Hazlett C. 2020. Kernel Balancing: A flexible non-parametric weighting procedure for estimating causal effects. *Statistica Sinica* 30: 1155-1189.

Heckman JJ, Humphries JE, Veramendi G. 2018. The nonmarket benefits of education and ability. *Journal of Human Capital* 12:282-304.

Heckman JJ, Robb R. 1986. Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. *Drawing Inferences from Self-Selected Sample* 63-107.

Heckman JJ, Urzua S, Vytlacil E. 2006. Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics* 88:389-432.

Holland PW. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81:945–60.

Hong G, Raudenbush S. 2015. Heterogeneous agents, social interactions, and causal inference. In *Handbook of Causal Analysis for Social Research*, ed. SL Morgan, pp. 331-352. Springer.

Hotz JV, Imbens GW, Mortimer JH. 2005. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics* 125, 241–270.

Huber M. 2021. Causal analysis: impact evaluation and causal machine learning with applications in R. https://drive.switch.ch/index.php/s/tNhKQmkGB48bjfz

Hudgens MG, Halloran ME. 2008. Toward causal inference with interference. *Journal of the American Statistical Association* 103: 832-842.

Imai K, Jiang Z. 2020. Identification and sensitivity analysis of contagion effects in randomized placebo-controlled trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 183: 1637-1657.

Imai K, Keele L, Yamamoto T. 2010. Identification, inference and sensitivity analysis for causal mediation effects. Statistical science, 25: 51-71.

Imai K, Ratkovic M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7: 443-470.

Imai K, Ratkovic M. 2014. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76: 243-263.

Imai K, Yamamoto T. 2013. Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments. *Political Analysis* 21:141–171.

Imai K, Kim IS. 2019. When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science* 63: 467-490.

Imbens GW. 2004. "Nonparametric estimation of average treatment effects under exogeneity: a review." *The Review of Economics and Statistics* 86: 4–29.

Imbens GW. 2015. Matching methods in practice. *Journal of Human Resources* 50; 373-419.

Imbens GW, Angrist JD. 1994. Identification and estimation of local average treatment effects. *Econometrica*, 62: 467-475.

Imbens GW, Lemieux T. 2008. Regression discontinuity designs: a guide to practice. *Journal of Econometrics* 142: 615-635.

Imbens GW, Rubin D. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

Josey KP, Yang F, Ghosh D, Raghavan S. 2022. A calibration approach to transportability and data-fusion with observational data. *Statistics in Medicine* 41: 4511-4531.

Kennedy EH, Ma Z, McHugh MD, Small DS. 2017. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 79: 1229-1245.

Kern HL, Stuart EA, Hill J, Green DP. 2016. Assessing methods for generalizing experimental impact estimates to target populations. *Journal of research on educational effectiveness*, 9: 103-127.

Klein M, Kühhirt M. 2021. Direct and indirect effects of grandparent education on grandchildren's cognitive development: the role of parental cognitive ability. *Sociological Science* 8: 265-284.

Künzel SR, Sekhon JS, Bickel PJ, Yu B. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* 116: 4156-4165.

Lee DS, Lemieux T. 2010. Regression discontinuity designs in economics. *Journal of economic literature* 48: 281-355.

Lee D, McLanahan S. 2015. Family structure transitions and child development: Instability, selection, and population heterogeneity. *American Sociological Review* 80: 738-763.

Lee BK, Lessler J, Stuart E. 2010. Improving propensity score weighting using machine learning. *Statistics in Medicine* 29:337-46

Lee Y, Ogburn EL. 2021. Network dependence can lead to spurious associations and invalid inference. *Journal of the American Statistical Association* 116: 1060-1074.

Levy BL, Owens A, Sampson RJ. 2019. The varying effects of neighborhood disadvantage on college graduation: Moderating and mediating mechanisms. *Sociology of Education* 92: 269-292.

Liu L, Hudgens MG, Saul B, Clemens JD, Ali M, Emch ME. 2019. Doubly robust estimation in observational studies with partial interference. *Stat* 8: e214.

Lundberg I. 2022. The gap-closing estimand: a causal approach to study interventions that close disparities across social categories. *Sociological Methods and Research*.

