



UNIVERSITÀ
DEGLI STUDI
FIRENZE

FLORE

Repository istituzionale dell'Università degli Studi di Firenze

From Controlled to Undisciplined Data: Estimating Causal Effects in the Era of Data Science Using a Potential Outcome Framework

Questa è la Versione finale referata (Post print/Accepted manuscript) della seguente pubblicazione:

Original Citation:

From Controlled to Undisciplined Data: Estimating Causal Effects in the Era of Data Science Using a Potential Outcome Framework / Dominici, Francesca; Bargagli-Stoffi, Falco J.; Mealli, Fabrizia. - In: HARVARD DATA SCIENCE REVIEW. - ISSN 2644-2353. - ELETTRONICO. - 3:(2021), pp. 0-0. [10.1162/99608f92.8102afed]

Availability:

This version is available at: 2158/1262520 since: 2022-03-26T14:37:10Z

Published version:

DOI: 10.1162/99608f92.8102afed

Terms of use:

Open Access

La pubblicazione è resa disponibile sotto le norme e i termini della licenza di deposito, secondo quanto stabilito dalla Policy per l'accesso aperto dell'Università degli Studi di Firenze (<https://www.sba.unifi.it/upload/policy-oa-2016-1.pdf>)

Publisher copyright claim:

(Article begins on next page)

Harvard Data Science Review • Issue 3.3, Summer 2021

From Controlled to Undisciplined Data: Estimating Causal Effects in the Era of Data Science Using a Potential Outcome Framework

Francesca Dominici¹, Falco J. Bargagli-Stoffi², Fabrizia Mealli³

¹Harvard Data Science Initiative, Harvard University, Cambridge, Massachusetts, United States of America; Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Cambridge, Massachusetts, United States of America,

²Department of Biostatistics, Harvard T.H. Chan School of Public Health, Cambridge, Massachusetts, United States of America,

³Department of Statistics, Informatics and Applications, University of Florence, Florence, Italy

Published on: Aug 31, 2021

DOI: 10.1162/99608f92.8102afed

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

This article discusses the fundamental principles of causal inference—the area of statistics that estimates the effect of specific occurrences, treatments, interventions, and exposures on a given outcome from experimental and observational data. We explain the key assumptions required to identify causal effects, and highlight the challenges associated with the use of observational data. We emphasize that experimental thinking is crucial in causal inference. The quality of the data (not necessarily the quantity), the study design, the degree to which the assumptions are met, and the rigor of the statistical analysis allow us to credibly infer causal effects. Although we advocate leveraging the use of big data and the application of machine learning (ML) algorithms for estimating causal effects, they are not a substitute for thoughtful study design. Concepts are illustrated via examples.

Keywords: causal inference, data science, potential outcomes, randomized control trials, observational studies, Bayesian modeling, machine learning

1. Introduction

Questions about cause and effect are ubiquitous in our everyday lives:

- Starting now, if I stop eating ice cream at night, how much weight will I lose in 3 months?
- If I adopt a puppy, will the severity of my depression symptoms improve in one year?
- If I give my patient new chemotherapy instead of the standard chemotherapy, how many more months will he/she live?
- If my government implements stricter regulatory policies for air pollution, how much longer can I expect to live?
- If my country or state had implemented a mask mandate 3 months ago to slow down the spread of COVID-19, how many lives would have been saved?

These are just a few situations ranging in complexity and importance where we would like to estimate the causal effect of a defined intervention W (e.g., not eating ice cream, adopting a puppy, taking a new drug, implementing air pollution regulations, enacting a stricter social distancing measure) on a specific outcome Y (e.g., body weight, depression symptoms, life expectancy, COVID-19 deaths).

This article discusses the fundamental ideas of causal inference under a potential outcome framework ([Neyman, 1923](#)); [D. B. Rubin \(1974, D. B. Rubin, 1978\)](#)) in relation

to new data science developments. As statisticians, we focus on study design and estimation of causal effects of a specified, well-defined intervention W on an outcome Y from observational data.

The article is divided into eight sections:

- Sections 1 and 2 – which heuristically contrast randomized controlled experiments with observational studies.
- Section 3 – the design phase of a study, including the illustration of the key assumptions required to define and identify a causal effect.
- Section 4 – a nontechnical overview of common approaches for estimating a causal effect, focusing on Bayesian methods.
- Section 5 – advantages and disadvantages of the most recent approaches for machine learning (ML) in causal inference.
- Section 6 – recent methods for estimating heterogeneous causal effects.
- Section 7 – a discussion of the critical role of sensitivity analysis to enhance the credibility of causal conclusions from observational data.
- Section 8 – a concluding discussion outlining future research directions.

1.1. The Potential Outcomes framework is just one of the many popular approaches to causal inference:

Before we present our panoramic view of the potential outcome framework for causal inference, we want to acknowledge that *causation* and *causality* are scientific areas that span many disciplines, including but not limited to statistics. Several approaches to causality have emerged and have become popular in the areas of computer, biomedical, and social sciences.

Although the focus of this article is on the potential outcome framework, we want to stress the importance of acknowledging the other approaches and viewpoints that exist in the general area of causality. A Google search on causality resulted in 45 books (<https://www.goodreads.com/shelf/show/causality>) of which fewer than 10 focus on statistical estimation of causal effects. Many of those books provide philosophical views of causal reasoning in the context of different disciplines and a comprehensive overview of causality as a discipline. Texts that discuss causal methods from the potential outcome perspective (even if they are not always exclusive) include ([Angrist & Pischke, 2008](#), ([M. D. Hernán & Robins, 2020](#), ([G. W. Imbens & Rubin, 2015](#)), ([Morgan & Winship, \(2007, 2014\)](#)), and ([2002](#)), ([2010](#))). The [Pearl \(2000\)](#) and [Pearl and Mackenzie \(2018\)](#) approach to causality has become very popular in computer science

(see for example [Peters et al., 2017](#)). In particular, the books by Pearl introduce the notion of causal graphs, which are graphical representations of causal models used to represent assumptions about potential causal relationships. These graphs represent assumptions regarding causal relationships as edges between nodes. In these instances, graphs are used to determine if data can identify causal effects and visually represent assumptions in causal models. The benefits of this approach are particularly evident when multiple variables could have causal relationships in a complex fashion (see also [\(M. D. Hernán & Robins, 2020\)](#)).

Other books cover the causality-philosophical meaningfulness of causation ([Beebe et al., 2009](#); [Cartwright, 2007](#); [Halpern, 2016](#)); deducing the causes of a given effect ([Dawid et al., 2017](#)); understanding the details of a causal mechanism ([T. VanderWeele, 2015](#)); or discovering or unveiling the causal structure ([Glymour et al. \(2014\)](#); [Spirtes et al., 2001](#)). Unfortunately, it is impossible to summarize all of these contributions in a single article.

For the type of studies we have encountered in our work in sociology, political science, economics, and environmental and health sciences, the typical setting is to estimate the causal effect of a pre-specified treatment or intervention W on an outcome Y . In these instances we have found the potential outcome approach useful to draw causal inferences. The potential outcome framework is also helpful to bridge experimental and observational thinking.

In this article, we provide a statistical view of the potential outcome framework for causal inference. We emphasize that there is a lot we can learn from the design of randomized controlled trials (RCTs) for estimating causal effects in the context of observational data. Furthermore, we will stress the importance of quantifying uncertainty around causal effects and how to conduct sensitivity analyses of causal conclusions with regard to violations of key assumptions.

2. The World of Data Science Is About Observational Data

Confounding bias is a key challenge when estimating causal effects from observational data. Let's assume that we are conducting an observational study to estimate the causal effect of a new drug compared to an older drug to lower blood pressure. Because the study is observational, it is highly likely that people who took the new drug are systematically different from people who took the older drug with respect to their socioeconomic and health status. For example, it is possible that individuals with a higher income might have easier access to the new treatment and at the same time

might be healthier than individuals with low income. Therefore, if we compare individuals taking the new drug to individuals taking the older drug without *adjusting* for income, we might conclude that the new drug is effective, when instead the difference we observe in blood pressure is due to individuals taking the new drug being richer and healthier to begin with.

For now, we can assume that a variable is a potential confounder if it is a pretreatment characteristic of the subjects (e.g., income) that is associated with the treatment (e.g., getting the new drug) and also associated with the outcome (e.g., blood pressure).¹

In our second example we define the treatment and the outcome as follows:

- Treatment - getting a dog $W = 1$, not getting a dog $W = 0$
- Outcome - severe depression symptoms $Y = 1$, mild depression symptoms $Y = 0$ measured after the treatment assignment W

Try our interactive timeline exploring whether adopting a puppy will improve the severity of depression symptoms in one year.

Visit the web version of this article to view interactive content.

Confounders could mask or *confound* the relation between W and Y , which complicates causal attribution or leads to potentially incorrect inferences. For the depression/dog example (Figure 1), a potential confounder is the severity of depression symptoms (denoted by X) before treatment assignment. It is reasonable to believe that individuals with severe symptoms of depression pretreatment ($X = 1$) are more likely to adopt a dog ($W = 1$) than people with mild symptoms of depression ($X = 0$). Furthermore, individuals with severe symptoms of depression before the treatment assignment ($X = 1$) are more likely to have severe symptoms of depression after the treatment assignment Y , than individuals with mild symptoms of depression ($X = 0$).

RCTs are the gold standard study design used to estimate causal effects. To assess the causal effect on survival of getting a new drug compared to a placebo, we could randomize half of the patients enrolled in our study. Half would receive the new drug ($W = 1$), and the other half would receive a placebo ($W = 0$). Randomization is particularly important to establish the efficacy and safety of drugs (new and existing) ([Collins et al., 2020](#)). This is because randomizing patients eliminates systematic

differences between treated and untreated observations. In other words, randomization ensures that these two sets of observations are as similar as possible with respect to all potential confounders, regardless of whether we measure these potential confounders, and are identical on average. If the distribution over the measured and unmeasured confounders are the same in the two groups, then we can use the treated observations to infer what would have happened to the untreated observations.

Unfortunately, randomization is often not possible, either because there are ethical conflicts (such as exposure to environmental contaminants) or because it is challenging to implement. In the latter case, the most constraining factors are the time and monetary expense of data collection. Additional limitations of randomization include inclusion criteria that are too strict and cannot study large and representative populations ([Athey & Imbens, 2017](#)). Moreover, inclusion criteria usually focus on simplified interventions (e.g., randomization to a drug versus placebo) that do not mirror the complexity of real-world decision-making. While the credibility (*internal validity*) and ability to advance scientific discovery of RCTs is well accepted (e.g., 2019 Nobel Memorial Prize in Economic Sciences, [Banerjee et al., 2015](#); [Duflo et al., 2007](#)), there are large classes of interventions and causal questions for which results that have a causal interpretation can only be gathered from observational data.

Fortunately, in this new era of data science, we have access to significant observational data. We can, for example, easily identify large and representative populations of cancer patients, and determine from medical records who received standard therapy or new therapy (or multiple concomitant therapies). Additionally, we can ascertain age, gender, behavioral variables, income, and health status before treatment assignment and assess cancer recurrence and survival ([Arvold et al., 2014](#)). However, because we *observe* who receives treatment instead of *randomizing* who receives treatment, the treated and untreated subpopulations are likely to be systematically different with respect to each of the potential confounders. Without adjusting for systematic differences between the treated and untreated populations, our inference on causal effects will be biased. Given the significant amounts of available data, it is tempting to use correlations observed in the data as evidence of causation; but strong correlation can lead to faulty conclusions when quantifying causal effects.

Figure 1 provides an example of confounding. Let's assume we compare two samples: one that adopts a dog ($W = 1$) and one that does not ($W = 0$). Then within each of

these two populations, we calculate the rate of experiencing severe symptoms of depression $Y = 1$. We found that adopting a dog appears to make the symptoms of depression worse: if you have a dog you are 5% more likely (83% versus 78%) to experience severe symptoms of depression. Should we advise people not to own dogs? The problem with this analysis is that we ignore the fact that the subjects might be different in ways that would bias the conclusions. As mentioned before, a key potential confounder is the degree of severity of their depression symptoms *before* they were assigned the treatment (X). For example, let's stratify the two populations (treated and untreated) based on whether they experience severe or milder depression symptoms of ($X = 1$ versus $X = 0$) before treatment assignment. We find that within these two population strata, adopting a dog reduces the rate of experiencing severe symptoms of depression.

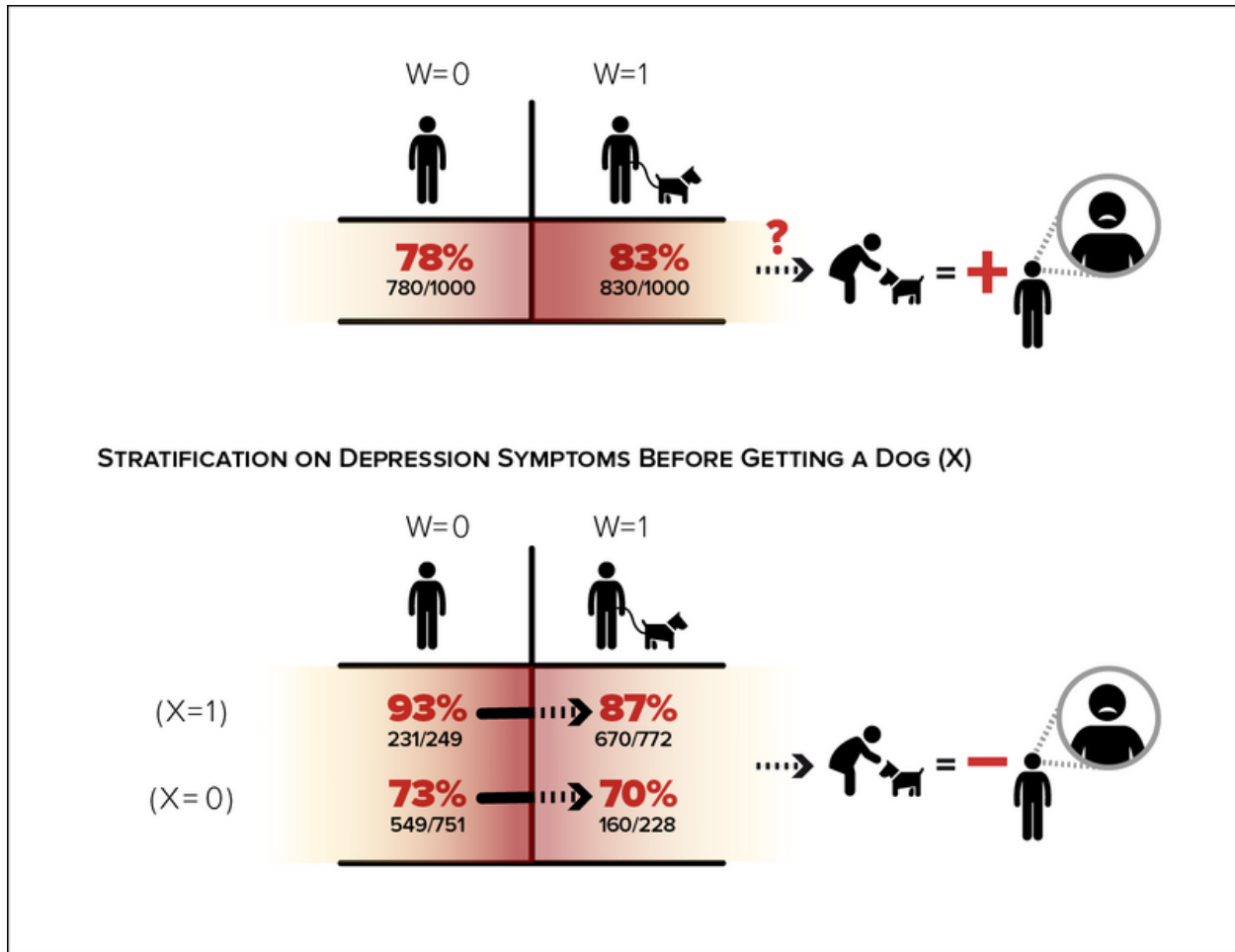


Figure 1. Simpson’s Paradox. Top panel: rate of experiencing depression symptoms one year after getting a dog $W = 1$ (83%) and one year after not getting a dog $W = 0$ (78%). Bottom panel: rate of experiencing severe depression symptoms one year after getting a dog $W = 1$ and one year after not getting the dog $W = 0$, but separately for the two subpopulations that experienced severe or mild symptoms of depression before treatment assignment, denoted by $X = 1$ and $X = 0$, respectively. In the top panel, where we do not stratify by levels of X , getting a dog seems to increase the degree of severity of depression symptoms. In the bottom panel, where we stratify by levels of X , getting a dog appears to decrease the severity of depression symptoms.

