

The Causal Cookbook: Recipes for Propensity Scores, G-Computation, and Doubly Robust Standardization

Arthur Chatton^{1,2}  and Julia M. Rohrer³ 

¹Faculté de Pharmacie, Université de Montréal, Montréal, QC, Canada; ²UMR Inserm 1246 - SPHERE, Nantes Université, Université de Tours, Nantes, France; and ³Wilhelm Wundt Institute for Psychology, Leipzig University, Leipzig, Germany

Advances in Methods and Practices in Psychological Science
 January-March 2024, Vol. 7, No. 1,
 pp. 1–22
 © The Author(s) 2024
 Article reuse guidelines:
 sagepub.com/journals-permissions
 DOI: 10.1177/25152459241236149
 www.psychologicalscience.org/AMPPS



Abstract

Recent developments in the causal-inference literature have renewed psychologists' interest in how to improve causal conclusions based on observational data. A lot of the recent writing has focused on concerns of causal identification (under which conditions is it, in principle, possible to recover causal effects?); in this primer, we turn to causal estimation (how do researchers actually turn the data into an effect estimate?) and modern approaches to it that are commonly used in epidemiology. First, we explain how causal estimands can be defined rigorously with the help of the potential-outcomes framework, and we highlight four crucial assumptions necessary for causal inference to succeed (exchangeability, positivity, consistency, and noninterference). Next, we present three types of approaches to causal estimation and compare their strengths and weaknesses: propensity-score methods (in which the independent variable is modeled as a function of controls), g-computation methods (in which the dependent variable is modeled as a function of both controls and the independent variable), and doubly robust estimators (which combine models for both independent and dependent variables). A companion R Notebook is available at github.com/ArthurChatton/CausalCookbook. We hope that this nontechnical introduction not only helps psychologists and other social scientists expand their causal toolbox but also facilitates communication across disciplinary boundaries when it comes to causal inference, a research goal common to all fields of research.

Keywords

causal inference, doubly robust estimator, g-formula, inverse probability weighting, potential outcomes.

Received 9/14/23; Revision accepted 2/13/24

Drawing causal inferences and quantifying them is a cornerstone of psychological research. Ever since the random assignment of individuals into different conditions was introduced in the social sciences in the late 19th century, it has been considered the “gold standard” for this purpose (Jamison, 2019). Therefore, psychologists are often reluctant to accept findings from nonrandomized studies that are explicitly presented as causal (Grosz et al., 2020). However, there are causal research questions for which randomization is unfeasible or unethical (Deaton & Cartwright, 2018)—for these questions, nonexperimental (i.e., observational) data can still be informative, but only if they are combined with the appropriate methods for causal inference. Unfortunately,

these methods are rarely taught in psychology curricula (D'Onofrio et al., 2020), leaving a certain knowledge gap.

Several articles from the last decade have aimed to fill this gap, including general introductions (e.g., Foster, 2010; Rohrer, 2018; Wysocki et al., 2022), work focusing on aspects such as mediation analysis (e.g., Nguyen et al., 2020; Rohrer et al., 2022) or longitudinal data modeling (e.g., Lucas, 2023; Rohrer & Murayama, 2023), and articles trying to build bridges between frameworks (e.g., Deffner et al., 2022; West & Thoemmes, 2010). With

Corresponding Author:

Arthur Chatton, Université de Montréal, Montréal, QC, Canada
 Email: arthur.chatton@umontreal.ca



Creative Commons NonCommercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits noncommercial use, reproduction, and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

some exceptions (e.g., Schafer & Kang, 2008; Thoemmes & Ong, 2016; Thoemmes & West, 2011), much of this work focuses on causal identification.

Causal identification focuses on whether and under which conditions it is possible to calculate a particular causal effect of interest from observational data (Elwert, 2013).¹ This involves determining whether researchers can find a set of control variables that allows them to correctly estimate the effect of interest. The question of causal identification usually worries psychologists most when it comes to observational data and rightfully so, because strong assumptions (e.g., no unobserved confounders) are necessary to conclude that a causal effect can be estimated. The step following causal identification is causal estimation, in which whatever data are available is used to compute an association that reflects the causal effect if all assumptions are met (Elwert, 2013).