Lundberg I, Brand JE, Jeon N. 2022. Researcher reasoning meets computational capacity: machine learning for social science. *Social Science Research*. 102807.

Lundberg I, Johnson R, Stewart BM. 2021. What is your estimand? defining the target quantity connects statistical evidence to theory. *American Sociological Review*. *86*, 532-565.

Manski CF. 1995. *Identification Problems in the Social Sciences*. Boston, MA: Harvard University Press.

Manski CF. Garfinkel I. 1992. Introduction. In *Evaluating Welfare and Training Programs*, ed. CF Manski, I Garfinkel. Cambridge, pp.1-21. MA: Harvard University Press.

Mason WM, Wong GY, Entwisle B. 1983. Contextual analysis through the multilevel linear model. *Sociological methodology* 14: 72-103.

McCaffrey DF, Ridgeway G, Morral AR. 2004. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* 9: 403.

Miles CH, Shpitser I, Kanki P, Meloni S, Tchetgen Tchetgen EJ. 2017. Quantifying an adherence path-specific effect of antiretroviral therapy in the Nigeria PEPFAR program. *Journal of the American Statistical Association*, *112*, 1443-1452.

Miles CH, Shpitser I, Kanki P, Meloni S, Tchetgen Tchetgen, EJ. 2020. On semiparametric estimation of a path-specific effect in the presence of mediator-outcome confounding. *Biometrika*, *107*, 159-172.

Mogstad M, Torgovitsky A. 2018. Identification and extrapolation of causal effects with instrumental variables. *Annual Review of Economics* 10: 577-613.

Molina M, Garip F. 2019. Machine learning for sociology. *Annual Review of Sociology* 45: 27-45.

Morgan S, Harding D. 2006. Matching estimators of causal effects: prospects and pitfalls in theory and practice. *Sociological Methods and Research* 35:3-60.

Morgan, S, Winship C. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge, UK: Cambridge University Press.

Naimi AI, Cole SR, Kennedy EH. 2017. An introduction to g methods. *International Journal of Epidemiology* 46: 756-762.

Neyman J. 1923. On the Application of Probability Theory to Agricultural Experiments. *Statistical Science* 5: 465-480.

Nie X, Wager S. 2021. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108: 299-319.

Nguyen TQ, Schmid I, Stuart EA. 2021. Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods* 26: 255.

Offer-Westort M, Coppock A, Green DP. 2021. Adaptive experimental design: prospects and applications in political science. *American Journal of Political Science* 65: 826-844.

Ogburn EL, Sofrygin O, Diaz I, Van Der Laan MJ. 2022. Causal inference for social network data. *Journal of the American Statistical Association* 1-46.

Park C, Kang H. 2022. Efficient semiparametric estimation of network treatment effects under partial interference. *Biometrika*.

Pearl J. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420. Morgan Kaufmann Publishers Inc.

Pearl J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pearl J, Bareinboim E. 2014. External validity: from do-calculus to transportability across populations. *Statistical Science* 29, 579– 595.

Quandt R. 1972. A new approach to estimating switching regression. *Journal of the American Statistical Association* 67:306-10.

Reardon SF, Raudenbush SW. 2013. Under what assumptions do site-by-treatment instruments identify average causal effects? *Sociological Methods & Research* 42: 143-163.

Robins JM. 1986. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling* 7:1393–1512.

Robins JM. 1997. Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, ed. M Berkane, pp. 69–117. New York: Springer.

Robins JM. 2003. Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems*, ed. PJ Green, NL Hjort, S Richardson, pp. 70–81. New York: Oxford University Press.

Robins JM, Greenland S. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 3: 143-155.

Robins JM, Rotnitzky A. 1995. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 90: 122-129.

Robins JM, Rotnitzky A, Zhao LP. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89:846–66.

Robins, JM., Migule Angel Hernan, and Babette Brumback. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* 11: 55—560.

Robins JM, Hernan MA, Brumback B. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*, 550-560.

Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.

Robins JM, Richardson TS, Shpitser I. 2022. An interventionist approach to mediation analysis."In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 713-764.

Rotnitzky A, Robins JM, Babino L. 2017. On the multiply robust estimation of the mean of the g-functional. *arXiv preprint:1705.08582*.

Roy AD. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Paper* 3:135-46.

Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66: 688–701.

Rubin DB. 1977. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* 2: 1-26.

Rubin DB. 1986. Which ifs have causal answers? Discussion of "Statistics and causal Inference" by Holland. *Journal of the American Statistical Association* 83:396.

Scharfstein DO, Rotnitzky A, Robins JM. 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 94: 1096-1120.

Scott SL. 2010. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry* 26: 639-658.

Semenova V, Chernozhukov V.2021. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal* 24: 264-289.

Shu X, Ye Y. 2022. knowledge discovery: methods from data mining and machine learning. *Social Science Research* 102817.

Sobel ME. 2006. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association* 101: 1398-1407.

Steiner PM, Kim Y, Hall CE, Su D. 2017. Graphical models for quasi-experimental designs. *Sociological methods & research* 46: 155-188.

Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. 2011. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174: 369-386.

Stuart EA, Bradshaw CP, Leaf PJ. 2015. Assessing the generalizability of randomized trial results to target populations. *Prevention Science* 16: 475-485.

Tchetgen Tchetgen EJ, Shpitser I. 2012. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics* 40: 1816-1845.

Tchetgen Tchetgen EJ, VanderWeele TJ. 2012. On causal inference in the presence of interference. *Statistical Methods in Medical Research* 21: 55-75.

Tipton E. 2013. Improving generalizations from experiments using propensity score subclassification: assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics* 38:239-266.

Tipton E, Hedges L, Vaden-Kiernan M, Borman G, Sullivan K, Caverly S. 2014. Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, *7*, 114-135.

van der Laan MJ, Rubin D. 2006. Targeted maximum likelihood learning. *The international journal of biostatistics*, *2*: Article 11.

van der Laan MJ, Rose S. 2018. *Targeted learning in data science*. Cham: Springer International Publishing.

VanderWeele TJ. 2009. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* 20: 18-26.

VanderWeele TJ. 2010. Bias formulas for sensitivity analysis for direct and indirect effects NIH Public Access. Epidemiology 21: 540–51.

VanderWeele TJ. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York, NY: Oxford University Press.

VanderWeele TJ. 2016. Mediation analysis: a practitioner's guide. *Annual Review of Public Health* 37:2.1–16.

VanderWeele TJ, Arah OA. 2011. unmeasured confounding for general outcomes, treatments, and confounders: bias formulas for sensitivity analysis. *Epidemiology* 22: 42–52.

VanderWeele, TJ, Vansteelandt S. 2009. Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface* 2: 457-468.

VanderWeele TJ, Vansteelandt S, Robins JM. 2014. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology* 25:300–306.

Vansteelandt S. 2009. Estimating direct effects in cohort and case–control studies. *Epidemiology* 20: 851-860.

Vansteelandt S, Bekaert M, Lange T. 2012. Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods* 1:131–158.

Vansteelandt S, Sjolander A. 2016. Revisiting g-estimation of the effect of a time-varying exposure subject to time-varying confounding. *Epidemiologic Methods* 5: 37-56.

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

Westreich D, Edwards JK, Lesko CR, Cole SR, Stuart EA. 2019. Target validity and the hierarchy of study designs. *American journal of epidemiology* 188: 438-443.

Westreich D, Lessler J, Funk MJ. 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology* 63: 826-833.

Winship C, Morgan SL. 1999. The estimation of causal effects from observational data. *Annual Review of Sociology* 25: 659-707.

Wodtke GT. 2020. Regression-based adjustment for time-varying confounders. *Sociological Methods and Research* 49:906-946.

Wodtke GT, Alaca Z, Zhou X. 2020. Regression-with-residuals estimation of marginal effects: a method of adjusting for treatment-induced confounders that may also be effect modifiers. *Journal of the Royal Statistical Society: Series A* 183: 311-332.

Wodtke, GT, Almirall D. 2017. Estimating moderated causal effects with time-varying treatments and time-varying moderators: Structural nested mean models and regression with residuals. *Sociological Methodology* 47: 212-245.

Wodtke, GT, Harding DJ, Elwert F. 2011. Neighborhood effects in temporal perspective: the impact of long-term exposure to concentrated disadvantage on high school graduation. *American Sociological Review* 76:713-736.