This is an example of what is known as Simpson’s paradox. Here the paradox occurs because people with severe depression symptoms before treatment assignment are more likely to adopt a dog. If we define by $e_i = P(W_i|X_i)$ the propensity of adopting a dog conditional to the level of depression symptoms pretreatment assignment, then in this example $P(W_i = 1|X_i = 1) = 772/(772 + 249)$ is higher than $P(W_i = 1|X_i = 0) = 228/(228 + 751)$. In other words, the assignment to treatment, who gets a dog and who does not, is not completely random, as in an RCT. It is influenced

by the preexisting level of depression of the study subjects. Situations like these are very common in observational studies. We argue that the potential outcome (PO) framework detailed following allows us to design an observational study and clarify the assumptions that are required to estimate the causal effects in these studies. Such assumptions translate expert knowledge into identifying conditions that are hard, if not impossible, to verify from data alone.

The PO framework is rooted in the statistical work on randomized experiments by [Fisher \(1918, 1925\)](#) and [Neyman, 1923](#)), extended by [D. Rubin \(1976\)](#) and [D. B. Rubin \(1974, 1977, D. B. Rubin, 1978\), 1990](#)) and subsequently by others to apply to observational studies. This perspective was called *Rubin's Causal Model* by [Holland \(1986\)](#) because it viewed causal inference as the result of missing data, and proposed explicit mathematical modeling of the assignment mechanism to reveal the observed data.

3. The Design Phase of a Study

We begin this section by describing the key distinctions between an RCT and an observational study. Table [1](#) summarizes and contrasts the main differences between RCTs and observational studies, and includes guidance on how to conduct causal inference in the context of observational studies in the last column.

The appeal of the RCT is that the design phase of the study (e.g., units, treatment, and timing of the assignment mechanism) is clearly defined a priori before data collection, including how to measure the outcome. In this sense, the RCT design is always prospective: Treatment assignment randomization always occurs before the outcome is measured. A key feature of RCTs is that the probability of getting the treatment or the placebo, defined as the propensity score, is known – under the experimenter's control – and it does not depend on unobserved characteristics of the study subjects.

Randomization of treatment assignment is also fundamentally useful because it balances observed and unobserved covariates between the treated and control subjects. Once the design phase is concluded, the experimenter can proceed with the analysis phase, that is, estimating the causal effects based on a statistical analysis protocol that was preselected without looking at the information on the outcome. The separation between design and analysis is critical as it guarantees *objective* causal inference. In other words, it will prevent the experimenter from picking and choosing the estimation method that would lead to their preferred conclusion [D. B. Rubin \(2008\)](#).

In observational studies the treatment conditions and the timing of treatment assignment are observed after the data have been collected. The data are often collected for other purposes – not explicitly for the study. As a result, the researcher does not control the treatment assignment mechanism. Moreover, the lack of randomization means there is no guarantee that covariates are balanced between treatment groups, which could result in systematic differences between the treatment group and the control group. Traditionally, practitioners do not draw a clear distinction between the design and analysis phase when they analyze observational data. They estimate causal effects using regression models arbitrarily choosing covariates. This lacks the clear protocol of RCTs. To make objective causal inference from observational studies, we must address these challenges. Luckily, it is possible to achieve objective causal inferences from observational studies with careful design that approximates a hypothetical, randomized experiment. A carefully designed observational study can duplicate many appealing features of RCTs, and provide an objective inference on causal effects ([M. A. Hernán & Robins, 2016](#); [D. B. Rubin, 2007](#), [D. B. Rubin \(2008\)](#)). In the context of causal inference, the design of an observational study involves at least three steps (presented following). These steps should be followed by an analysis phase where the estimation approach is defined according to a pre-specified protocol as in the RCT. The three steps in the design phase are:

1. define experimental conditions and potential outcomes (subsection 3.1);
2. define causal effect of interest, including assumptions for identifiability (subsection 3.2);
3. construct a comparison group (subsection 3.3).

Table 1. Randomized Control Trials Versus Observational Studies

Study characteristics	Randomized Control Trials	Observational Studies	Best practices for causal inference in observational studies
<i>Basic concepts</i>			
Units	Defined before data are collected	Generally not specified	Define units and treatments
Treatment			Determine timing of treatment assignment relative to measured variables
Timing of treatment assignment			
<i>Design phase versus Analysis phase</i>	Separated	The separation is unclear	Hide outcome data until design phase is complete
<i>Treatment assignment mechanism</i>			
Unconfounded	✓	?	Do data contain key confounders?
Probabilistic	✓	?	Yes: Assume unconfoundedness
Known	✓	?	No: Give up, find a better data source, or use a different identification strategy
<i>Covariate balancing</i>	Guaranteed (in expectation)	Not guaranteed	<ol style="list-style-type: none"> 1. Assess overlap in covariates distribution 2. Remove units not similar to any units in the opposite treatment group 3. Find subgroups in which the treatment groups are balanced on covariates
<i>Type of analysis</i>	By protocol	Regression	Analyze according to a pre-specified protocol Complement results with sensitivity analysis to deviations from assumptions

3.1. Define the experimental conditions and the potential outcomes.

The first step to address a causal question is to identify the conditions (actions, treatments, or exposure levels) required to assess causality. To define the causal effect of $W = 1$ versus $W = 0$ on Y , one must postulate the existence of two potential outcomes: $Y(W = 1)$ and $Y(W = 0)$. As the name implies, both variables are potentially observable, but only the variable associated with the observed (or assigned) action will be observed. The critical feature of the notion of a cause is that the value of W for each unit can be manipulated. To illustrate this idea, we now introduce the subscripts $(t - 1)$ and t to indicate time at or before the treatment assignment and time after treatment assignment (e.g., one year later). For example, let's say Amy (subject i at a specific time $t - 1$) adopted a dog ($W_{i,t-1} = 1$) and we observe her depression symptoms one year later t ($Y_{i,t}(W_{i,t-1} = 1) = Y_{i,t}(1)$). Now we must assume

that W can be manipulated, that is, we must be able to hypothesize a situation where Amy would not get a dog ($W_{i,t-1} = 0$). So, $Y_{i,t}(1)$ is observed and $Y_{i,t}(0)$ is unobserved.

3.2. Define the causal effect of interest.

We define a unit-level causal effect as the comparison between the potential outcome under treatment and the potential outcome under control. For example, the causal effect of W on Y for unit i is defined as $Y_{i,t}(1) - Y_{i,t}(0)$.

As we can see from Figure 2, the causal effect of interest is not a pre/post comparison of the depression symptoms for Amy (defined as $Y_{i,t}(1) - X_{i,t-1}$). It is, instead, the difference between her two potential outcomes evaluated at time t , defined as $Y_{i,t}(1) - Y_{i,t}(0)$, where $Y_{i,t}(1)$ is observed, whereas $Y_{i,t}(0)$ is not. These differences are called individual treatment effects (ITE) ([D. B. Rubin, 2005](#)).

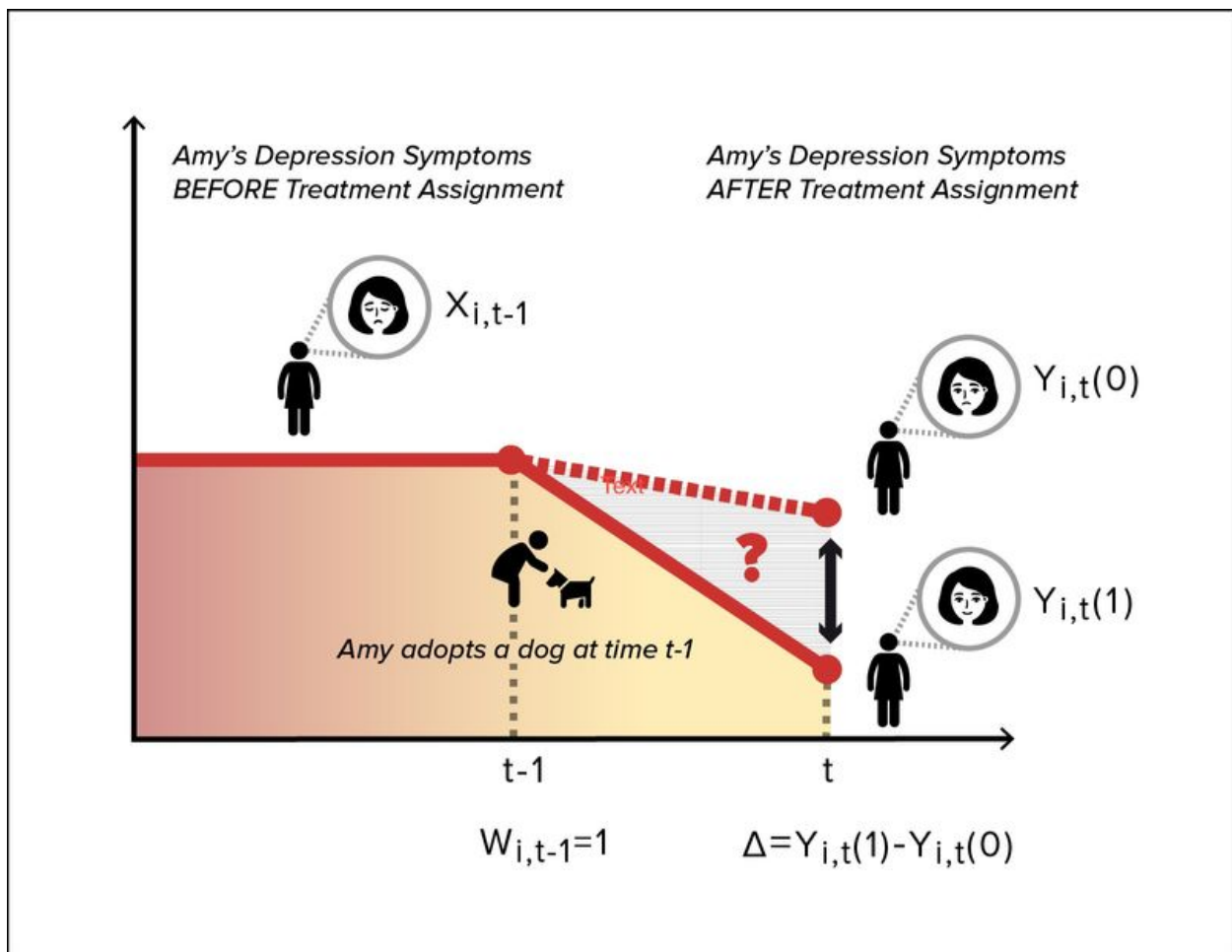


Figure 2. Causal effects as contrasts of potential outcomes at a given time point t .

While difficult to estimate [Funk et al., 2011](#), ITEs inform treatment effect heterogeneity ([e.g., Dahabreh et al., 2016](#)) and facilitate decision making in individualized settings where an estimate of the causal effect averaged across all the subjects in the sample may not be very practical ([e.g., J. Li et al., 2019](#)). We will return to this issue in Section 5. The last column of Table 2 introduces the concept of summarizing individual causal effects across the population of interest. For example, we can summarize the unit-level causal effects by taking the average difference, or by taking the difference in median values of the $Y_i(1)$ s and $Y_i(0)$ s, respectively. We typically focus on causal estimands that contrast potential outcomes on a common set of units (our target sample of size N), for example the average treatment effect (ATE): $\frac{1}{N} \sum_{i=1}^N (Y_i(1) - Y_i(0)) = \bar{Y}(1) - \bar{Y}(0)$.

Table 2. The Science

Units	Covariates	Potential Outcomes	Potential Outcomes	Unit-level Causal Effects	Summary of Causal Effects
1	X_1	$Y_1(1)$	$Y_1(0)$	$Y_1(1)$ vs $Y_1(0)$	Comparison of $Y_i(1)$ vs $Y_i(0)$ for a common set of units
⋮	⋮	⋮	⋮	⋮	
i	X_i	$Y_i(1)$	$Y_i(0)$	$Y_i(1)$ vs $Y_i(0)$	
⋮	⋮	⋮	⋮	⋮	
N	X_N	$Y_N(1)$	$Y_N(0)$	$Y_N(1)$ vs $Y_N(0)$	

The fundamental challenge is that we will never observe a potential outcome under a condition other than the one that actually occurred, so that we will never observe an individual causal effect (see Table 3). [Holland \(1986\)](#) refers to this as the *fundamental problem of causal inference*. We typically refer to the missing potential outcome as the *counterfactual* and to the observed outcome as the *factual* outcome. Causal inference relies on the ability to predict the counterfactual outcome. It is important to note that methods used to predict or impute counterfactuals are different than off-the-shelf prediction or imputation often used for missing values. This is because we will never be able to find data where both potential outcomes ($Y_i(1), Y_i(0)$) are simultaneously observed on a common set of units. Table 3 highlights other fundamental implications of this representation of causal parameters: a) Uncertainty remains even if the N units are our finite population of interest, because of the missingness in the potential outcomes; b) The inclusion of more units provides additional information (more factual outcomes) but also increases missingness (more counterfactual outcomes).

Data alone are not sufficient to predict the counterfactual outcome. We need to introduce several assumptions that essentially embed subject matter expert knowledge

(Angrist & Pischke, 2008). This is why machine learning alone cannot resolve causal inference problems, an issue discussed further in Section 5. To identify a causal effect from the observed data, we have to make several assumptions.

Table 3. What we are able to observe about the Science

Units	Covariates	Treatment	Potential Outcomes		Unit-level Causal Effects
	X	W	$Y(1)$	$Y(0)$	
1	X_1	1	$Y_1(1)$?	?
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
i	X_i	0	?	$Y_i(0)$?
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
N	X_N	1	$Y_N(1)$?	?

Assumption 1: Stable Unit Treatment Value Assumption (SUTVA). SUTVA, introduced and formalized in a series of papers by Rubin (see D. B. Rubin, 1980, 1986, 1990), requires that there is no interference and no hidden version of the treatment. No interference assumes that the potential outcomes of a unit i only depend on the treatment unit i receives, and are not affected by the treatment received by other units.

For example, epidemiological studies of the causal effects of nonpharmaceutical interventions (e.g., stay-at-home advisory) on the probability of getting COVID19 violate the assumption of no interference. This is because the individual level outcome (whether or not a subject is infected) depends on whether he/she complies with the stay-at-home advisory, but also on whether or not others in the same household also comply with the stay-at-home advisory.

Spillover effects are a violation of the no-interference assumption. These examples often occur when the observations are correlated in time or space. In our ice cream case study this assumption is likely to hold as it is reasonable to assume that the only person who benefits from weight loss is the person who stopped eating ice cream. SUTVA violations can be particularly challenging in air pollution regulation studies as pollution moves through space and presents a setting for interference. Intervening at one location (e.g., a pollution source) likely affects pollution and health across many locations, meaning that potential outcomes at a given location are probably functions

of local interventions as well as interventions at other locations [Forastiere et al., 2020](#); [Papadogeorgou et al., 2019](#). The condition of no hidden version of treatments requires that potential outcomes not be affected by *how* unit i received treatment. This assumption is related to the notion of consistency ([M. D. Hernán & Robins, 2020](#); [M. A. Hernán, 2016](#)). For example, how Amy adopted a dog (a friend giving away puppies or driving to a breeder) does not affect Amy's outcome.

Our ability to estimate the missing potential outcomes depends on the treatment assignment mechanism. That is, it depends upon the probabilistic rule $W = 1$ versus $W = 0$, which determines whether $Y(1)$ or $Y(0)$ is observed. The assignment mechanism is defined as the probability of getting the treatment conditional on $X, Y(1), Y(0)$, for example, $P(W|X, Y(1), Y(0))$. This expression will be simplified under the next assumption.