In many ways, causal estimation is familiar terrain for many psychologists because the standard curriculum usually covers methods that can be used to estimate causal effects in some scenarios, such as analysis of variance (ANOVA) and analysis of covariance (ANCOVA), linear and nonlinear regression analysis, multilevel models, and structural equation models. With the present article, we intend to expand this toolbox by introducing readers to modern estimators that are often applicable under more general circumstances and that have been explicitly developed with the aim of causal estimation. We cover methods based on propensity scores, g-computation, and their combination (so-called doubly robust estimators; Kang & Schafer, 2007). Propensity scores have already been used in the social sciences since the last decade (Thoemmes & Kim, 2011), and several tutorials have been published in the psychological literature (e.g., Austin, 2011; Harder et al., 2010; Lanza et al., 2013). Nonetheless, we include them here because we want to discuss different usages with their respective strengths and weaknesses and because they are one ingredient used in doubly robust estimators.

An advantage of these modern estimators is that they are designed to recover clearly defined estimands. In psychology, it is common to talk about “*the* effect of X on Y,” which implies that there is one number that captures how X affects Y. If researchers want to summarize the effects of the action, they may do so in different ways, for example, by averaging over the whole population or subgroups of interest. Given the central role of the estimand, we start with how causal effects are defined in the potential-outcomes framework—a framework that was developed in the context of randomized experiments (Neyman, 1923) and may thus be quite accessible to many psychologists.

What Is a “Causal Effect”?

An example: Alcoholics Anonymous attendance and abstinence

Consider the following example: We are interested in whether Alcoholics Anonymous (AA) attendance successfully leads to abstinence 1 year after starting (Ye & Kaskutas, 2009). Imagine we had access to a large number of individuals whom we made follow the AA program. At the 12-month mark, we observe that 70% abstain from alcohol. Now, we take a time machine and ensure that the same people do not follow the AA program. At the 12-month mark, we observe that only 40% are abstinent. This would be the best possible proof that the intervention has a causal effect, increasing the prevalence of abstinence by 30 percentage points.

Unfortunately, this time machine does not exist. Quantifying the causal effect, therefore, requires two groups, one that does attend AA and another one that does not. But apart from that, the groups should be as similar as possible so that they emulate the time-machine study design as well as possible. Suppose the two groups we pick are “people who attend AA” and “people who do not attend AA.” This is not a good way to emulate the time machine: AA attendees may differ from nonattendees in many ways, including age, gender, religiosity, motivation, and so forth. Thus, any direct comparison of attendees and nonattendees will allow us to discern only whether AA attendance and abstinence are associated. To move from association to causation, something else is needed.

Randomized trials are considered the “gold standard” because random allocation (e.g., by the flip of a coin) leads to two more suitable groups. The only systematic difference between those groups will be whether they received the action (e.g., whether they attended AA). Of course, especially with small sample sizes, random allocation may still result in groups that are noticeably unbalanced on some characteristics—for example, more motivated individuals may end up in the action group by chance alone. However, conventional statistical analysis already takes this into account and correctly reflects the resulting imprecision with larger standard errors and wider confidence intervals for smaller samples (Senn, 2013). Other factors can induce bias after random assignment—inevitably, some will drop out of the study before reaching the 1-year mark (loss to follow-up); maybe not everybody randomly assigned into the action group shows up to the meetings (partial adherence).² Nonetheless, randomization greatly reduces the sources of bias that researchers need to worry about.

Table 1. Toy Data Set Illustrating the Potential Outcomes

A	Y	Y^1	Y^0
1	1	1	?
1	0	0	?
0	1	?	1
0	0	?	0

Counterfactuality

The potential-outcomes framework was initially developed in the context of agricultural experiments by the statistician Jerzy Neyman (1923). It was later expanded to observational settings (Rubin, 1974) and then also to longitudinal data (Robins, 1986). The powerful notation has been widely adopted and thus provides “more or less the lingua franca for thinking about and expressing causal statements” (Cunningham, 2021, p. 85). West and Thoemmes (2010) provided an accessible introduction for psychologists. We consider a scenario in which both the action and the outcome are binary to simplify the explanations, but the potential-outcomes framework applies much more broadly.