Wodtke GT, Parbst M. 2017. Neighborhoods, schools, and academic achievement: a formal mediation analysis of contextual effects on reading and mathematics abilities. *Demography* 54 : 1653–1676.

Wodtke GT, Ramaj S, Schachner J. 2022. Toxic neighborhoods: the effects of concentrated poverty and environmental lead contamination on early childhood development. Demography 59: 1275–1298.

Wodtke, GT, Yildirim U, Harding DJ, Elwert F. 2020. Are neighborhood effects explained by differences in school quality?

Wodtke GT, Zhou X. 2020. Effect decomposition in the presence of treatment-induced confounding: a regression-with-residuals approach. *Epidemiology* 31:369-375.

Xia F, Chan KCG. 2021. "Identification, semiparametric efficiency, and quadruply robust estimation in mediation analysis with treatment-induced confounding." *Journal of the American Statistical* Association,1-10

Xie Y. 2013. Population heterogeneity and causal inference. *Proceedings of the National Academy of Sciences* 110: 6262-6268.

Xie Y, Brand JE, Jann B. 2012. Estimating heterogeneous treatment effects with observational data. *Sociological Methodology* 42:314-347.

Xie Y, Near C, Xu H, Song X. 2020. Heterogeneous treatment effects on children's cognitive/non-cognitive skills: a reevaluation of an influential early childhood intervention. *Social Science Research* 86: 102389.

Yadlowsky S, Fleming S, Shah N, Brunskill E, Wager S. 2021. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966*.

Zhou, X. 2019. Equalization or selection? reassessing the "meritocratic power" of a college degree in intergenerational income mobility. *American Sociological Review* 84:459–485.

Zhou, X. 2022a. Attendance, completion, and heterogeneous returns to college: a causal mediation approach. *Sociological Methods & Research.*

Zhou, X. 2022b. Semiparametric estimation for causal mediation analysis with multiple causally ordered mediators". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 84: 794-821.

Zhou X, Pan G. Forthcoming. Higher education and the black-white earnings gap. *American Sociological Review.*

Zhou X, Wodtke GT. 2019. A regression-with-residuals method for estimating controlled direct effects. *Political Analysis* 27:360-369.

Zhou X, Wodtke GT. 2020. Residual balancing: a method of constructing weights for marginal structural models. *Political Analysis* 28:487-506.

Zhou X, Xie Y. 2019. Marginal treatment effects from a propensity score perspective. *Journal of Political Economy* 127: 3070-3084.

Zhou X, Xie Y. 2020. Heterogeneous treatment effects in the presence of self-selection: a propensity score perspective. *Sociological Methodology* 50: 350-385.

Zhou X, Yamamoto T. 2022. Tracing causal paths from experimental and observational data. *Journal of Politics.*

Zubizarreta JR. 2015. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association* 110: 910-922.

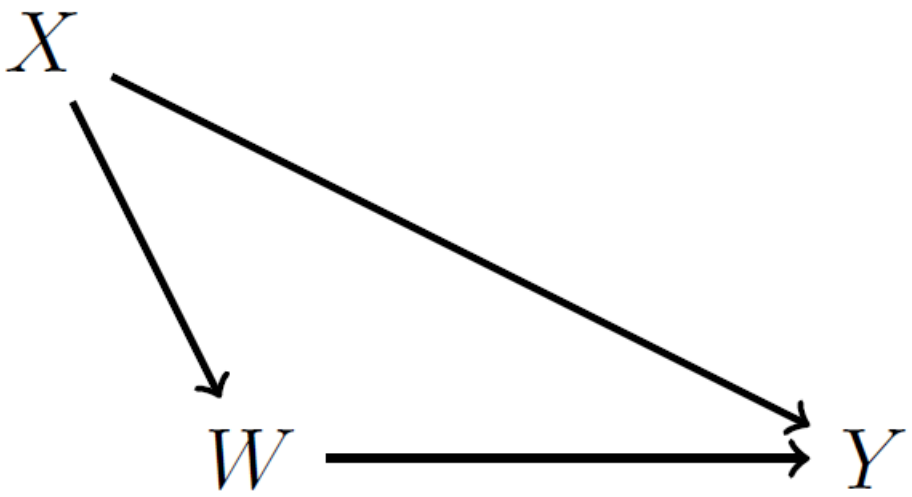

Figure 1: Direct Acyclic Graph under Unconfoundedness

Note: W denotes treatment status, Y denotes the outcome of interest, and X denotes observed pretreatment confounders.