Assumption 2: No unmeasured confounding. The assignment mechanism is unconfounded if: $P(W | X, Y(1), Y(0)) = P(W | X)$. Unconfoundedness is also known as no unmeasured confounding assumption, or conditional independence assumption. This means that if we can stratify the populations within subgroups that have the same covariate values (e.g., same age, gender, race, income), then within each of these strata, the treatment assignment (e.g., who gets the drug and who does not) is random.

The assumption allows us to provide a formal definition of a confounder. Although there is no consensus regarding a unique and formal definition, we adopt the one proposed by [T. J. VanderWeele & Shpitser \(2013\)](#): a preexposure covariate X is said to be a confounder for the effect of W on Y if there exists a set of covariates X^* such that the effect of W on Y is unconfounded conditional on (X^*, X) , but it is not for a subset of (X^*, X) . Equivalently, a confounder is a member of a minimally sufficient adjustment set.

Unconfoundedness is critical to estimate causal effects in observational studies. As $Y_i(1)$ is never observed on subjects with $W_i = 0$ and $Y_i(0)$ is never observed on subjects with $W_i = 1$, we cannot test this assumption, and so its plausibility relies on subject-matter knowledge. As a result, sensitivity analysis should be conducted routinely to assess how the conclusions will change under specific deviations from this assumption (discussed in Section 6). Moreover, this assumption may fail to hold if some relevant covariates are not observed, or if decisions are based on information on potential outcomes. For example a *perfect doctor* ([G. W. Imbens & Rubin, 2015](#)) gives a drug to patients based on who benefits from the treatment (e.g., $W_i = I(Y_i(1) > Y_i(0))$): the assignment is confounded – that is, depends on the potential outcomes,

irrespective of the covariates we are able to condition on. We discuss these situations in our final remarks.

Assumption 3: Overlap or positivity. We define the propensity score for subject i as the probability of getting the treatment given the covariates ([Rosenbaum & Rubin, 1983](#)), $e_i = P(W_i = 1 \mid X_i)$. The assumption of overlap requires that all units have a propensity score that is between 0 and 1, that is, they all have a positive chance of receiving one of the two levels of the treatment.

In the depression/dog example, this may be violated if some people in the population of interest are allergic to dogs and therefore their probability of getting a dog is zero. In the clinical example, this hypothesis is commonly violated if a patient has a genetic mutation that prevents them from receiving the treatment being tested. Because the propensity score can be estimated from data, we can check if overlap holds. If for some units the estimated e_i is very close to either 1 or 0, then these units are only observed under a single experimental condition and therefore contain very little information about the causal effect. In this situation, strong assumptions are necessary. For example, a strong assumption is to say that the functional form that relates the covariates with the outcome holds also outside of the observed range of the covariates. A more formal approach of how to overcome violations of the positivity assumption is presented in [Netherly et al., 2019](#). If assumptions 2 and 3 are both met, then we conclude that the assignment mechanism is *strongly ignorable* ([Rosenbaum & Rubin, 1983](#)). Classic randomized experiments are special cases of strongly ignorable assignment mechanisms.

3.3. How to construct an adequate comparison group.

Once you have identified relevant potential confounders, and assuming they are sufficient for unconfoundedness to hold, the issue of confounding can be resolved by constructing an adequate comparison group. This is a crucial step in the design of an observational study. Our goal is to synthetically recreate a setting that is very similar to a randomized experiment, so the joint distribution of all the potential confounders is as similar as possible between the treatment and control groups ([Ho et al., 2007](#); [Stuart, 2010](#); [Stuart & Rubin, 2008](#)). For instance, let's return to our example about Amy (in this case we drop the subscript t). Amy (subject i) got a dog $W_i = 1$, so we observe $Y_{i,obs} = Y_i(W_i = 1)$, whereas $Y_{i,mis} = Y_i(W_i = 0)$. To estimate a causal effect we need to predict $Y_i(W_i = 0)$, that is, what we would expect the severity of Amy's symptoms to be under the hypothetical (not observed) scenario in which Amy did not get the dog ($W_i = 0$). To predict the missing counterfactual $Y_i(W_i = 0)$, we need to

define an adequate control group. Based on the considerations made so far, we need to find subjects similar to Amy with respect to the potential confounders (age, race, income, health status, severity of depression symptoms) before treatment assignment. The only difference between the matched subjects and Amy is that they did not get a dog. This should be done for Amy and for any subject in our target population who got a dog. With a large number of confounders, matching on all confounders *exactly* may not be feasible. A common approach to address this challenge is to use propensity scores (e_i) and match subjects with respect to e_i . The estimated propensity score is a univariate summary of all covariates and is crucial to estimating causal effects under unconfoundedness ([Imai & Ratkovic, 2014](#); ([G. W. Imbens & Rubin, 2015](#)); [Robins & Rotnitzky, 1995](#); [Robins et al., 1995](#)). Subjects sharing the same value of the propensity score have the same distribution of the observed potential confounders whether they are treated or not. Estimated propensity scores can be applied in the design phase to assess overlap and construct a comparison group through matching, stratifying, or weighting observations [D. B. Rubin \(2008\)](#). Covariate balance can also be viewed as an optimization problem. Procedures based on this idea either directly optimize weights or find optimal subsets of controls such that the mean or other characteristics of the covariates is the same in the treatment and control group ([Diamond & Sekhon, 2013](#); [F. Li et al., 2018](#); [F. Li & Thomas, 2018](#); [Zubizarreta, 2012](#), [Zubizarreta \(2015\)](#); [Zubizarreta, Paredes, et al., 2014](#)).

So far we have only discussed what we are interested in estimating, the study design, and the problem of confounding. Now we will discuss how we estimate the causal effects. Being able to identify causal effects is a feature of the causal reasoning used in the potential outcome framework. Without including causal reasoning in the design phase you cannot recover the causal effect even with the most sophisticated machine learning or nonparametric methods (*e.g.*, [Mattei & Mealli, 2015](#)).

4. Estimation

Causal estimands such as $ATE = \bar{Y}(1) - \bar{Y}(0)$ are a function of the observed Y_{obs} and the missing potential outcomes (the counterfactuals) Y_{mis} . Therefore, an estimation strategy needs to implicitly or explicitly impute Y_{mis} . What follows is not a comprehensive review of all estimation methods of causal effects, but rather key ideas of Bayesian estimation of the average treatment effect ([Ding & Li, 2018](#); ([G. W. Imbens & Rubin, 2015](#)); [D. B. Rubin, 1978](#)).

4.1. Bayesian methods for the imputation of the missing counterfactuals after the design phase is concluded.

Within the model-based Bayesian framework for causal inference ([D. B. Rubin, 1975](#), [D. B. Rubin, 1978](#)), the Y_{mis} are considered unknown parameters. The goal is to sample from their posterior predictive distribution conditionally to the observed data defined as:

$$(4.1) \quad P(Y_{\text{mis}} | Y_{\text{obs}}, X, W) \propto P(X, Y(1), Y(0))P(W | X, Y(1), Y(0))$$

where $P(X, Y(1), Y(0))$ denotes the model for the potential outcomes, while $P(W | X, Y(1), Y(0))$ denotes the model for the treatment assignment. By sampling from this posterior distribution we can multiply impute Y_{mis} , and then estimate ATE, or any other causal contrast, and its posterior credible interval ([Mealli et al., 2011](#); [D. B. Rubin, 1978](#)). Note that this missing data imputation exercise is critically different from a usual prediction task: the two models introduced above contain expert knowledge (for example, the assumption of strong ignorability or the inclusion of relevant covariates) that cannot be retrieved from data alone. Under unconfoundedness or no unmeasured confounding (Assumption 2: $P(W | X, Y(1), Y(0)) = P(W | X)$), and assuming the parameters of the model for the potential outcomes are a priori independent of the parameters of the model for the assignment mechanism, then the posterior distribution of the missing potential outcomes only depends on the parameters of the model for the outcomes. Specifically, assuming *exchangeability* ([de Finetti, 1963](#)), there exists a parameter vector θ having a known prior distribution $p(\theta)$ such that:

$$(4.2) \quad P(Y(0), Y(1), W, X) = \int \prod_i P(Y_i(0), Y_i(1), W_i, X_i, \theta)p(\theta)$$

The posterior predictive distribution of the missing data, $P(Y_{\text{mis}} | Y_{\text{obs}}, W, X)$, can be written as

$$(4.3) \quad P(Y_{\text{mis}} | Y_{\text{obs}}, W, X) = \frac{P(Y(0), Y(1), W, X)}{\int P(Y(0), Y(1), W, X)dY_{\text{mis}}} = \frac{\int \prod_i P(W_i | Y_i(0), Y_i(1), X_i, \theta)P(Y_i(0), Y_i(1) | X_i, \theta)P(X_i | \theta)p(\theta)d\theta}{\iint \prod_i P(W_i | Y_i(0), Y_i(1), X_i, \theta)P(Y_i(0), Y_i(1) | X_i, \theta)P(X_i | \theta)p(\theta) d\theta dY_{\text{mis}}}$$

Let $\theta_{W|X}$, $\theta_{Y|X}$ and θ_X , denote the unknown parameters corresponding to the distribution of the treatment assignment mechanism (e.g., the propensity score), the distribution of potential outcomes, and the distribution of covariates, respectively. Then, given ignorability, the propensity score $P(W_i | X_i, \theta_{W|X})$ and the covariates' distribution $P(X_i | \theta_X)$ cancel out in Equation (4.3), which simplifies to:

$$(4.4) \quad P(Y_{\text{mis}} | Y_{\text{obs}}, W, X) \propto \int \prod_i P(Y_i(0), Y_i(1) | X_i, \theta_{Y|X}) p(\theta_{Y|X})$$

Equation (4.4) shows that, under ignorable treatment assignments, the potential outcome model needs to be specified $P(Y_i(w) | X_i, \theta_{Y|X})$ for $w = 0, 1$, as well as the prior distribution $p(\theta_{Y|X})$, to derive the posterior distribution of the causal effects.² Therefore, the most straightforward Bayesian approach to estimate causal effects under ignorability is to specify models for $Y(1)$ and $Y(0)$ that are conditional to covariates and some parameters and then draw the missing potential outcomes from their posterior predictive distribution, which will also derive the posterior distribution of any causal estimand.

As such, it seems like propensity scores, which are central to balance the covariates in the design stage, do not affect Bayesian inference for causal effects under ignorability. However, as noted by [\(D. B. Rubin, 1985\)](#), for effective calibration of Bayesian inference of causal effects (i.e., good frequentist properties), good covariate balance is necessary [\(Ding & Li, 2019, see also\)](#). That is, if covariates are balanced between treatment groups, inference on causal effects derived from the estimation of the outcome model are robust and not very sensitive to model assumptions. Flexible semi- and non-parametric specifications of the outcome model can be found in literature including Bayesian Additive Regression Models (BART; [Hahn et al., 2020](#); [J. L. Hill, 2011](#)) or Bayesian Cubic Splines [\(Chib & Greenberg, 2010\)](#).

4.2. Bayesian methods for joint estimation of the outcome and propensity score models and the feedback problem.

Some Bayesian methods explicitly include the estimation of the propensity score in the causal effects' estimation procedure. Some of these approaches involve the specification of a regression model for the outcome with the propensity score as a single regressor, arguing that this modeling task is simpler than specifying a model for the outcome conditional on the whole set of (high-dimensional) covariates [\(D. B. Rubin, 1985\)](#). This approach can be improved by adjusting for the residual covariance between X and Y at each value of $e(X)$ [\(Gutman & Rubin, 2015\)](#), or by specifying an

outcome model conditional on the covariates X and the propensity score ([Zheng & Little, 2005](#)).

These are two-stage methods that separate design (estimation of the propensity score) from analysis. Some authors have proposed a single-step Bayesian approach to merge the two stages. For example, [McCandless et al. \(2009\)](#) proposed a joint modeling approach to estimate the outcomes and the propensity score in this setting. However [C. M. Zigler et al. \(2013\)](#) has shown that merging the two steps may induce the *feedback problem* if the parameters of the two models are a priori dependent. In this instance, the outcome information enters the estimation of the propensity score, contradicting ignorability, and can lead to a distorted estimate of the propensity scores and compromise estimates of causal effects. Consequently, these types of Bayesian models are not often used. [Liao and Zigler \(2020\)](#) considers how uncertainty associated with the design stage impacts estimation of causal effects in the analysis stage. Uncertainty in the design stage may stem from the propensity score estimation but also from how the propensity score is used. Liao and Zigler propose a procedure that obtains the posterior distribution of causal effects after marginalizing distributed over design-stage outputs.

4.3. Bayesian methods to overcome violations of the positivity.

To address violations of the positivity assumption Bayesian methods, including BART and splines, have also been developed. For example, [Nethery et al., 2019](#) have recently developed a novel Bayesian framework to estimate population average causal effects with minor model dependence and appropriately large uncertainties in the presence of nonoverlap. In this approach, the tasks of estimating causal effects in the overlap and nonoverlap regions are delegated to two distinct models, suited to the degree of data in each region. In this case, tree ensembles are used to nonparametrically estimate individual causal effects in the overlap region, where the data can speak for themselves. In the nonoverlap region, where insufficient data support makes reliance on model specification necessary, individual causal effects are estimated by extrapolating trends from the overlap region via a spline model. The authors showed that their proposed approach has a good performance compared to approaches that perform estimation only in the area of overlap ([J. Hill & Su, 2013](#); [C. Zigler & Cefalu, 2017](#)).

4.4. Bayesian and non-Bayesian methods that are model-free.

Estimation methods under ignorability that are model-free (i.e., matching, stratification, weighting) are usually frequentist and treat potential outcomes as fixed values instead of random variables ([G. Imbens, 2004](#); [G. W. Imbens & Rubin, 2015](#)); [Yao et al., 2020](#)). Some authors have proposed quasi-Bayesian approaches that involve only the Bayesian estimation of the propensity score, and then estimate the causal effect via matching ([An, 2010](#)), stratification or weighting ([Saarela et al., 2015](#)). These approaches are not fully Bayesian in that they incorporate only the uncertainty in the propensity score estimation and not in the imputation of the missing potential outcomes. The frequentist methods have been improved by combining them with outcome regression adjustments (e.g., [Abadie & Imbens, 2011](#)). Robins and colleagues (e.g., [Bang & Robins, 2005](#); [Funk et al., 2011](#); [Knaus, 2020](#); [Lunceford & Davidian, 2004](#); [Robins, 2000](#); [Robins & Rotnitzky, 1995](#); [Robins et al., 1995](#); [Scharfstein et al., 1999](#)) have proposed a class of double robust (DR) estimators that combine inverse probability weighting estimator (IPW) with an outcome regression. Interestingly, [Gustafson \(2012\)](#) casted DR estimator from a Bayesian perspective as a weighted average of a parametric model and a saturated model for the outcome conditional on covariates, with weights that depend on how well the parametric model fits the data. As a result, it can also be viewed as a Bayesian model average estimator ([Cefalu et al., 2016](#)).

4.5. Bayesian and non-Bayesian methods to account for variable selection either in the propensity or in the outcome model.

[C. M. Zigler & Dominici \(2014\)](#) proposed a Bayesian model averaging method to adjust for the uncertainty in the selection of the covariates in the propensity score model, extending [Wang et al. \(2012\)](#) [Wang et al., 2015](#)). When the number of the potential confounders is larger than the number of observations, approaches for dimension reduction and penalization are required. The standard approaches (e.g., the LASSO) generally aim to predict the outcome, and are less suited for estimation of causal effects. Under standard penalization approaches, if a variable X is strongly associated with the treatment W but weakly associated with the outcome Y , its coefficient will be reduced toward zero leading to confounding bias. ([Belloni et al., 2014a](#), [Belloni et al. \(2014b\)](#)) proposed a modified version of LASSO, called double LASSO, to reduce confounding bias. There are several Bayesian alternatives that usually outperform such approaches, for example using continuous spike and slab priors on the covariates' coefficients in the outcome and propensity score models ([Antonelli & Dominici, 2018](#); [Antonelli et al., 2019](#); [Cefalu et al., 2016](#); [Wang et al., 2015](#)).