Let A be the action, AA attendance, which takes the value 1 in the case of attendance and 0 otherwise. Let Y be the outcome, alcohol abstinence, which takes the value 1 in the case of abstinence and 0 otherwise. The potential-outcomes framework is based on the following question: What if the individual experiences $A = 1$ rather than $A = 0$? Each individual has a pair of potential outcomes (Table 1). One of them, $Y^{A=1}$ (condensed to Y^1), reflects the observed outcome if the individual experiences $A = 1$. The other one, $Y^{A=0}$ (or Y^0), reflects the observed outcome if the individual experiences $A = 0$. Only one potential outcome can actually be observed; the other one remains counterfactual (Holland, 1986). For example, consider the first row in Table 1: This is a person who did attend AA ($A = 1$) and who was abstinent a year later ($Y = Y^1 = 1$). We do not know whether the person would have been abstinent if the person had not attended ($Y^0 = ?$).

Estimands

We can use this notation to define a targeted causal effect, also referred to as a “theoretical (or causal) estimand.” Two components define such a theoretical estimand (Lundberg et al., 2021): a unit-specific quantity, such as a specific contrast between the potential outcomes (e.g., their difference or their ratio), and a target population over which we want to aggregate (e.g., the adult population of a particular country or all people with alcohol use disorders). Considering our AA example, the unit-specific quantity may be the difference in

abstinence ($Y^1 - Y^0$). The target population may be people with alcohol use disorder who meet any additional study eligibility criteria (i.e., the entire study population). The resulting estimand is the so-called average treatment effect (ATE) on the entire population, $E[Y^1 - Y^0]$, the most common estimand, which answers the question, “How would abstinence differ, on average, if all participants attended AA meetings versus if no participants attended AA meetings?”

Other common estimands include the ATE on the treated (ATT; $E[Y^1 - Y^0 | A = 1]$) and the ATE on the untreated (ATU; $E[Y^1 - Y^0 | A = 0]$). The ATT targets a population made up of the treated individuals, as defined by the study’s eligibility criteria; for example, the average effect of AA attendance among people who did attend AA. For these people, we observe their outcome under treatment (Y^1) but need to infer their outcome without treatment (Y^0). How would AA attendees’ abstinence differ, on average, had they (counter to fact) not attended the meetings? This gives us the effect of withholding the treatment from those individuals who would otherwise experience it (with the sign reversed). The ATU is the flip side of this. How would AA nonattendees’ abstinence differ, on average, had they attended the AA meetings? The ATU is the effect of expanding treatment to those individuals who would otherwise not experience it. When who attends AA has not been randomly assigned, the ATT and the ATU may plausibly differ. For example, maybe individuals who are most likely to benefit from AA are also the most likely to attend meetings (e.g., because the social component is particularly motivating to them); in such a scenario, the ATT would be larger than the ATU. Or maybe the people who are most likely to benefit from AA (e.g., individuals who suffer from social isolation) are actually the least likely to attend meetings, rendering the ATU larger than the ATT. The ATE averages over both the treated and the untreated and can thus be considered a weighted average of the two (for a thoughtful discussion of these estimands, see Greifer & Stuart, 2021).

ATE, ATT, and ATU are so-called marginal effects because they aim at certain populations, thus averaging (“marginalizing”) across people who may vary on other features that can also matter for the magnitude of the causal effect. For example, women may profit more from AA than men. The resulting ATE is a weighted average over this heterogeneity and thus also depends on the gender ratio in the population. The notion of marginal effects is often conflated with the notion of causal effects, maybe because a randomized experiment will yield a marginal causal effect. However, causal effects are not limited to marginal effects; they can also be so-called conditional effects (Box 1). We focus on the marginal causal effects in the rest of the article.