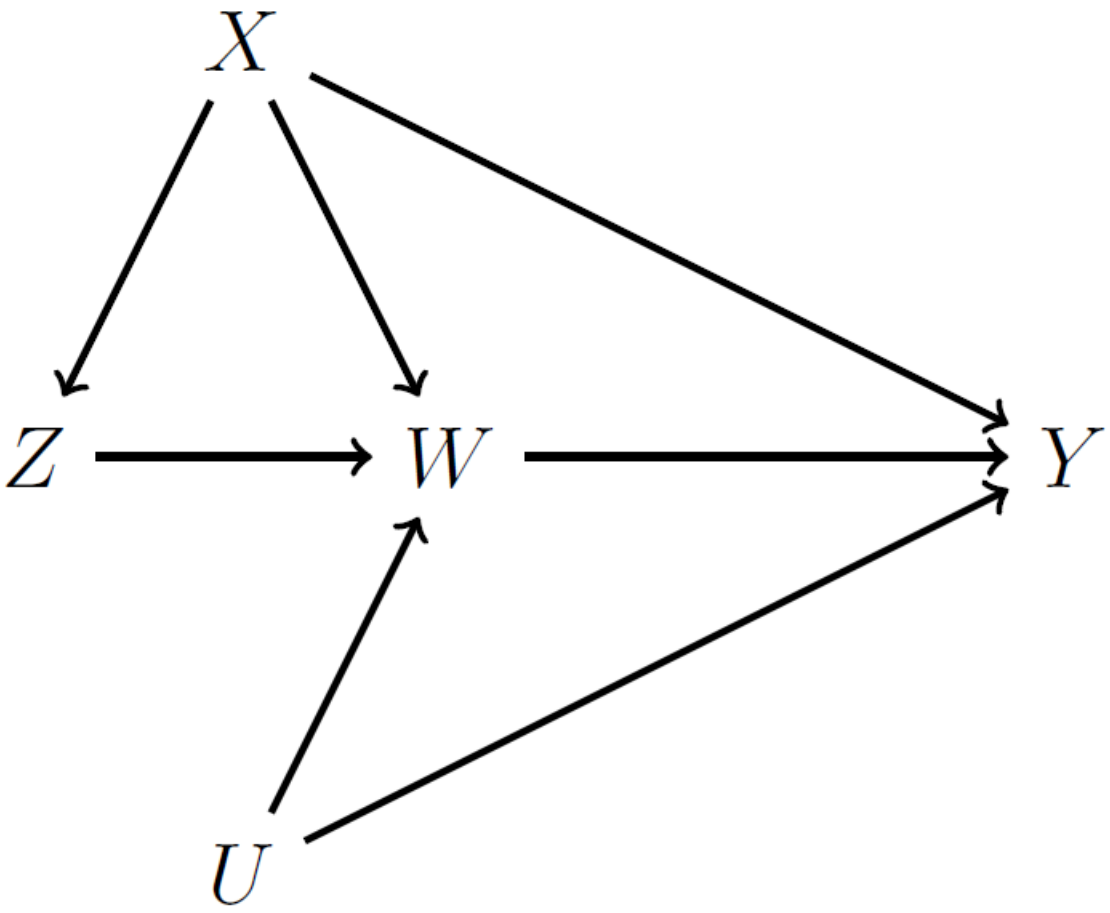

Figure 2: Direct Acyclic Graph under the Instrumental Variable (IV) Design

Note: W denotes treatment status, Y denotes the outcome of interest, Z denotes an instrumental variable, and X denotes observed pretreatment confounders.