5. The Perils and the Strengths of Machine Learning Methods in Causal Inference

We have presented approaches for the estimation of Y_{mis} with high-dimensional covariates and nonparametric methods, and the related issue of variable selection. At first glance, these tasks could be addressed by implementing off-the-shelf machine learning methods, but there are challenges. In this section we provide critical insights regarding the application of machine learning methods in causal inference.

On a high level, there are two broad categories of machine learning methods: unsupervised and supervised learning. Unsupervised learning ([see Hastie et al., 2009](#)) refers to those techniques used to draw inferences from data sets consisting of input data (e.g., features X) without labeled outcomes. These methods are used to perform tasks such as clustering, pattern mining, dimension reduction, anomaly detection. Clustering algorithms essentially group units based on their mathematical similarities and dissimilarities of features X . These tools can be used for example to find groups of basketball or soccer players with similar attributes and then interpret and use these clusters to form teams or to target coaching. In supervised learning (henceforth in this work, when using *machine learning* we will refer to these techniques), the predictors (i.e., covariates, features) X and the outcome Y are both observed. The goal is often to estimate the conditional mean of an outcome Y given a set of covariates or features X , to ultimately predict Y . These methods include decision trees ([Breiman et al., 1984](#)), random forests ([Breiman, 2001](#)), gradient boosting ([Friedman, 2001](#)), support vector machines ([Cortes & Vapnik, 1995](#); [Suykens & Vandewalle, 1999](#)), deep neural networks ([e.g., Farrell et al., 2018](#); [LeCun et al., 2015](#)), ensemble methods ([e.g., Dietterich, 2000](#)), and variable selection tools such as LASSO ([Hastie et al., 2015](#)); [Tibshirani, 1996](#)). Regression trees, and random forests as their extension, have become very popular methods for estimating regression functions in settings where out-of-sample predictive power is important. When the outcome Y is an unordered discrete response, these supervised learning algorithms attempt to resolve classification problems—for example, detecting spam emails. In this instance, a machine learning algorithm is trained with a set of spam-emails labeled as spam and not-spam emails labeled as not-spam, so that a new email can be classified as either spam or not-spam. Deep learning methods are another general and flexible approach to estimate regression functions. They perform very well in settings with extremely large number of features—like image recognition or image diagnostics ([He et al., 2016](#); [Simonyan & Zisserman, 2014](#))—when the size of the data sets is large enough. These methods typically require a large amount of tuning to work well in practice, relative to

other methods such as random forests, and as a result we will not discuss them further in this article.

Since machine learning methods aim to estimate the conditional mean of an outcome Y given X it would on the surface appear to be a good fit to exploit machine learning methods to estimate the missing potential outcomes and as a result the causal effects, especially in high-dimensional settings. However, it is not that simple. The following section presents the circumstances under which off-the-shelf machine learning methods might not be appropriate with regard to causal inference. We also discuss how these methods can be adapted to achieve estimation of causal effects; how machine learning and the literature on statistical causal inference can cross-fertilize; and the open questions and problems that machine learning cannot handle on its own.

5.1. Why off-the shelf machine learning techniques might not be appropriate for causal inference.

A key distinction between causal inference and machine learning is that the former focuses on estimation of the missing potential outcomes, average treatment effects, and/or other causal estimands, and machine learning focuses on prediction tasks. Therefore, machine learning could dismiss covariates with limited prediction importance while in causal inference, if these same covariates are correlated with the treatment assignment, they can be important confounders ([Belloni et al., 2014a](#)). As previously discussed, omitting covariates from the analysis that are highly correlated with the treatment can introduce substantial bias in the estimation of the causal effects even if their predicting power is weak. Another major difference is that machine learning methods are typically assessed on their out-of-sample predictive power. This approach has two major drawbacks in the context of causal inference. First, as pointed out by [Athey and Imbens \(2016\)](#), a fundamental difference between a prediction and the estimation of a causal effect is that in causal inference we can never observe the ground truth (e.g., in this context the counterfactual). That is, in our example, because Amy has adopted a dog, we will never be able to measure the severity of the Amy's depression symptoms under the alternative hypothetical scenario where Amy did not adopt a dog. Therefore, standard approaches for quantifying the performance of machine learning algorithms cannot be implemented to assess the quality of prediction of the missing potential outcomes and therefore the causal effects ([Gao et al., 2020](#)). Second, in causal inference we must provide valid confidence intervals for the causal estimands of interest. This is required to make decisions regarding which treatment or treatment regime is best for a given unit or subset of

units, and whether a treatment is worth implementing. Another limitation of machine learning techniques in causal inference is that they are developed, for the most part, in settings where the observations are independent and therefore have limited ability to handle data that is correlated in time and/or in space, such as time series, panel data, and spatially correlated processes. Additionally, most techniques are not able to handle specific structural restrictions suggested by subject-matter knowledge, such as monotonicity, exclusion restrictions, and endogeneity of some variables ([Angrist & Pischke, 2008](#); [Athey & Imbens, 2019](#)).

5.2. How machine learning methods can adapt to the goal of estimation of causal effects.

Machine learning methods have been adapted to address causal inference. One popular approach is to redefine the optimization criteria, which typically are expressed as functions of the prediction errors, to prioritize issues arising in causal inference, such as the controlling for confounders and the discovering of treatment effect heterogeneity ([Chernozhukov et al., 2017](#); [Chernozhukov et al., 2018](#)). For example, the causal tree method proposed by [Athey and Imbens \(2016\)](#) is based on a rework of the criterion function of classification and regression trees ([Breiman et al., 1984](#))—originally aimed at minimizing the predictive error—to maximize the variation in the treatment effects and, in turn, discover the subgroups with the highest heterogeneity in the causal effects (further details are presented in Section 6). Typical regularizing algorithms used in machine learning, such as LASSO, elastic net, and ridge ([Hastie et al., 2015](#)), must prioritize confounding adjustment to avoid missing relevant covariates, as seen in [Belloni et al. \(2014b\)](#) and ([Belloni et al., 2014a](#)) and discussed in Section 4.5. Additional research efforts are required to study the statistical properties of machine learning techniques, as in [Wager & Athey, 2018](#)) and [Athey et al. \(2019\)](#) for random forests and [Jeong and Rockova \(2020\)](#) for BART. Treatment effect estimators based on causal random forests have been shown to be pointwise consistent and asymptotically normal, while estimators based on BART are asymptotically optimal. Both these asymptotic results hold true under a set of specific conditions that are not necessarily mild. However, research has shown that machine learning algorithms show promise when the hardest issues are resolved during design ([Bargagli-Stoffi et al., 2019](#); [Bargagli-Stoffi & Gnecco, 2020](#); [Dorie et al., 2019](#); [Hahn et al., 2019](#); [Knaus et al., 2020](#)). Specifically, it is critical to first identify a good control population, for example using propensity score matching as discussed earlier. After we have achieved this critically important task, then we can be confident that the assumptions of identifiability of causal effects are met. It is only at this stage that the machine

learning techniques provide an excellent tool to predict the missing counterfactuals [Chernozhukov et al., 2018](#). However, it is important to keep in mind that performance will ultimately rely on the flexible parametrization that machine learning methods impose on the data, the plausibility of the unconfoundedness assumption, and the extent of the overlap in the distribution of the covariates ([M. A. Hernán et al., 2019](#)). Even in the presence of a high-dimensional set of covariates, the study design is important ([D'Amour et al., 2017](#)).

5.3. Cross-fertilization between machine learning and causal inference problems.

Several machine learning techniques have been adapted to improve traditional causal inference methods. These include approaches to regularization (the process of adding information to solve ill-posed problems or to prevent overfitting) that scale well to large data sets ([Chang et al., 2018](#)) and the related concept of sparsity (the idea that some variables may be dropped from the analysis without affecting the performance of estimation of causal effects). The use of model averaging and ensemble methods common in machine learning is a practice that is now exploited in causal inference (see Section 4) ([Cefalu et al., 2016](#); [Van der Laan & Rose, 2011](#)).

Framing data analysis as an optimization problem has inspired the development of causal inference methods based on direct covariate balancing. For example, [Zubizarreta \(2015\)](#) which optimizes weights for each observation instead of trying to estimate the propensity score so that the covariate distribution is the same in the treatment and control group. Matrix completion methods were originally developed in machine learning for the imputation of the missing entries of a partially observed matrix ([Candes & Recht, 2009](#); [Recht, 2011](#)). These methodologies can be used to improve causal inference methods for panel data and synthetic control methods ([Abadie et al., 2010](#)) in settings with large N and T. In particular, matrix completion can be successfully adapted for the imputation of missing counterfactuals when a large proportion of potential outcomes is missing ([Athey et al., 2018](#)).

5.4. Open questions and problems that machine learning alone cannot handle.

However, even when adapted to treatment effect estimation, machine learning algorithms must be implemented with extreme caution when there are unresolved key issues surrounding the study design. For example:

- The sample under study is not representative of the population about which we need or want to draw conclusions;
- The number of potential confounders that are measured may not be sufficient: there is nothing in the data that tells us that unconfoundedness holds; causal effect estimation should be followed by a well-designed sensitivity analysis;
- The presence of post-treatment variables that must be excluded (i.e., variables that can be affected by the treatment and are strong predictor of the outcome);
- The lack of overlap ([D'Amour et al., 2017](#); [Nethery et al., 2019](#)) in the distribution of the estimated propensity scores for the treated and untreated, demands the machine extrapolate beyond what is observed;
- If interference is detected, causal estimands (direct and indirect - or spillover - effects) need to be redefined and different estimation strategies need to be implemented ([Arpino & Mattei, 2016](#); [Bargagli-Stoffi et al., 2020](#); [Forastiere et al., 2020](#); [Papadogeorgou et al., 2019](#); [Tortú et al., 2020](#)).

Even in settings where machine learning can accurately predict potential outcomes ([Belloni et al., 2014a](#); [Chernozhukov et al., 2017](#); [Chernozhukov et al., 2018](#); [Hahn et al., 2020](#); [Wager & Athey, 2018](#)), they cannot handle other problems such as missing outcomes ([J. Hill, 2004](#); [Mattei et al., 2014](#)), (nonrandom) measurement error or misclassification of outcome or treatment ([Funk & Landi, 2014](#); [Imai & Yamamoto, 2010](#); [Pierce & VanderWeele, 2012](#)), censoring due to death of the outcome ([Mattei & Mealli, 2007](#); [D. B. Rubin, 2006](#)), and disentangling different causal mechanisms ([Baccini et al., 2017](#); [Forastiere et al., 2018](#); [Mattei & Mealli, 2011](#); [T. VanderWeele, 2015](#)).

6. Treatment Effect Heterogeneity: Is the Treatment Beneficial to Everyone?

Suppose we found statistically significant evidence that a new drug prolonged life expectancy on average for the population under study. Should we encourage anyone, regardless of their age, income, or other diseases to take this new drug? It is often highly desirable to characterize which subgroups of the population would benefit the most or the least from a treatment. These types of questions require us to analyze treatment effect heterogeneity based on pretreatment variables. There is extensive literature on assessing heterogeneity of causal effects that is based on estimating the conditional average treatment effect (CATE), which is defined as

$E[Y(W = 1) | X = x] - E[Y(W = 0) | X = x]$ where $Y(W = 1) | X = x$ and $Y(W = 0) | X = x$ are the potential outcomes in the subgroups of the population

defined by $X = x$. Conditionally to $X = x$, the CATE can be estimated with the same set of the causal assumptions that are needed for estimating the ATE [Athey and Imbens \(2016\)](#). Recently, machine learning methods such as random forests, BART ([Chipman et al., 2010](#)), and forest based algorithms ([Foster et al., 2011](#); [J. L. Hill, 2011](#)), have been used to estimate CATE, especially in the presence of high dimensional X . Despite accurately estimating the CATE using machine learning methods, these methods offer little guidance about which population subgroups are important in the treatment effect heterogeneity. Their parametrization of the covariate space is complicated and difficult to interpret even by human experts. We define interpretability as the degree to which a human can understand the cause of a decision or consistently predict the results of the model ([Kim et al., 2016](#); [Miller, 2019](#)). Decision rules fit well with this nonmathematical definition of interpretability. A decision rule consists of a set of conditions about the covariates that define a subset of the features space and correspond to a specific subgroup. In our recent work ([Lee et al., 2020](#)), we propose a novel causal rule ensemble (CRE) method that ensures interpretability of the causal rules while maintaining a high level of accuracy of the estimated treatment effects for each rule. We argue that in the context of treatment effect heterogeneity, we want to achieve at least two main goals: to (1) discover de novo the rules (that is, interpretable population subgroups) that lead to heterogeneity of causal effects, and (2) make valid inference about the CATE with respect to the newly discovered rules. [Athey and Imbens \(2016\)](#) has introduced a clever approach to make valid inferences in this context. They introduced the idea of a sample-splitting approach that divides the total sample into two smaller samples: (1) to discover a set of interpretable decision rules that could lead to treatment effect heterogeneity (i.e., discovery sample), and (2) to estimate the rule-specific treatment effects and associated statistical uncertainty (i.e., inference sample). This is a very active area of research and one where the integration of machine learning and causal inference could provide important advances.

7. Sensitivity Analysis

Sensitivity analyses can be conducted to bound the magnitude of the causal effects as a function of the degree to which the assumptions are violated to evaluate the robustness of causal conclusions to violations of unverifiable assumptions [Ding and VanderWeele \(2016\)](#); ([G. W. Imbens & Rubin, 2015](#)). Sensitivity analysis is different from model assessment or model diagnostic, because identifying assumptions, such as unconfoundedness, are intrinsically untestable: the observed data are uninformative about the distribution of $Y(0)$ for treated units ($W = 1$) and about the distribution of $Y(1)$ for control units ($W = 0$). We like to argue that the larger the set of measured

covariates, the smaller the chance of unmeasured confounding bias, and it is thus often sensible to assume unconfoundedness. However, in practice, even if we have adjusted for all measured covariates, we can never rule out the possibility of an unmeasured confounder. A sensitivity analysis assumes the existence of at least one unmeasured confounder and posits assumptions on how it may relate to both treatment assignment and outcome. It then examines how the estimated causal effect varies as the degree of confounding bias, resulting from this unmeasured covariate, increases. Different strategies for sensitivity analysis have been proposed in the literature and include [Ding and VanderWeele \(2016\)](#), [Franks et al. \(2019\)](#), [Gustafson et al. \(2010\)](#), [Ichino et al. \(2008\)](#), [G. W. Imbens \(2003\)](#), [Liu et al. \(2013\)](#), [Rosenbaum \(1987b, 2002\)](#), [T. J. VanderWeele & Ding \(2017\)](#), and [Zhang and Tchetgen \(2019\)](#). Software has been developed to perform sensitivity analyses that assesses the strength of conclusions to violations of unverifiable assumptions ([Gang, 2004](#); [Keele, 2014](#); [Nannicini, 2007](#); [Ridgeway et al., 2004](#)).

One type of sensitivity analysis is the so-called *placebo* or *negative control* ([G. W. Imbens & Rubin, 2015](#)). Here the goal is to identify a variable that cannot be affected by the treatment. This variable is then used as an outcome (say Y^*) and the causal effect of W on Y^* is estimated, when we know that the true causal effect, say Δ , should be zero. If we estimate that Δ is statistically significantly different from zero, when we have adjusted for all the measured confounders, then we can conclude that there is unmeasured confounding bias. For additional details see [Schuemie et al. \(2020\)](#).

Sometimes we can check the quality of an observational control group by exploiting useful comparisons. For example, we can have access to two different pools of controls and use them both to check robustness of results to the use of either one ([Rosenbaum, 1987a](#)). We can also assess the plausibility of unconfoundedness and the performance of methods for causal effects in observational studies by checking the extent to which we can reproduce experimental findings using observational data. For example, [Dehejia and Wahba \(2002\)](#) use data from the [LaLonde \(1986\)](#) evaluation of nonexperimental methods that combine the treated units from a randomized evaluation of the National Supported Work (NSW) Demonstration, a labor training program, with nonexperimental comparison units drawn from survey data sets. They show that estimates obtained with propensity score based estimators of the treatment effect are closer to the experimental benchmark estimate than other methods.