Box 1. Regression Coefficients, Conditional Causal Effects, and Collapsibility

Although a marginal effect is the effect on the population (defined by the eligibility criteria; e.g., adults who have been diagnosed with an alcohol use disorder), a conditional effect is the effect for a particular subgroup of the population (e.g., 40- to 50-year-old women who have been drinking for more than 10 years). Psychological researchers routinely handle conditional effect estimates without being necessarily aware of it because of the widespread practices of directly interpreting regression coefficients. A regression coefficient reflects the change in the outcome variable when the predictor of interest is changed, “holding constant” all other predictors—that is, conditional on the other predictors. In simple linear models, this conditional effect corresponds to the marginal effect. However, even in linear models, things get more complicated if we include a multiplicative interaction term. In such a scenario, the coefficient of the predictor of interest may end up reflecting only the effect in one particular reference group (i.e., a conditional effect); psychologists usually deal with this by centering predictors so that the coefficient instead reflects the “main effect” (i.e., a marginal effect; Rohrer et al., 2022).

Outside of linear models, the correspondence between conditional and marginal effects can break down even further. For example, reasonably, one may expect that if one splits the population into subgroups and calculates effects in those subgroups, then the marginal effect across all subgroups should be some weighted average of the resulting conditional effects. But this is true only for so-called collapsible causal contrasts, which include risk differences and relative risks. In contrast, odds ratios are noncollapsible. This means that when effects are expressed in odds ratios, the marginal effect may be larger or smaller than any individual conditional effect (Whitcomb & Naimi, 2021).

The noncollapsibility issue is especially important in experiments (i.e., randomized trials). Including strong prognostic factor(s) of the outcome (e.g., the baseline, i.e., the outcome before the action) in a multiple regression analysis increases the statistical power (Kahan et al., 2014), but it changes the meaning of the regression coefficient of the action from the marginal effect to a conditional effect. To recover, for example, a marginal odds ratio (while keeping the benefit of increased power), methods designed to target marginal effects are necessary. See Morris et al. (2022) for a more complete discussion on using the methods discussed in the present article for experiments.

Another issue can arise when interpreting regression coefficients: The coefficients of controls may be misinterpreted as causal effects. Even if the model has been correctly specified so that the coefficient of the action has a causal interpretation, this does not extend to the coefficient of controls. In epidemiology, this is known as the “Table 2 fallacy” because it is usually Table 2 that displays these coefficients (Westreich & Greenland, 2013). For an accessible explanation in social sciences, see Hünernmund and Louw (2023) or Keele et al. (2020).

The causal estimand is theoretical—researchers cannot estimate it because they observe only one potential outcome per individual, with the other half of the potential outcomes remaining unobserved (i.e., counterfactual). In contrast, a statistical estimand (also called an “empirical estimand”), such as the mean difference in terms of observed outcomes between the two exposure groups ($E[Y|A = 1] - E[Y|A = 0]$), can be estimated. A causal estimand is identifiable if it maps onto a statistical estimand. In such a scenario, an observable metric allows one to make statements about an unobservable metric.

The distinction between the causal estimand and the statistical estimand allows one to define two different families of bias: identification bias and estimation bias (Díaz, 2020). Identification bias occurs when one of the assumptions necessary for identification is not met and the statistical estimand thus no longer maps onto the causal estimand. This type of bias is common to all causal methods and requires expert knowledge (Hernán et al., 2019); sometimes, it may even require one to target a different theoretical estimand altogether. Estimation

bias occurs when there are modeling issues. It is method-specific and can require additional assumptions (e.g., correct model specification in ordinary least squares regression). We discuss identifiability assumptions and the resulting potential biases in the next section and turn to potential estimation bias when discussing the respective estimators.

Identifiability

Four central assumptions are necessary to map the causal estimand to a statistical estimand.³ Exchangeability is usually the most contentious of these, and it is likely the one that psychologists are most aware of. It implies that individuals experiencing the action and individuals not experiencing the action are essentially “the same”: They had the same average risk of the outcome before experiencing the action. The two groups are thus exchangeable; the same effect estimate would have been obtained if one had swapped the action group and the control group. This assumption requires the absence of confounding and selection biases. If sources of

confounding or selection bias exist, one may “control” for them by adjusting for control variables (see Box 2); this results in the modified assumption of “conditional exchangeability” (also known as “no unmeasured confounding”), which is much more relevant in practice. Exchangeability would be violated if, for example, AA attendees are, on average, more motivated to change their behavior than nonattendees and, thus, more likely to be abstinent 1 year later. Controls may also include risk factors (as defined in Box 2); these do not contribute to exchangeability but can reduce variance (Chatton et al., 2020).