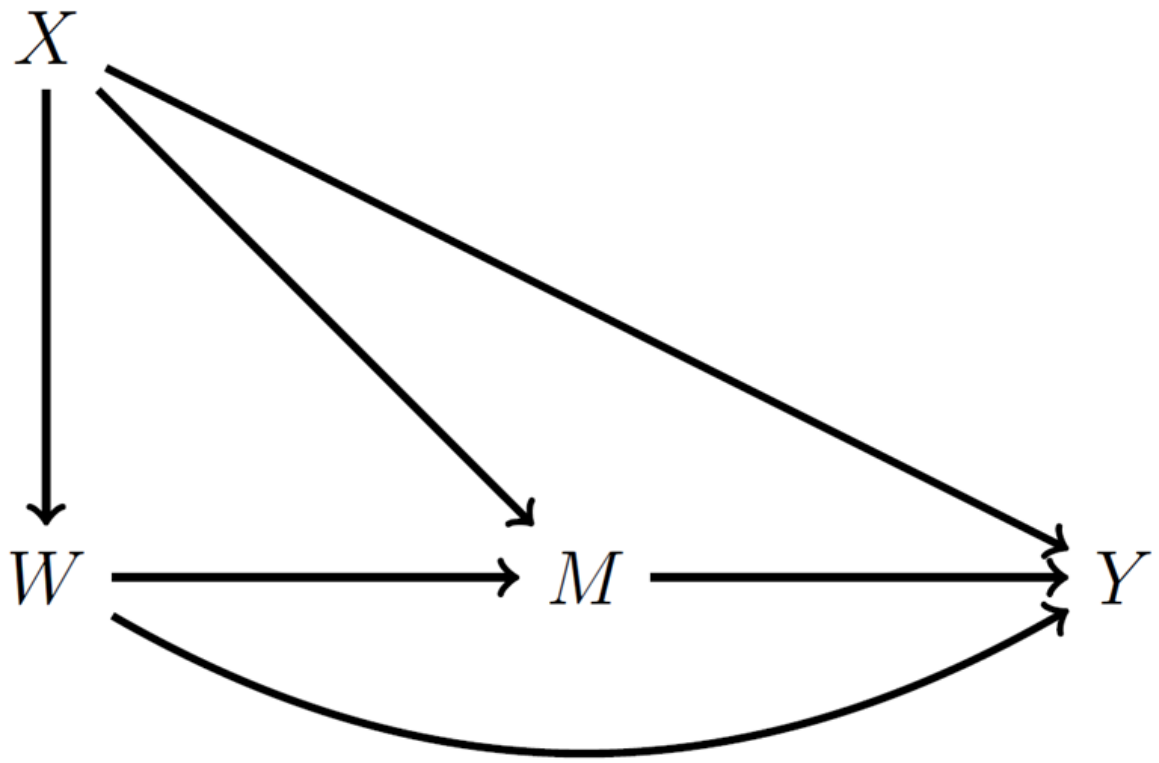

Figure 3: Causal Mediation Analysis without Treatment-Induced Confounding

Note: W denotes treatment status, Y denotes the outcome of interest, M denotes a putative mediator, and X denotes observed pretreatment confounders.