In addition to the challenge of adjusting for confounding bias, there is another key challenge that is often overlooked, and that is due to overadjustment. For example, controlling for posttreatment variables (such as mediators) will bias the estimation of the causal effects; see, for example, [Ding and Miratrix \(2015\)](#), [Greenland et al. \(1999\)](#), [M. A. Hernán et al. \(2004\)](#), [D. B. Rubin \(2008\)](#), [Schisterman et al. \(2009\)](#), and [Shrier \(2008\)](#). In addition, it is common to include as many covariates as possible to control for confounding bias; however, *M-bias* is a problem that may arise when the correct estimation of treatment effects requires that certain variables are not adjusted for (i.e., are simply neglected from inclusion in the model). The topic of over adjustment is also broadly discussed in the causal graphical models literature (see [Pearl, 1995](#); [Perkovic et al., 2017](#); [Shpitser et al., 2010](#); [T. J. VanderWeele & Shpitser, 2011](#)).

8. Discussion

This article has focused on the potential outcome framework as one of the many approaches to define and estimate the causal effects of a given intervention on an outcome. We have presented:

- Thoughts regarding the central role of the study design when estimating causal effects (Table 1).
- Why machine learning is not a substitute for a thoughtful study design, and that it cannot overcome data quality, missing confounders, interference, or extrapolation.
- That machine learning algorithms can be very useful to estimate missing potential outcomes after issues related to study design have been resolved.
- Machine learning algorithms show great promise in discovering *de novo* subpopulations with heterogeneous causal effects.
- The importance of sensitivity analysis to assess how the conclusions are affected by deviations from identifying assumptions.

This review is focused on data science methods for *prospective* causal questions, that is, on assessing the causal effect of a particular, sometimes hypothetical, manipulation.

This is a different goal than that of causal discovery, which investigates the causal structure in the data, without starting with a specified causal model. To read more about causal discovery we refer you to [Glymour et al. \(2014\)](#), [Mooij et al. \(2016\)](#), and [Spirtes and Zhang \(2016\)](#).

We briefly outlined the Bayesian approach as one of the many statistical methods to estimate missing potential outcomes and the ATE. Alternative approaches such as those proposed by Fisher and Neyman are based on the randomization distribution of

statistics induced by a classical randomized assignment mechanism ([Fisher, 1937](#); [Neyman, 1923](#)) and their sampling distribution. The key feature of these approaches is that potential outcomes are treated as fixed but unknown, while the vector of treatment assignments, W , and the sampling indicators are the random variables. The concepts created by these methods, p -values, significance levels, unbiased estimation, confidence coverage, remain fundamental today ([D. Rubin, 2010](#)). However, we believe that the Bayesian thinking provides a straightforward approach to summarize the current state of knowledge in complex situations, to make informed decisions about which interventions look most promising for future application, and to properly quantify uncertainty around such decisions.

There are settings or instances where adjusting for measured covariates is not enough, that is, we cannot rule out dependence of the assignment mechanisms on the potential outcomes. In these *irregular* settings, another important area of research is relying on identification strategies that differ from strong ignorability, some of which are called quasi-experimental approaches. A natural experiment can be thought of as an observational study where treatment assignment, though not randomized, seems to resemble random assignment in that it is haphazard, not confounded by the typical attributes that determine treatment assignment in a particular empirical field ([Zubizarreta et al., 2014](#)). An example is instrumental variable (IV) methods ([Angrist et al., 1996](#)) where a variable, the instrument, plays the role of a randomly assigned incentive of treatment receipt and can be used to estimate treatment effect with a non-ignorable treatment assignment. For example, suppose we want to study the causal effect of an additional child on female labor supply; fertility decisions are typically *endogenous* and plausibly determined by observed and unobserved characteristics. [Angrist and Evans \(1998\)](#) used the sex of the first two children as an instrument for the decision to have a third child, and estimated the effect of having an additional child on a women's employment status. Such strategies are very popular in socioeconomic applications. In addition, other examples include regression discontinuity designs ([G. Imbens & Lemieux, 2008](#); [F. Li et al., 2015](#)), synthetic controls ([Abadie et al., 2010](#)), and their combinations ([Arkhangelsky et al., 2019](#); [Athey et al., 2018](#)). These designs typically focus on narrow effects (e.g., of compliers, of units at the threshold) with high internal validity, that need to be extrapolated to people or the population we are interested in. These methods can also be improved using machine learning ideas (for making IV stronger, or using more lagged values in a nonparametric way) but they require subject-matter knowledge and a level of creativity that, at least now, machine learning does not have.

Another area of active research is how to extend or generalize results from an RCT to a larger population ([Stuart et al., 2018](#)). [Hartman et al. \(2015\)](#) spells out the assumptions for this generalization ([Pearl & Bareinboim, 2011](#)) and proposes an approach to estimate effects for the larger population.

There are other key areas of causal inference not covered in this article. These include mediation analysis ([Huber, 2020](#); [T. VanderWeele, 2015](#)) and principal stratification ([Frangakis & Rubin, 2002](#); [Mealli & Mattei, 2012](#)), both of which provide understanding of causal mechanisms and causal pathways. We have also contributed to the literature in these areas (e.g., [Baccini et al., 2017](#); [Forastiere et al., 2016](#); [Mattei et al., 2013](#); [Mattei et al., 2020](#); [Mealli & Pacini, 2013](#); [Mealli et al., 2016](#)). These methods attempt to address questions such as: Why does having a dog reduce the level of severity of the symptoms of depression? Is it because they make you happier, or when you have a dog you are outside and exercising more, which helps depression, or that dogs help boost our immune systems? These questions are about exploring treatment effect heterogeneity with respect to post-treatment variables (e.g., spending more time outdoors as a consequence of walking the dog is a post-treatment variable). Questions regarding mediation require different tools from regression, matching, and machine learning prediction. In this new era of data science, where we have access to a deluge of data on pretreatment variables, complex treatments, post-treatment variables, and outcomes, this area of causality will become more prominent.

Can we envision a future where all these steps, characterizing the design and the analysis of observational data, and the unavoidable subject-matter knowledge, be translated into meta-data and automatized? Perhaps. But this is a challenge and there is still a lot of fascinating work ahead of us.

Disclosure Statement

The authors have no disclosures to share for this manuscript.

Acknowledgements

We sincerely thank Giorgio Gnecco, Kevin Josey, Leila Kamareddine, Alessandra Mattei and Xiao Wu and Lena Goodwin for providing feedback and editing the manuscript. Funding was provided by the Health Effects Institute (HEI) grant 4953-RFA14-3/16-4, National Institute of Health (NIH) grants R01, R01ES026217, R01MD012769, R01ES028033, 1R01AG060232-01A1, 1R01ES030616, 1R01AG066793, R01ES029950, the

2020 Starr Friedman Award, Sloan Foundation, EPA 83587201-0 and *Dipartimenti Eccellenti* 2018-2022 Italian Ministerial funds.

References

- Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105 (490), 493–505. <https://doi.org/10.1198/jasa.2009.ap08746>
- Abadie, A., & Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29 (1), 1–11. <https://doi.org/10.1198/jbes.2009.07333>
- An, W. (2010). Bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*, 40 (1), 151–189. <https://doi.org/10.1111/j.1467-9531.2010.01226.x>
- Angrist, J. D., & Evans, W. N. (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *The American Economic Review*, 88 (3), 450–477. <http://www.jstor.org/stable/116844>
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91 (434), 444–455. <https://doi.org/10.2307/2291629>
- Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press. <https://doi.org/10.2307/j.ctvc4j72>
- Antonelli, J., & Dominici, F. (2018). A Bayesian semiparametric framework for causal inference in high-dimensional data. *arXiv preprint arXiv:1805.04899*.
- Antonelli, J., Parmigiani, G., & Dominici, F. (2019). High dimensional confounding adjustment using continuous spike and slab priors. *Bayesian Analysis*, 14 (3), 805. <https://doi.org/10.1214/18-ba1131>
- Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. (2019). *Synthetic difference in differences* (tech. rep.). National Bureau of Economic Research. <https://doi.org/10.3386/w25532>

- Arpino, B., & Mattei, A. (2016). Assessing the causal effects of financial aids to firms in Tuscany allowing for interference. *The Annals of Applied Statistics*, 10 (3), 1170–1194. <https://doi.org/10.1214/15-AOAS902>
- Arvold, N., Wang, Y., Zigler, C., Schrag, D., & Dominici, F. (2014). Hospitalization burden and survival among elderly patients with glioblastoma. *Neuro-Oncology*, 16 (11), 1530–40. <https://doi.org/10.1093/neuonc/nou060>
- Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113 (27), 7353–7360. <https://doi.org/10.1073/pnas.1510489113>
- Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2018). *Matrix completion methods for causal panel data models* (tech. rep.). National Bureau of Economic Research. <https://doi.org/10.3386/w25132>
- Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. *Handbook of economic field experiments* (pp. 73–140). Elsevier.
- Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 685–725. <https://doi.org/10.1146/annurev-economics-080217-053433>
- Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *The Annals of Statistics*, 47 (2), 1148–1178. <https://doi.org/10.1214/18-AOS1709>
- Baccini, M., Mattei, A., & Mealli, F. (2017). Bayesian inference for causal mechanisms with application to a randomized study for postoperative pain control. *Biostatistics*, 18 (4), 605–617. <https://doi.org/10.1093/biostatistics/kxx010>
- Banerjee, A., Duflo, E., Glennerster, R., & Kinnan, C. (2015). The miracle of microfinance? Evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, 7 (1), 22–53. <https://doi.org/10.1257/app.20130533>
- Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4), 962–973. <https://doi.org/10.1111/j.1541-0420.2005.00377.x>
- Bargagli-Stoffi, F. J., De-Witte, K., & Gnecco, G. (2019). Heterogeneous causal effects with imperfect compliance: A novel Bayesian machine learning approach. *arXiv preprint arXiv:1905.12707v3*.

Bargagli-Stoffi, F. J., & Gnecco, G. (2020). Causal tree with instrumental variable: An extension of the causal tree framework to irregular assignment mechanisms.

International Journal of Data Science and Analytics, 9 (3), 315–337.

<https://doi.org/10.1007/s41060-019-00187-z>

Bargagli-Stoffi, F. J., Tortu, C., & Forastiere, L. (2020). Heterogeneous treatment and spillover effects under clustered network interference. *arXiv preprint arXiv:2008.00707v2*.

Beebe, H., Hitchcock, C., & Menzies, P. (2009). *The oxford handbook of causation*. Oxford University Press UK.

<https://doi.org/10.1093/oxfordhb/9780199279739.001.0001>

Belloni, A., Chernozhukov, V., & Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28 (2), 29–50. <https://doi.org/10.1257/jep.28.2.29>

Belloni, A., Chernozhukov, V., & Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81 (2), 608–650. <https://doi.org/10.1093/restud/rdt044>

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth; Brooks. <https://cds.cern.ch/record/2253780>

Breiman, L. (2001). Random forests. *Machine Learning*, 45 (1), 5–32.

Candes, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9 (6), 717–763.

<https://doi.org/10.1007/s10208-009-9045-5>

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*.

Cambridge University Press. <https://doi.org/10.1017/CBO9780511618758>

Cefalu, M., Dominici, F., Arvold, N., & Parmigiani, G. (2016). Model averaged double robust estimation. *Biometrics*, 73 (2), 410–421. <https://doi.org/10.1111/biom.12622>

Chang, C., Kundu, S., & Long, Q. (2018). Scalable Bayesian variable selection for structured high- dimensional data. *Biometrics*, 74 (4), 1372–1382.

<https://doi.org/10.1111/biom.12882>

- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, 107 (5), 261–265. <https://doi.org/10.1257/aer.p20171038>
- 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 (1), 1–68. <https://doi.org/10.1111/ectj.12097>
- Chib, S., & Greenberg, E. (2010). Additive cubic spline regression with Dirichlet process mixture errors. *Journal of Econometrics*, 156 (2), 322–336. <https://doi.org/10.1016/j.jeconom.2009.11.002>
- Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4 (1), 266–298. <https://doi.org/10.1214/09-AOAS285>
- Collins, R., Bowman, L., Landray, M., & Peto, R. (2020). The magic of randomization versus the myth of real-world evidence. *New England Journal of Medicine*, 382 (7), 674–678. <https://doi.org/10.1056/NEJMsb1901642>
- Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine learning*, 20 (3), 273–297. <https://doi.org/10.1023/A:1022627411411>
- Dahabreh, I. J., Hayward, R., & Kent, D. M. (2016). Using group data to treat individuals: Understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *International Journal of Epidemiology*, 45 (6), 2184–2193. <https://doi.org/10.1093/ije/dyw125>
- D’Amour, A., Deng, P., Feller, A., Lei, L., & Sekhon, J. (2017). Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*.
- Dawid, A. P., Musio, M., & Murtas, R. (2017). The probability of causation. *Law, Probability and Risk*, 16 (4), 163–179. <https://doi.org/10.1093/lpr/mgx012>
- de Finetti, B. (1963). Foresight: Its logical laws, its subjective sources. In H. Kyburg & H. Smokler (Eds.), *Studies in subjective probability*. Wiley. https://doi.org/10.1007/978-1-4612-0919-5_10
- Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84 (1), 151–161. <https://doi.org/10.1162/003465302317331982>

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95 (3), 932–945.

https://doi.org/10.1162/rest_a_00318

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems, LCBS- 1857*, 1–15. https://doi.org/10.1007/3-540-45014-9_1

Ding, P., & Li, F. (2018). Causal inference: A missing data perspective. *Statistical Science*, 33 (2), 214–237. <https://doi.org/10.1111/rssb.12124>

Ding, P., & Li, F. (2019). A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis*, 27 (4), 605–615.

<https://doi.org/10.1017/pan.2019.25>

Ding, P., & Miratrix, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of m -bias and butterfly-bias. *Journal of Causal Inference*, 3 (1), 41–57.

<https://doi.org/10.1515/jci-2013-0021>

Ding, P., & VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology*, 27 (3), 368–377. <https://doi.org/10.1097/EDE.0000000000000457>

Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D., et al. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34 (1), 43–68. <https://doi.org/10.1214/18-STS667>

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, 4 (1), 3895–3962. [https://doi.org/10.1016/S1573-4471\(07\)04061-2](https://doi.org/10.1016/S1573-4471(07)04061-2)

Farrell, M. H., Liang, T., & Misra, S. (2018). Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953v3*.

Fisher, R. A. (1918). The causes of human variability. *Eugenics Review*, 10 (1), 213–220. Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

Fisher, R. A. (1937). *The design of experiments*. Oliver; Boyd.

Forastiere, L., Airoidi, E. M., & Mealli, F. (2020). Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, *116* (534), 901–918.

<https://doi.org/10.1080/01621459.2020.1768100>

Forastiere, L., Mattei, A., & Ding, P. (2018). Principal ignorability in mediation analysis: Through and beyond sequential ignorability. *Biometrika*, *105* (4), 979–986.

<https://doi.org/10.1093/biomet/asy053>

Forastiere, L., Mealli, F., & VanderWeele, T. J. (2016). Identification and estimation of causal mechanisms in clustered encouragement designs: Disentangling bed nets using Bayesian principal stratification. *Journal of the American Statistical Association*, *111* (514), 510–525. <https://doi.org/10.1080/01621459.2015.1125788>

Foster, J. C., Taylor, J. M., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, *30* (24), 2867–2880.

<https://doi.org/10.1002/sim.4322>

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference.

Biometrics, *58* (1), 21–29. <https://doi.org/10.1111/j.0006-341x.2002.00021.x>

Franks, A., D’Amour, A., & Feller, A. (2019). Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, *115* (532), 1730–1746.

<https://doi.org/10.1080/01621459.2019.1604369>

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine.

The Annals of Statistics, *29* (9), 1189–1232. <https://doi.org/10.1214/aos/1013203451>

Funk, J. M., & Landi, S. (2014). Misclassification in administrative claims data: Quantifying the impact on treatment effect estimates. *Current Epidemiology Reports*, *1* (4), 175–185.

<https://doi.org/10.1007/s40471-014-0027-z>

Funk, J. M., Westreich, D., Wiesen, C., Sturmer, T., Brookhart, M., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, *173* (7), 761–767.