Positivity essentially means that individuals can theoretically experience all levels of the action. Structural violations of positivity occur if researchers include individuals whose probability of receiving a particular action level is zero. For example, for some people in rural areas, there may simply exist no AA-meeting opportunity. Positivity is needed for all variables required to achieve conditional exchangeability and also for any additional control variables included (e.g., to reduce variance; Chatton et al., 2020), but not beyond that (Westreich, 2020, p. 53). To use an example from Hernán and Robins (2020, p. 30), researchers do not have to ask themselves whether the probability of attending AA meetings is greater than zero for individuals with blue eyes because “having blue eyes” is (very likely) not necessary to achieve conditional exchangeability. Note that positivity can also be violated by chance, especially in small samples. For example, it may happen that in our particular sample, none of the men of a particular age group attend AA. Such violations do not threaten identifiability; however, they can result in estimation issues—we may end up with unstable estimates or may have to extrapolate in missing subgroups. Sometimes, such random violations are referred to as “sparsity,” with the term “positivity violation” exclusively used for structural violations.

Consistency implies that the observed outcomes actually match (are consistent with) the potential outcomes of interest. In practice, this means that the different action levels must be well defined and be manipulable in principle. For example, our current definition of AA attendance is underspecified and may yield nonconsistency: Attending one meeting in 12 months will not lead to the same potential outcome as attending one meeting per week during the same period, but both may count as “AA attendance” unless we clarify our criteria. How specific we need to be to ensure consistency ultimately is a judgment call based on domain expertise (Hernán, 2016)—for example, we may assume that which brand of coffee is served at the AA group does not matter and thus does not need to be specified; however, at least in principle, future research could prove this assumption wrong.

Noninterference means that the outcome of an individual is not affected by the intervention assignment or the outcome of other individuals. For example, noninterference may be violated if our study includes several people living together: In such a scenario, an attendee may counsel a nonattendee, thus leading to a “spillover” of the action. Although such spillover is a nuisance when estimating action effects, it may be of interest in its own right because it leads to other (causal) research questions (Loh & Ren, 2022). Consistency and noninterference are often jointly summarized as the stable unit treatment value assumption (Rubin, 1974).

Causal Estimation as Cake Baking

The causal estimation workflow is a bit like baking a challenging cake (Fig. 2). Imagine you want to bake a cake resembling a character from a popular (noncopyrighted) children’s TV series—that is the causal estimand, the abstract goal of your efforts. On the Internet, you find a cake (the statistical estimand) that is close enough to what you imagined (identifiability). This cake comes with a recipe (the estimator), which you use to create your cake (the estimate). This is an ambitious project that will require a lot of experience and/or collaboration with a baking expert (a statistician). During the baking process (causal estimation), you may strictly follow the recipe provided, or you may adapt it to the ingredients (the data) available to you. You may also want to make other changes, such as adjusting the cooking time or temperature (varying the assumptions of the estimator). There are no guarantees that the cake you will end up with will resemble the cake you imagined, but you can still try your best.

Causal Estimators

We present two families of causal estimators that can be distinguished by their nuisance function. Although the nuisance function is not of direct interest to us, we use it to estimate the causal effect. To sustain the cake comparison, the nuisance function may be an essential part of the cake (e.g., the cake base) that must be prepared according to its own recipe.