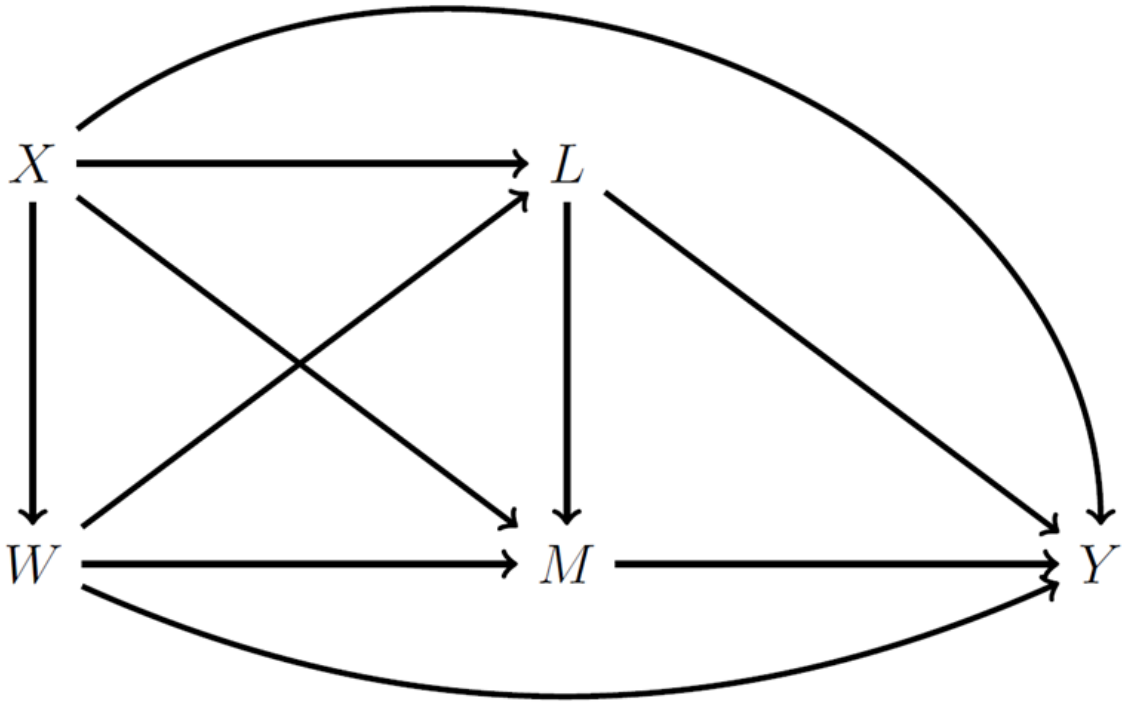

Figure 4: Causal Mediation Analysis with Treatment-Induced Confounding

Note: W denotes treatment status, Y denotes the outcome of interest, M denotes a putative mediator, X denotes observed pretreatment confounders, and L denotes treatment-induced confounders.

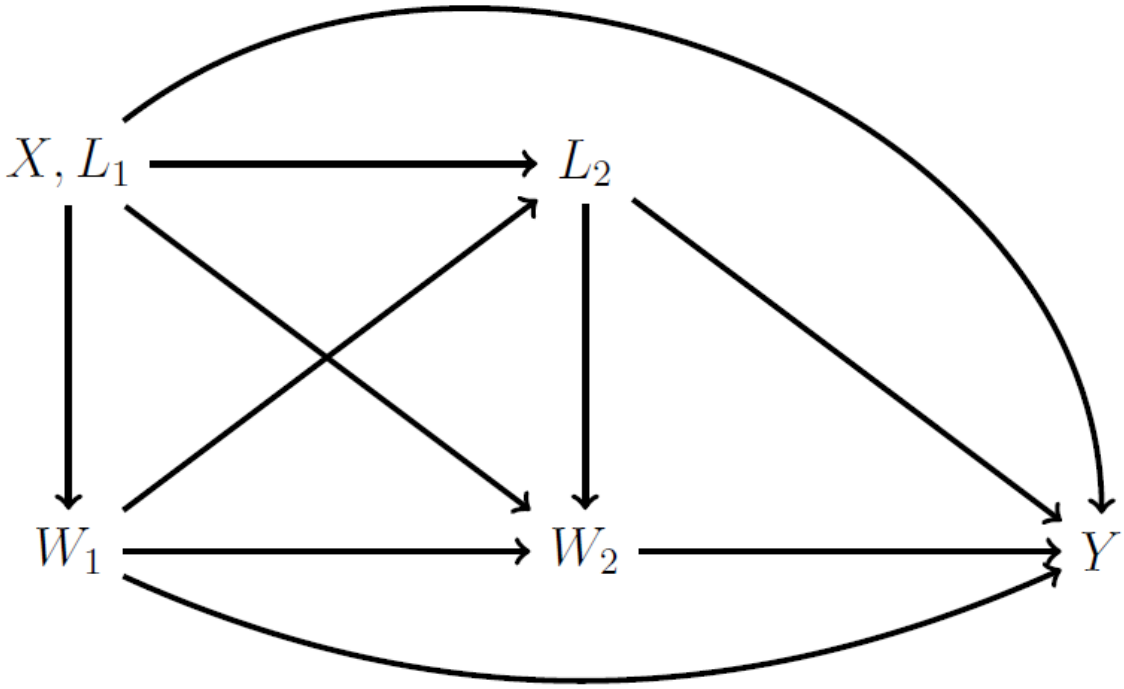

Figure 5: Causal Inference with Temporal Interference

Note: W_t denotes treatment status at time t, Y denotes the outcome of interest, X denotes baseline confounders, and L_t denotes time-varying confounders at time t.