<https://doi.org/10.1093/aje/kwq439>

Gang, M. (2004). RBOUNDS: Stata module to perform Rosenbaum sensitivity analysis for average treatment effects on the treated.

- Gao, Z., Hastie, T., & Tibshirani, R. (2020). Assessment of heterogeneous treatment effect estimation accuracy via matching. *arXiv preprint arXiv:2003.03881v1*.
- Glymour, C., Scheines, R., & Spirtes, P. (2014). *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. Academic Press. <https://doi.org/10.2307/1164612>
- Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 37–48. <http://www.jstor.org/stable/3702180>
- Gustafson, P. (2012). Double-robust estimators: Slightly more Bayesian than meets the eye? *The International Journal of Biostatistics*, 8 (2), 1. <https://doi.org/10.2202/1557-4679.1349>
- Gustafson, P., McCandless, L. C., Levy, A. R., & Richardson, S. (2010). Simplified Bayesian sensitivity analysis for mismeasured and unobserved confounders. *Biometrics*, 66 (4), 1129–1137. <https://doi.org/10.1111/j.1541-0420.2009.01377.x>
- Gutman, R., & Rubin, D. B. (2015). Estimation of causal effects of binary treatments in unconfounded studies. *Statistics in Medicine*, 34 (26), 3381–3398. <https://doi.org/10.1002/sim.6532>
- Hahn, P. R., Dorie, V., & Murray, J. S. (2019). Atlantic Causal Inference Conference (ACIC) data analysis challenge 2017. *arXiv preprint arXiv:1905.09515v1*.
- Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15 (3), 965–1056. <https://doi.org/10.1214/19-BA1195>
- Halpern, J. Y. (2016). *Actual causality*. MIT Press. <https://muse.jhu.edu/book/47917>
- Hartman, E., Grieve, R., Ramsahai, R., & Sekhon, J. S. (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 (Statistics in Society)*, 178 (3), 757–778. <https://doi.org/10.1111/rssa.12094>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

- Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The LASSO and generalizations*. CRC press. <https://doi.org/10.1201/b18401>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Hernán, M. D., & Robins, J. M. (2020). *Causal inference: What if*. Boca Raton: Chapman & Hall/CRC. <https://www.hsph.harvard.edu/miguel-herman/causal-inference-book/>
- Hernán, M. A. (2016). Does water kill? A call for less casual causal inferences. *Annals of epidemiology*, 26 (10), 674–680. <https://doi.org/10.1016/j.annepidem.2016.08.016>
- Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 615–625. <https://doi.org/10.1097/01.ede.0000135174.63482.43>
- Hernán, M. A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *Chance*, 32 (1), 42–49. <https://doi.org/10.1080/09332480.2019.1579578>
- Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183 (8), 758–764. <https://doi.org/10.1093/aje/kwv254>
- Hill, J. (2004). Reducing bias in treatment effect estimation in observational studies suffering from missing data. *ISERP Working Papers*, 04-01. <https://doi.org/10.7916/D8B85G11>
- Hill, J., & Su, Y. S. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children’s cognitive outcomes. *The Annals of Applied Statistics*, 1386–1420. <https://doi.org/10.1214/13-AOAS630>
- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20 (1), 217–240. <https://doi.org/10.1198/jcgs.2010.08162>
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political*

Analysis, 15 (3), 199–236. <https://doi.org/10.1093/pan/mpl013>

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81 (396), 945–960. <https://doi.org/10.2307/2289064>

Huber, M. (2020). Mediation analysis. *Handbook of Labor, Human Resources and Population Economics*, 1–38. https://doi.org/10.1007/978-3-319-57365-6_162-2

Ichino, A., Mealli, F., & Nannicini, T. (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics*, 23 (3), 305–327. <https://doi.org/10.1002/jae.998>

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76 (1), 243–263. <https://doi.org/10.1111/rssb.12027>

Imai, K., & Yamamoto, T. (2010). Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis. *American Journal of Political Science*, 54 (2), 543–560. <https://doi.org/10.1111/j.1540-5907.2010.00446.x>

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86 (1), 4–29. <https://doi.org/10.1162/003465304323023651>

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139025751>

Imbens, G., & Lemieux, T. (2008). The regression discontinuity design—Theory and applications. *Journal of Econometrics*, 142 (2), 611–850. <https://doi.org/10.1016/j.jeconom.2007.05.008>

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93 (2), 126–132. <https://doi.org/10.1257/000282803321946921>

Imbens, G., & Rubin, D. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25 (1), 305–327. <https://doi.org/10.1214/aos/1034276631>

- Jeong, S., & Rockova, V. (2020). The art of BART: On flexibility of Bayesian forests. *arXiv preprint arXiv:2008.06620v2*.
- Keele, L. J. (2014). Rbounds: An R package for sensitivity analysis with matched data. <https://cran.r-project.org/web/packages/rbounds/index.html>
- Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. *Advances in Neural Information Processing Systems*, 2280–2288. <https://doi.org/10.5555/3157096.3157352>
- Knaus, M. C. (2020). Double machine learning based program evaluation under unconfoundedness. *arXiv preprint arXiv:2003.03191*.
- Knaus, M. C., Lechner, M., & Strittmatter, A. (2020). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal*, 24. <https://doi.org/10.1093/ectj/utaa014>
- LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76 (4), 604–620. <http://www.jstor.org/stable/1806062>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521 (7553), 436–444. <https://doi.org/10.1038/nature14539>
- Lee, K., Bargagli-Stoffi, F. J., & Dominici, F. (2020). Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. *arXiv preprint arXiv:2009.09036*.
- Li, F., Mattei, A., & Mealli, F. (2015). Evaluating the causal effect of university grants on student dropout: Evidence from a regression discontinuity design using principal stratification. *The Annals of Applied Statistics*, 9 (4), 1906–1931. <https://doi.org/10.1214/15-AOAS881>
- Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113 (521), 390–400. <https://doi.org/10.1080/01621459.2016.1260466>
- Li, F., & Thomas, L. E. (2018). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, 188 (1), 250–257. <https://doi.org/10.1093/aje/kwy201>

- Li, J., Ma, S., Liu, L., Le, T. D., Liu, J., & Han, Y. (2019). Identify treatment effect patterns for personalised decisions. *arXiv preprint arXiv:1906.06080v1*.
- Liao, S., & Zigler, C. (2020). Uncertainty in the design stage of two-stage Bayesian propensity score analysis. *Statistics in Medicine*, *39* (17), 2265–2290. <https://doi.org/10.1002/sim.8486>
- Liu, W., Kuramoto, S. J., & Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention science: The official journal of the Society for Prevention Research*, *14* (6), 570–580. <https://doi.org/10.1007/s11121-012-0339-5>
- Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23* (19), 2937–2960. <https://doi.org/10.1002/sim.1903>
- Mattei, A., & Mealli, F. (2007). Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination. *Biometrics*, *63* (2), 437–446. <https://doi.org/10.1111/j.1541-0420.2006.00684.x>
- Mattei, A., Li, F., & Mealli, F. (2013). Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program. *The Annals of Applied Statistics*, *7* (4), 2336–2360. <https://doi.org/10.1214/13-AOAS674>
- Mattei, A., & Mealli, F. (2011). Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73* (5), 729–752. <https://doi.org/10.1111/j.1467-9868.2011.00780.x>
- Mattei, A., & Mealli, F. (2015). Discussion of “On Bayesian estimation of marginal structural models”. *Biometrics*, *71* (2), 293–296. <https://doi.org/10.1111/biom.12272>
- Mattei, A., Mealli, F., & Ding, P. (2020). Assessing causal effects in the presence of treatment switching through principal stratification. *arXiv preprint arXiv:2002.11989*.
- Mattei, A., Mealli, F., & Pacini, B. (2014). Identification of causal effects in the presence of non-ignorable missing outcome values. *Biometrics*, *70* (2), 278–288. <https://doi.org/10.1111/biom.12136>
- McCandless, L., Gustafson, P., & Austin, P. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, *15* (1), 94–112.

<https://doi.org/10.1002/sim.3460>

Mealli, F., & Mattei, A. (2012). A refreshing account of principal stratification. *The International Journal of Biostatistics*, 8 (1), 1–17. <https://doi.org/10.1515/1557-4679.1380>

Mealli, F., Pacini, B., & Rubin, D. B. (2011). Statistical inference for causal effects. *Modern analysis of customer satisfaction surveys*. Wiley. <https://doi.org/10.1002/9781119961154>

Mealli, F., & Pacini, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, 108 (503), 1120–1131. <https://doi.org/10.1080/01621459.2013.802238>

Mealli, F., Pacini, B., & Stanghellini, E. (2016). Identification of principal causal effects using additional outcomes in concentration graphs. *Journal of Educational and Behavioral Statistics*, 41 (5), 463–480. <https://doi.org/10.3102/1076998616646199>

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267 (1), 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., & Schölkopf, B. (2016). Distinguishing cause from effect using observational data: Methods and benchmarks. *The Journal of Machine Learning Research*, 17 (1), 1103–1204.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511804564>

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9781107587991>

Nannicini, T. (2007). A simulation-based sensitivity analysis for matching estimators. *Stata Journal*, 7 (3), 334–350. <https://doi.org/10.1177/1536867X0700700303>

Nethery, R. C., Mealli, F., & Dominici, F. (2019). Estimating population average causal effects in the presence of non-overlap: A Bayesian approach. *The Annals of Applied Statistics*, 13 (2), 1242–1267. <https://doi.org/10.1214/18-AOAS1231>

- Neyman, J. (1923). On the application of probability theory to agricultural experiments, section 9 [reprinted in *Statistical Science*, 1990, 5, 463–485]. *Roczniki Nauk Rolniczych*, 10, 1–51. <https://doi.org/10.1214/ss/1177012031>
- Papadogeorgou, G., Mealli, F., & Zigler, C. M. (2019). Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, 75(3), 778–787. <https://doi.org/10.1111/biom.13049>
- Pearl, J., & Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. *AAAI Conference on Artificial Intelligence, Proceedings of the Twentieth Conference*, 247–254. <https://doi.org/10.1109/ICDMW.2011.169>
- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4), 669–688. <https://doi.org/10.1093/biomet/82.4.669>
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press. <https://doi.org/10.1017/S0266466603004109>
- Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.
- Perkovic, E., Textor, J., Kalisch, M., & Maathuis, M. H. (2017). Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *The Journal of Machine Learning Research*, 18(1), 8132–8193.
- Peters, J., Janzing, D., & Schoelkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. MIT Press. <https://library.oapen.org/handle/20.500.12657/26040>
- Pierce, B. L., & VanderWeele, T. J. (2012). The effect of non-differential measurement error on bias, precision and power in Mendelian randomization studies. *International Journal of Epidemiology*, 41(5), 1383–1393. <https://doi.org/10.1093/ije/dys141>
- Recht, B. (2011). A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(104), 3413–3430. <https://doi.org/10.5555/1953048.2185803>
- Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., & Burgette, L. (2004). Twang: Toolkit for weighting and analysis of nonequivalent groups. <https://cran.r-project.org/web/packages/twang/index.html>

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999*, 6–10.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90 (429), 122–129. <https://www.jstor.org/stable/2291135>

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90 (429), 106–121. <http://www.jstor.org/stable/2291134>

Rosenbaum, P. R. (1987a). The role of a second control group in an observational study (with discussion). *Statistical Science*, 2 (3), 292–316. <http://www.jstor.org/stable/2245766>

Rosenbaum, P. R. (1987b). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74 (1), 13–26. <https://doi.org/10.1093/biomet/74.1.13>

Rosenbaum, P. R. (2002). *Observational studies*. New York: Springer. <https://doi.org/10.1007/978-1-4757-2443-1>

Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer. <https://doi.org/10.1007/978-1-4419-1213-8>

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70 (1), 41–55. <https://doi.org/10.1093/biomet/70.1.41>

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66 (5), 688–701. <https://doi.org/10.1037/h0037350>

Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. *The Proceedings of the Social Statistics Section of the American Statistical Association*, 233, 239. <http://www.jstor.org/stable/2958688>

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63 (3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>

- Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2 (1), 1–26. <https://doi.org/10.3102/10769986002001001>
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6 (1), 34–58. <https://doi.org/10.1214/aos/1176344064>
- Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75 (371), 591–593. <https://doi.org/10.2307/2287653>
- Rubin, D. B. (1985). The use of propensity scores in applied Bayesian inference. *Bayesian Statistics 2*, 463–472.
- Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81 (396), 961–962. <https://doi.org/10.1080/01621459.1986.10478355>
- Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25 (3), 279–292. [https://doi.org/10.1016/0378-3758\(90\)90077-8](https://doi.org/10.1016/0378-3758(90)90077-8)
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100 (469), 322–331.
- Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with “censoring” due to death. *Statistical Science*, 21 (3), 299–309. <https://doi.org/10.1214/088342306000000114>
- Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26 (1), 20–36. <https://doi.org/10.1002/sim.2739>
- Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2 (3), 808–840. <https://doi.org/10.1214/08-AOAS187>
- Rubin, D. B. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods*, 15 (1), 38–46. <https://doi.org/10.1037/a0018537>
- Saarela, O., Stephens, D. A., Moodie, E. E. M., & Klein, M. B. (2015). On Bayesian estimation of marginal structural models (with discussion). *Biometrics*, 71 (2), 279–

301. <https://doi.org/10.1111/biom.12269>

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94 (448), 1096–1120.

<https://doi.org/10.1080/01621459.1999.10473862>

Schisterman, E. F., Cole, S. R., & Platt, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology (Cambridge, Mass.)*, 20 (4), 488. <https://doi.org/10.1097/EDE.0b013e3181a819a1>

Schuemie, M. J., Cepeda, M. S., Suchard, M. A., Yang, J., Schuler, Y. T. A., Ryan, P. B., Madigan, D., & Hripcsak, G. (2020). How confident are we about observational findings in health care: A benchmark study. *Harvard Data Science Review*, 2 (1).

<https://doi.org/10.1162/99608f92.147cc28e>

Shpitser, I., VanderWeele, T., & Robins, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, 527–536.

Shrier, I. (2008). Letter to the editor: Propensity scores letter to the editor. *Statistics in Medicine*, 27, 2740–41. <https://doi.org/10.1002/sim.3554>

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Spirtes, P., Glymour, C., & R, S. (2001). *Causation, prediction, and search*, 2nd ed. MIT Press, Cambridge.

Spirtes, P., & Zhang, K. (2016). Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, 3 (1), 3. <https://doi.org/10.1186/s40535-016-0018-x>

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25 (1), 1–21. <https://doi.org/10.1214/09-STS313>

Stuart, E. A., Ackerman, B., & Westreich, D. (2018). Generalizability of randomized trial results to target populations: Design and analysis possibilities. *Research on Social Work Practice*, 28 (5), 532–537. <https://doi.org/10.1177/1049731517720730>

- Stuart, E. A., & Rubin, D. B. (2008). Best practices in quasi-experimental designs: Matching methods for causal inference. *Best practices in quantitative methods* (pp. 155–176). SAGE Publications, Inc., Thousand Oaks, CA. <https://doi.org/10.1.1.584.1057>
- Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9 (3), 293–300. <https://doi.org/10.1023/A:1018628609742>
- Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58 (1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Tortú, C., Forastiere, L., Crimaldi, I., & Mealli, F. (2020). Modelling network interference with multi-valued treatments: The causal effect of immigration policy on crime rates. *arXiv preprint arXiv:2003.10525*.
- Van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data*. Springer Science & Business Media. <https://doi.org/10.1007/978-1-4419-9782-1>
- VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press. <https://doi.org/10.1093/ije/dyw277>
- VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the E-Value. *Annals of Internal Medicine*, 167 (4), 268–274. <https://doi.org/10.7326/M16-2607>
- VanderWeele, T. J., & Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, 67 (4), 1406–1413. <https://doi.org/10.1111/j.1541-0420.2011.01619.x>
- VanderWeele, T. J., & Shpitser, I. (2013). On the definition of a confounder. *The Annals of Statistics*, 41 (1), 196. <https://doi.org/10.1214/12-AOS1058>
- Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113 (523), 1228–1242. <https://doi.org/10.1080/01621459.2017.1319839>
- Wang, C., Dominici, F., Parmigiani, G., & Zigler, C. (2015). Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models: Accounting for uncertainty in confounder and effect

modifier selection when estimating aces in GLMs. *Biometrics*, 71 (3), 654–665.

<https://doi.org/10.1111/biom.12315>

Wang, C., Parmigiani, G., & Dominici, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68 (3), 661–671. <https://doi.org/10.1111/j.1541-0420.2011.01731.x>

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., & Zhang, A. (2020). A survey on causal inference. *arXiv preprint arXiv:2002.02770*.

Zhang, B., & Tchetgen, E. J. T. (2019). A semiparametric approach to model-based sensitivity analysis in observational studies. *arXiv preprint arXiv:1910.14130*.

Zheng, H., & Little, R. J. A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21 (1), 1–20.

Zigler, C., & Cefalu, M. (2017). Posterior predictive treatment assignment for estimating causal effects with limited overlap. *arXiv preprint arXiv:1710.0874*.

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., & Dominici, F. (2013). Model feedback in Bayesian propensity score estimation. *Biometrics*, 69 (1), 263–273.

<https://doi.org/10.1111/j.1541-0420.2012.01830.x>

Zigler, C. M., & Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, 109 (505), 95–107.

<https://doi.org/10.1080/01621459.2013.869498>

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107 (500), 1360–1371. <https://doi.org/10.1080/01621459.2012.703874>

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110 (511), 910–922. <https://doi.org/10.1080/01621459.2015.1023805>

Zubizarreta, J. R., Paredes, R. D., Rosenbaum, P. R., et al. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *The Annals of Applied Statistics*, 8 (1), 204–231.

<https://doi.org/10.1214/13-AOAS713>

Zubizarreta, J. R., Small, D. S., & Rosenbaum, P. R. (2014). Isolation in the construction of natural experiments. *The Annals of Applied Statistics*, 8 (4), 2096–2121.
<https://doi.org/10.1214/14-AOAS770>

This article is © 2021 by the author(s). The editorial is licensed under a Creative Commons Attribution (CC BY 4.0) International license (<https://creativecommons.org/licenses/by/4.0/legalcode>), except where otherwise indicated with respect to particular material included in the article. The article should be attributed to the authors identified above.

Footnotes

1. A formal definition of a confounder is offered in Section 3.2, after the introduction of notation, postulates, and assumptions. [↵](#)
2. Please refer to Section 7 in D. B. Rubin (1990) for a specific example, and G. Imbens and Rubin (1997) for the dependence of the posterior distribution of causal effects on association parameters in the joint distribution of $Y(1)$ and $Y(0)$. The discussion is beyond the scope of this paper. [↵](#)

Citations

1. Neyman, J. (1923). On the application of probability theory to agricultural experiments, section 9. *Roczniki Nauk Rolniczych*, X, 1–51.
<https://doi.org/10.1214/ss/1177012031> [↵](#)
2. Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.
<https://doi.org/https://doi.org/10.1037/h0037350> [↵](#)
3. Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1), 34–58.
<https://doi.org/10.1214/aos/1176344064> [↵](#)
4. Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.
<https://doi.org/10.2307/j.ctvc4j72> [↵](#)

5. Hernán, M. D., & Robins, J. M. (2020). *Causal inference: What if*. Boca Raton: Chapman & Hall/CRC. <https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/> ↵
6. Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press. <https://doi.org/https://doi.org/10.1017/CBO9781139025751> ↵
7. Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511804564> ↵
8. Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9781107587991> ↵
9. Rosenbaum, P. R. (2002). *Observational studies*. New York: Springer. <https://doi.org/10.1007/978-1-4757-2443-1> ↵
10. Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer. <https://doi.org/10.1007/978-1-4419-1213-8> ↵
11. Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press. <https://doi.org/10.1017/S0266466603004109> ↵
12. Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books. ↵
13. Peters, J., Janzing, D., & Schoelkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. MIT press. <http://library.oapen.org/handle/20.500.12657/26040> ↵
14. Beebe, H., Hitchcock, C., & Menzies, P. (2009). *The oxford handbook of causation*. Oxford University Press UK. <https://doi.org/10.1093/oxfordhb/9780199279739.001.0001> ↵
15. Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511618758> ↵
16. Halpern, J. Y. (2016). *Actual causality*. MIT Press. muse.jhu.edu/book/47917 ↵

17. Dawid, A. P., Musio, M., & Murtas, R. (2017). The probability of causation. *Law, Probability and Risk*, 16(4), 163–179. <https://doi.org/10.1093/lpr/mgx012> ↵
18. VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press. <https://doi.org/10.1093/ije/dyw277> ↵
19. Glymour, C., Scheines, R., & Spirtes, P. (2014). *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. Academic Press. <https://doi.org/10.2307/1164612> ↵
20. Spirtes, P., Glymour, C., & R, S. (2001). *Causation, prediction, and search, 2nd edn*. MIT Press, Cambridge. ↵
21. Collins, R., Bowman, L., Landray, M., & Peto, R. (2020). The magic of randomization versus the myth of real-world evidence. *New England Journal of Medicine*, 382(7), 674–678. <https://doi.org/10.1056/NEJMSb1901642> ↵
22. Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments* (Vol. 1, pp. 73–140). Elsevier. ↵
23. Banerjee, A., Duflo, E., Glennerster, R., & Kinnan, C. (2015). The miracle of microfinance? Evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, 7(1), 22–53. <https://doi.org/10.1257/app.20130533> ↵
24. Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, 4(1), 3895–3962. [https://doi.org/10.1016/S1573-4471\(07\)04061-2](https://doi.org/10.1016/S1573-4471(07)04061-2) ↵
25. Arvold, N., Wang, Y., Zigler, C., Schrag, D., & Dominici, F. (2014). Hospitalization burden and survival among elderly patients with glioblastoma. *Neuro-Oncology*, 16(11), 1530–1540. <https://doi.org/10.1093/neuonc/nou060> ↵
26. Fisher, R. A. (1918). The causes of human variability. *Eugenics Review*, 10(1), 213–220. ↵
27. Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd. ↵
28. Rubin, D. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. <https://doi.org/https://doi.org/10.1093/biomet/63.3.581> ↵

29. Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1), 1-26. <https://doi.org/https://doi.org/10.3102/10769986002001001> [↵](#)
30. Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25(3), 279-292. [https://doi.org/https://doi.org/10.1016/0378-3758\(90\)90077-8](https://doi.org/https://doi.org/10.1016/0378-3758(90)90077-8) [↵](#)
31. Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945-960. <https://doi.org/10.2307/2289064> [↵](#)
32. Rubin, D. B., & others. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3), 808-840. <https://doi.org/10.1214/08-AOAS187> [↵](#)
33. Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8), 758-764. <https://doi.org/10.1093/aje/kwv254> [↵](#)
34. Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26(1), 20-36. <https://doi.org/https://doi.org/10.1002/sim.2739> [↵](#)
35. Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469), 322-331. [↵](#)
36. Funk, J. M., Westreich, D., Wiesen, C., Sturmer, T., Brookhart, M., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7), 761-767. <https://doi.org/10.1093/aje/kwq439> [↵](#)
37. Dahabreh, I. J., Hayward, R., & Kent, D. M. (2016). Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *International Journal of Epidemiology*, 45(6), 2184-2193. <https://doi.org/10.1093/ije/dyw125> [↵](#)
38. Li, J., Ma, S., Liu, L., Le, T. D., Liu, J., & Han, Y. (2019). Identify treatment effect patterns for personalised decisions. *ArXiv Preprint ArXiv:1906.06080v1*. [↵](#)
39. Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371), 591-593. <https://doi.org/https://doi.org/10.2307/2287653> [↵](#)

40. Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396), 961–962.
<https://doi.org/https://doi.org/10.1080/01621459.1986.10478355> ↵
41. Forastiere, L., Airoidi, E. M., & Mealli, F. (2020). Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 1–18.
<https://doi.org/10.1080/01621459.2020.1768100> ↵
42. Papadogeorgou, G., Mealli, F., & Zigler, C. M. (2019). Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, 75(3), 778–787. <https://doi.org/10.1111/biom.13049> ↵
43. Hernán, M. A. (2016). Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*, 26(10), 674–680.
<https://doi.org/10.1016/j.annepidem.2016.08.016> ↵
44. VanderWeele, T. J., & Shpitser, I. (2013). On the definition of a confounder. *The Annals of Statistics*, 41(1), 196. <https://doi.org/10.1214/12-AOS1058> ↵
45. Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
<https://doi.org/https://doi.org/10.1093/biomet/70.1.41> ↵
46. Nethery, R. C., Mealli, F., & Dominici, F. (2019). Estimating population average causal effects in the presence of non-overlap: A Bayesian approach. *The Annals of Applied Statistics*, 13(2), 1242–1267. <https://doi.org/10.1214/18-AOAS1231> ↵
47. Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3), 199–236. <https://doi.org/10.1093/pan/mp1013> ↵
48. Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1), 1–21. <https://doi.org/10.1214/09-STS313> ↵
49. Stuart, E. A., & Rubin, D. B. (2008). Best practices in quasi-experimental designs: Matching methods for causal inference. In *Best practices in quantitative methods* (Vol. 18, pp. 155–176). SAGE Publications, Inc., Thousand Oaks, CA.
<https://doi.org/10.1.1.584.1057> ↵

50. Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 243–263. <https://doi.org/10.1111/rssb.12027> [↵](#)
51. Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429), 122–129. <http://www.jstor.org/stable/2291135> [↵](#)
52. Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429), 106–121. <http://www.jstor.org/stable/2291134> [↵](#)
53. Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3), 932–945. https://doi.org/10.1162/REST_a_00318 [↵](#)
54. Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521), 390–400. <https://doi.org/10.1080/01621459.2016.1260466> [↵](#)
55. Li, F., & Thomas, L. E. (2018). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, 188(1), 250–257. <https://doi.org/10.1093/aje/kwy201> [↵](#)
56. Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500), 1360–1371. <https://doi.org/https://doi.org/10.1080/01621459.2012.703874> [↵](#)
57. Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511), 910–922. <https://doi.org/https://doi.org/10.1080/01621459.2015.1023805> [↵](#)
58. Zubizarreta, J. R., Paredes, R. D., Rosenbaum, P. R., & others. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *The Annals of Applied Statistics*, 8(1), 204–231. <https://doi.org/10.1214/13-AOAS713> [↵](#)

59. Mattei, A., & Mealli, F. (2015). Discussion of “On Bayesian estimation of marginal structural models.” *Biometrics*, 71(2), 293–296. <https://doi.org/10.1111/biom.12272> ↵
60. Ding, P., & Li, F. (2018). Causal inference: A missing data perspective. *Statistical Science*, 33(2), 214–237. <https://doi.org/10.1111/rssb.12124> ↵
61. Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. *The Proceedings of the Social Statistics Section of the American Statistical Association*, 233, 239. <http://www.jstor.org/stable/2958688> ↵
62. Mealli, F., Pacini, B., & Rubin, D. B. (2011). Statistical inference for causal effects. In *Modern analysis of customer satisfaction surveys*. Wiley. <https://doi.org/10.1002/9781119961154> ↵
63. Finetti, de B. (1963). Foresight: Its logical laws, its subjective sources. In H. Kyburg & H. Smokler (Eds.), *Studies in subjective probability*. Wiley. https://doi.org/10.1007/978-1-4612-0919-5_10 ↵
64. Rubin, D. B. (1985). The use of propensity scores in applied Bayesian inference. *Bayesian Statistics 2*, 463–472. ↵
65. Ding, P., & Li, F. (2019). A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis*, 27(4), 605–615. <https://doi.org/10.1017/pan.2019.25> ↵
66. Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15(3), 965–1056. <https://doi.org/10.1214/19-BA1195> ↵
67. Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 217–240. <https://doi.org/10.1198/jcgs.2010.08162> ↵
68. Chib, S., & Greenberg, E. (2010). Additive cubic spline regression with dirichlet process mixture errors. *Journal of Econometrics*, 156(2), 322–336. <https://doi.org/10.1016/j.jeconom.2009.11.002> ↵
69. Gutman, R., & Rubin, D. B. (2015). Estimation of causal effects of binary treatments in unconfounded studies. *Statistics in Medicine*, 34(26), 3381–3398. <https://doi.org/10.1002/sim.6532> ↵

70. Zheng, H., & Little, R. J. A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21(1), 1-20. [↵](#)
71. McCandless, L., Gustafson, P., & Austin, P. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, 15(1), 94-112. <https://doi.org/10.1002/sim.3460> [↵](#)
72. Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., & Dominici, F. (2013). Model feedback in Bayesian propensity score estimation. *Biometrics*, 69(1), 263-273. <https://doi.org/10.1111/j.1541-0420.2012.01830.x> [↵](#)
73. Liao, S., & Zigler, C. (2020). Uncertainty in the design stage of two-stage Bayesian propensity score analysis. *Statistics in Medicine*, 39(17), 2265-2290. <https://doi.org/10.1002/sim.8486> [↵](#)
74. Hill, J., & Su, Y. S. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, 1386-1420. <https://doi.org/10.1214/13-AOAS630> [↵](#)
75. Zigler, C. M., & Cefalu, M. (2017). Posterior predictive treatment assignment for estimating causal effects with limited overlap. *ArXiv Preprint ArXiv:1710.0874*. [↵](#)
76. Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1), 4-29. <https://doi.org/10.1162/003465304323023651> [↵](#)
77. Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., & Zhang, A. (2020). A survey on causal inference. *ArXiv Preprint ArXiv:2002.02770*. [↵](#)
78. An, W. (2010). Bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*, 40(1), 151-189. <https://doi.org/10.1111/j.1467-9531.2010.01226.x> [↵](#)
79. Saarela, O., Stephens, D. A., Moodie, E. E. M., & Klein, M. B. (2015). On Bayesian estimation of marginal structural models (with discussion). *Biometrics*, 71(2), 279-301. <https://doi.org/10.1111/biom.12269> [↵](#)
80. Abadie, A., & Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1), 1-11.

<https://doi.org/10.1198/jbes.2009.07333>

81. Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962–973. <https://doi.org/10.1111/j.1541-0420.2005.00377.x>
82. Knaus, M. C. (2020). Double machine learning based program evaluation under unconfoundedness. *ArXiv Preprint ArXiv:2003.03191*.
83. Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19), 2937–2960. <https://doi.org/10.1002/sim.1903>
84. Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999*, 6–10.
85. Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448), 1096–1120. <https://doi.org/10.1080/01621459.1999.10473862>
86. Gustafson, P. (2012). Double-robust estimators: Slightly more Bayesian than meets the eye? *The International Journal of Biostatistics*, 8(2, 1), 1. <https://doi.org/10.2202/1557-4679.1349>
87. Cefalu, M., Dominici, F., Arvold, N., & Parmigiani, G. (2016). Model averaged double robust estimation. *Biometrics*, 73(2), 410–421. <https://doi.org/10.1111/biom.12622>
88. Zigler, C. M., & Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, 109(505), 95–107. <https://doi.org/10.1080/01621459.2013.869498>
89. Wang, C., Parmigiani, G., & Dominici, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68(3), 661–671. <https://doi.org/10.1111/j.1541-0420.2011.01731.x>
90. Wang, C., Dominici, F., Parmigiani, G., & Zigler, C. (2015). Accounting for uncertainty in confounder and effect modifier selection when estimating average

- causal effects in generalized linear models: Accounting for uncertainty in confounder and effect modifier selection when estimating acs in GLMs. *Biometrics*, 71(3), 654–665. <https://doi.org/10.1111/biom.12315>
91. Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2), 29–50. <https://doi.org/10.1257/jep.28.2.29>
92. Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 608–650. <https://doi.org/10.1093/restud/rdt044>
93. Antonelli, J., & Dominici, F. (2018). A Bayesian semiparametric framework for causal inference in high-dimensional data. *ArXiv Preprint ArXiv:1805.04899*. <https://arxiv.org/abs/1805.04899>
94. Antonelli, J., Parmigiani, G., & Dominici, F. (2019). High dimensional confounding adjustment using continuous spike and slab priors. *Bayesian Analysis*, 14(3), 805. <https://doi.org/10.1214/18-ba1131>
95. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.
96. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth; Brooks. <https://cds.cern.ch/record/2253780>
97. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
98. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(9), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
99. Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1023/A:1022627411411>
100. Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3), 293–300. <https://doi.org/10.1023/A:1018628609742>
101. Farrell, M. H., Liang, T., & Misra, S. (2018). Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands.

ArXiv Preprint ArXiv:1809.09953v3. [↵](#)

102. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539> [↵](#)

103. Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems, Lcbs-1857*, 1–15. https://doi.org/10.1007/3-540-45014-9_1 [↵](#)

104. Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The LASSO and generalizations*. CRC press. <https://doi.org/10.1201/b18401> [↵](#)

105. Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1), 267–288. <https://doi.org/https://doi.org/10.1111/j.2517-6161.1996.tb02080.x> [↵](#)

106. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition*, 770–778. [↵](#)

107. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*. [↵](#)

108. Athey, S., & Imbens, G. (2016). Resursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360. <https://doi.org/10.1073/pnas.1510489113> [↵](#)

109. Gao, Z., Hastie, T., & Tibshirani, R. (2020). Assessment of heterogeneous treatment effect estimation accuracy via matching. *ArXiv Preprint ArXiv:2003.03881v1*. [↵](#)

110. Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 685–725. <https://doi.org/10.1146/annurev-economics-080217-053433> [↵](#)

111. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, 107(5), 261–265. <https://doi.org/10.1257/aer.p20171038> [↵](#)

112. 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(1), 1–68.
<https://doi.org/10.1111/ectj.12097> [↵](#)
113. Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242.
<https://doi.org/https://doi.org/10.1080/01621459.2017.1319839> [↵](#)
114. Athey, S., Tibshirani, J., Wager, S., & others. (2019). Generalized random forests. *The Annals of Statistics*, 47(2), 1148–1178. <https://doi.org/10.1214/18-AOS1709> [↵](#)
115. Jeong, S., & Rockova, V. (2020). The art of BART: On flexibility of Bayesian forests. *ArXiv Preprint ArXiv:2008.06620v2*. [↵](#)
116. Bargagli-Stoffi, F. J., De-Witte, K., & Gnecco, G. (2019). Heterogeneous causal effects with imperfect compliance: A novel Bayesian machine learning approach. *ArXiv Preprint ArXiv:1905.12707v3*. [↵](#)
117. Bargagli-Stoffi, F. J., & Gnecco, G. (2020). Causal tree with instrumental variable: An extension of the causal tree framework to irregular assignment mechanisms. *International Journal of Data Science and Analytics*, 9(3), 315–337.
<https://doi.org/10.1007/s41060-019-00187-z> [↵](#)
118. Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D., & others. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1), 43–68. <https://doi.org/10.1214/18-STS667> [↵](#)
119. Hahn, P. R., Dorie, V., & Murray, J. S. (2019). Atlantic Causal Inference Conference (ACIC) data analysis challenge 2017. *ArXiv Preprint ArXiv:1905.09515v1*. [↵](#)
120. Knaus, M. C., Lechner, M., & Strittmatter, A. (2020). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, 24(1). <https://doi.org/10.1093/ectj/utaa014> [↵](#)
121. Hernán, M. A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *Chance*, 32(1), 42–49.
<https://doi.org/10.1080/09332480.2019.1579578> [↵](#)

122. D'Amour, A., Deng, P., Feller, A., Lei, L., & Sekhon, J. (2017). Overlap in observational studies with high-dimensional covariates. *ArXiv Preprint ArXiv:1711.02582*. [↵](#)
123. Chang, C., Kundu, S., & Long, Q. (2018). Scalable Bayesian variable selection for structured high-dimensional data. *Biometrics*, 74(4), 1372–1382. <https://doi.org/10.1111/biom.12882> [↵](#)
124. Van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data*. Springer Science & Business Media. <https://doi.org/https://doi.org/10.1007/978-1-4419-9782-1> [↵](#)
125. Candes, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6), 717–763. <https://doi.org/10.1007/s10208-009-9045-5> [↵](#)
126. Recht, B. (2011). A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(104), 3413–3430. <https://doi.org/10.5555/1953048.2185803> [↵](#)
127. Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490), 493–505. <https://doi.org/https://doi.org/10.1198/jasa.2009.ap08746> [↵](#)
128. Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2018). *Matrix completion methods for causal panel data models*. National Bureau of Economic Research. <https://doi.org/10.3386/w25132> [↵](#)
129. Arpino, B., & Mattei, A. (2016). Assessing the causal effects of financial aids to firms in tuscany allowing for interference. *The Annals of Applied Statistics*, 10(3), 1170–1194. <https://doi.org/10.1214/15-AOAS902> [↵](#)
130. Bargagli-Stoffi, F. J., Tortu, C., & Forastiere, L. (2020). Heterogeneous treatment and spillover effects under clustered network interference. *ArXiv Preprint ArXiv:2008.00707v2*. [↵](#)
131. Tortú, C., Forastiere, L., Crimaldi, I., & Mealli, F. (2020). Modelling network interference with multi-valued treatments: The causal effect of immigration policy on crime rates. *ArXiv Preprint ArXiv:2003.10525*. [↵](#)

132. Hill, J. (2004). Reducing bias in treatment effect estimation in observational studies suffering from missing data. *ISERP Working Papers, 04-01*.
<https://doi.org/https://doi.org/10.7916/D8B85G11> [↵](#)
133. Mattei, A., Mealli, F., & Pacini, B. (2014). Identification of causal effects in the presence of nonignorable missing outcome values. *Biometrics, 70*(2), 278–288.
<https://doi.org/10.1111/biom.12136> [↵](#)
134. Funk, J. M., & Landi, S. (2014). Misclassification in administrative claims data: Quantifying the impact on treatment effect estimates. *Current Epidemiology Reports, 1*(4), 175–185. <https://doi.org/10.1007/s40471-014-0027-z> [↵](#)
135. Imai, K., & Yamamoto, T. (2010). Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis. *American Journal of Political Science, 54*(2), 543–560. <https://doi.org/10.1111/j.1540-5907.2010.00446.x> [↵](#)
136. Pierce, B. L., & VanderWeele, T. J. (2012). The effect of non-differential measurement error on bias, precision and power in Mendelian randomization studies. *International Journal of Epidemiology, 41*(5), 1383–1393.
<https://doi.org/10.1093/ije/dys141> [↵](#)
137. Mattei, A., & Mealli, F. (2007). Application of the principal stratification approach to the faenza randomized experiment on breast self-examination. *Biometrics, 63*(2), 437–446. <https://doi.org/10.1111/j.1541-0420.2006.00684.x> [↵](#)
138. Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with “censoring” due to death. *Statistical Science, 21*(3), 299–309. <https://doi.org/10.1214/088342306000000114> [↵](#)
139. Baccini, M., Mattei, A., & Mealli, F. (2017). Bayesian inference for causal mechanisms with application to a randomized study for postoperative pain control. *Biostatistics, 18*(4), 605–617. <https://doi.org/10.1093/biostatistics/kxx010> [↵](#)
140. Forastiere, L., Mattei, A., & Ding, P. (2018). Principal ignorability in mediation analysis: through and beyond sequential ignorability. *Biometrika, 105*(4), 979–986.
<https://doi.org/10.1093/biomet/asy053> [↵](#)
141. Mattei, A., & Mealli, F. (2011). Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73*(5), 729–752. <https://doi.org/10.1111/j.1467-9868.2011.00780.x> [↵](#)

142. Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266–298.
<https://doi.org/10.1214/09-AOAS285> [↵](#)
143. Foster, J. C., Taylor, J. M. G., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24), 2867–2880.
<https://doi.org/10.1002/sim.4322> [↵](#)
144. Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. *Advances in Neural Information Processing Systems*, 2280–2288. <https://doi.org/10.5555/3157096.3157352> [↵](#)
145. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267(1), 1–38.
<https://doi.org/10.1016/j.artint.2018.07.007> [↵](#)
146. Lee, K., Bargagli-Stoffi, F. J., & Dominici, F. (2020). Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. *ArXiv Preprint ArXiv:2009.09036*. [↵](#)
147. Ding, P., & VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology*, 27(3), 368–377. <https://doi.org/10.1097/EDE.0000000000000457> [↵](#)
148. Franks, Alexander M., D’Amour, A., & Feller, A. (2019). Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 1–33.
<https://doi.org/10.1080/01621459.2019.1604369> [↵](#)
149. Gustafson, P., McCandless, L. C., Levy, A. R., & Richardson, S. (2010). Simplified Bayesian sensitivity analysis for mismeasured and unobserved confounders. *Biometrics*, 66(4), 1129–1137. <https://doi.org/10.1111/j.1541-0420.2009.01377.x> [↵](#)
150. Ichino, A., Mealli, F., & Nannicini, T. (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics*, 23(3), 305–327.
<https://doi.org/10.1002/jae.998> [↵](#)
151. Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2), 126–132.
<https://doi.org/10.1257/000282803321946921> [↵](#)

152. Liu, W., Kuramoto, S. J., & Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science: The Official Journal of the Society for Prevention Research*, 14(6), 570–580. <https://doi.org/10.1007/s11121-012-0339-5> [↵](#)
153. Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1), 13–26. <https://doi.org/https://doi.org/10.1093/biomet/74.1.13> [↵](#)
154. VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the E-Value. *Annals of Internal Medicine*, 167(4), 268–274. <https://doi.org/10.7326/M16-2607> [↵](#)
155. Zhang, B., & Tchetgen, E. J. T. (2019). A semiparametric approach to model-based sensitivity analysis in observational studies. *ArXiv Preprint ArXiv:1910.14130*. [↵](#)
156. Gang, M. (2004). *RBOUNDS: Stata module to perform Rosenbaum sensitivity analysis for average treatment effects on the treated*. [↵](#)
157. Keele, L. J. (2014). *Rbounds: An R package for sensitivity analysis with matched data*. <https://cran.r-project.org/web/packages/rbounds/index.html> [↵](#)
158. Nannicini, T. (2007). A simulation-based sensitivity analysis for matching estimators. *Stata Journal*, 7(3), 334–350. <https://doi.org/10.1177/1536867X0700700303> [↵](#)
159. Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., & Burgette, L. (2004). *Twang: Toolkit for weighting and analysis of nonequivalent groups*. <https://cran.r-project.org/web/packages/twang/index.html> [↵](#)
160. Schuemie, M. J., Cepeda, S. M., Suchard, M. A., Yang, J., Schuler, Y. T. A., Ryan, P. B., Madigan, D., & Hripcsak, G. (2020). How confident are we about observational findings in health care: A benchmark study. *Harvard Data Science Review*, 2(1). <https://doi.org/https://doi.org/10.1162/99608f92.147cc28e> [↵](#)
161. Rosenbaum, P. R. (1987). The role of a second control group in an observational study (with discussion). *Statistical Science*, 2(3), 292–316. <http://www.jstor.org/stable/2245766> [↵](#)
162. Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1), 151–161.

<https://doi.org/10.1162/003465302317331982> ↵

163. LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4), 604–620.

<http://www.jstor.org/stable/1806062> ↵

164. Ding, P., & Miratrix, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of m-bias and butterfly-bias. *Journal of Causal Inference*, 3(1), 41–57.

<https://doi.org/10.1515/jci-2013-0021> ↵

165. Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 37–48. <http://www.jstor.org/stable/3702180> ↵

166. Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 615–625.

<https://doi.org/10.1097/01.ede.0000135174.63482.43> ↵

167. Schisterman, E. F., Cole, S. R., & Platt, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology (Cambridge, Mass.)*, 20(4), 488. <https://doi.org/10.1097/EDE.0b013e3181a819a1> ↵

168. Shrier, I. (2008). Letter to the editor: Propensity scores letter to the editor. *Statistics in Medicine*, 27, 2740–2741.

<https://doi.org/https://doi.org/10.1002/sim.3554> ↵

169. Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4), 669–688. <https://doi.org/10.1093/biomet/82.4.669> ↵

170. Perkovic, E., Textor, J., Kalisch, M., & Maathuis, M. H. (2017). Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *The Journal of Machine Learning Research*, 18(1), 8132–8193. <https://doi.org/10.5555/3122009.3242077> ↵

171. Shpitser, I., VanderWeele, T., & Robins, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, 527–536. ↵

172. VanderWeele, T. J., & Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, 67(4), 1406–1413. <https://doi.org/10.1111/j.1541-0420.2011.01619.x> ↵

173. Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., & Schölkopf, B. (2016). Distinguishing cause from effect using observational data: Methods and benchmarks. *The Journal of Machine Learning Research*, 17(1), 1103–1204. [↵](#)
174. Spirtes, P., & Zhang, K. (2016). Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, 3(1), 3. <https://doi.org/https://doi.org/10.1186/s40535-016-0018-x> [↵](#)
175. Fisher, R. A. (1937). *The design of experiments*. Oliver And Boyd. [↵](#)
176. Rubin, D. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods*, 15(1), 38–46. <https://doi.org/10.1037/a0018537> [↵](#)
177. Zubizarreta, J. R., Small, D. S., & Rosenbaum, P. R. (2014). Isolation in the construction of natural experiments. *The Annals of Applied Statistics*, 8(4), 2096–2121. <https://doi.org/10.1214/14-AOAS770> [↵](#)
178. Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444–455. <https://doi.org/10.2307/2291629> [↵](#)
179. Angrist, J. D., & Evans, W. N. (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *The American Economic Review*, 88(3), 450–477. <http://www.jstor.org/stable/116844> [↵](#)
180. Imbens, G., & Lemieux, T. (2008). The regression discontinuity design—theory and applications. *Journal of Econometrics*, 142(2), 611–850. <https://doi.org/10.1016/j.jeconom.2007.05.008> [↵](#)
181. Li, F., Mattei, A., & Mealli, F. (2015). Evaluating the causal effect of university grants on student dropout: Evidence from a regression discontinuity design using principal stratification. *The Annals of Applied Statistics*, 9(4), 1906–1931. <https://doi.org/10.1214/15-AOAS881> [↵](#)
182. Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. (2019). *Synthetic difference in differences*. National Bureau of Economic Research. <https://doi.org/10.3386/w25532> [↵](#)
183. Stuart, E. A., Ackerman, B., & Westreich, D. (2018). Generalizability of randomized trial results to target populations: Design and analysis possibilities.

Research on Social Work Practice, 28(5), 532–537.

<https://doi.org/10.1177/1049731517720730>

184. Hartman, E., Grieve, R., Ramsahai, R., & Sekhon, J. S. (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 (Statistics in Society)*, 178(3), 757–778. <https://doi.org/10.1111/rssa.12094>

185. Pearl, J., & Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. *AAAI Conference on Artificial Intelligence, Proceedings of the Twentieth Conference*, 247–254. <https://doi.org/10.1109/ICDMW.2011.169>

186. Huber, M. (2020). Mediation analysis. *Handbook of Labor, Human Resources and Population Economics*, 1–38. https://doi.org/10.1007/978-3-319-57365-6_162-2

187. Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1), 21–29. <https://doi.org/10.1111/j.0006-341x.2002.00021.x>

188. Mealli, F., & Mattei, A. (2012). A refreshing account of principal stratification. *The International Journal of Biostatistics*, 8(1), 1–17. <https://doi.org/10.1515/1557-4679.1380>

189. Forastiere, L., Mealli, F., & VanderWeele, T. J. (2016). Identification and estimation of causal mechanisms in clustered encouragement designs: Disentangling bed nets using Bayesian principal stratification. *Journal of the American Statistical Association*, 111(514), 510–525. <https://doi.org/10.1080/01621459.2015.1125788>

190. Mattei, A., Li, F., & Mealli, F. (2013). Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program. *The Annals of Applied Statistics*, 7(4), 2336–2360. <https://doi.org/10.1214/13-AOAS674>

191. Mattei, A., Mealli, F., & Ding, P. (2020). Assessing causal effects in the presence of treatment switching through principal stratification. *ArXiv Preprint ArXiv:2002.11989*.

192. Mealli, F., & Pacini, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, 108(503), 1120–1131. <https://doi.org/10.1080/01621459.2013.802238>

193. Mealli, F., Pacini, B., & Stanghellini, E. (2016). Identification of principal causal effects using additional outcomes in concentration graphs. *Journal of Educational and Behavioral Statistics*, 41(5), 463–480.

<https://doi.org/10.3102/1076998616646199> ↵