The first family of estimators is based on propensity scores (Rosenbaum & Rubin, 1983). The corresponding nuisance function is typically denoted with $e(C)$. This function takes as ingredient C , the set of controls—which should include all variables needed to achieve conditional exchangeability (and may include more to improve the precision of the estimate).⁴ Because we model a binary action (recall that this may also be referred to as the “treatment,” the “intervention,” or the “exposure”), the function returns an individual’s propensity (i.e.,

Box 2. Association Versus Causation

Directed acyclic graphs (DAGs) conceptualize expert knowledge about a problem with the help of arrows and nodes. Each node represents a variable (measured or unmeasured); each arrow presents a causal effect (of any possible form). Ultimately, a DAG is the graphical representation of a system of nonparametric equations called a structural causal model (Pearl, 1995) and can be used as a conceptual tool. Rohrer (2018) provided an introduction for psychologists; Kunicki et al. (2023) explained how DAGs and structural equation models differ.

Let us start with a minimal DAG that includes an action node A and an outcome node Y.

We now add additional nodes (Figure 1a) that can, depending on their relationship to A and Y, induce different types of association between A and Y. A confounder is a common cause of the action and the outcome and can induce a spurious (i.e., noncausal) association. A mediator is causally affected by the action and causally affects the outcome, thus inducing a causal association. A collider is a common consequence of both the action and the outcome, and it does not induce an association between A and Y (unless it is conditioned on, see below). In addition, a variable may directly affect only the action (a so-called instrument), or it may affect only the outcome (a so-called risk factor). We may add more nodes, resulting in increasingly more complex paths between the nodes.

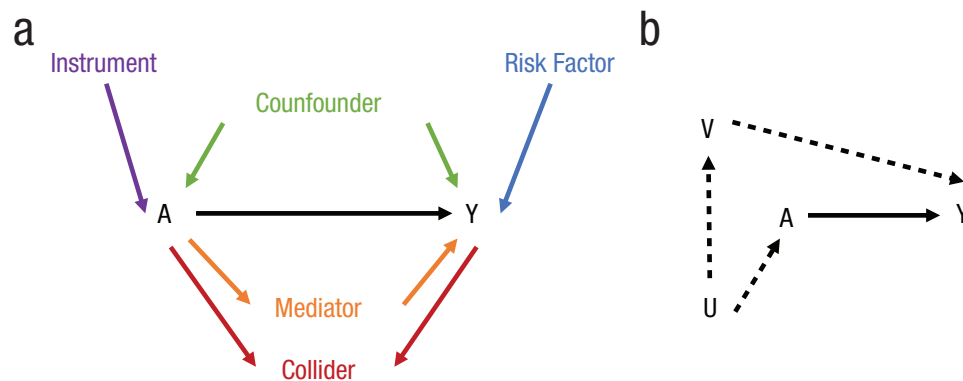


Fig. 1. Directed acyclic graphs illustrating (A) the possible components and (B) the backdoor (dotted lines) and frontdoor (solid line) paths.

“Frontdoor paths” link action A to outcome Y in the direction of arrows. These paths can include multiple mediators, and figuring out their causal contributions is the goal of mediation analysis (Imai et al., 2010). Frontdoor paths transmit causal associations and are therefore no cause for concern when it comes to causal identification.

“Backdoor paths” start with an arrow pointing to action A and end with an arrow pointing to outcome Y, such as $A \leftarrow U \rightarrow V \rightarrow Y$ in Figure 1b. Backdoor paths can transmit noncausal associations; we therefore need to block them if we want to identify the effect of A on Y.

Whether or not a path is blocked will depend on the causal structures and which variables we are conditioning on (by, e.g., stratifying on them, including them as a predictor in a regression, including them in the model underlying propensity scores). A set of four rules known as “d(irected)-separation” can be used to figure out whether a path is open or blocked (Geiger et al., 1990; Thoemmes et al., 2018):

1. Usually (that is, without conditioning on other variables) a path is open unless it contains a collider. If it contains a collider, it is blocked.
2. A path is blocked if we condition on a confounder or on a mediator (i.e., on a noncollider) that lies on the path.
3. A path opens if we condition on a collider that lies on the path.
4. A path opens if we condition on a variable that is affected by a collider that lies on the path.

Our goal is to find a set of controls (i.e., variables on which we condition) that ensures that all backdoor paths are closed while all frontdoor paths remain open so that only causal associations flow freely. We can determine such sets by “manually” applying the d-separation rules; software such as the website DAGitty.net (Textor et al., 2011) can also do the job. In any case, we need the DAG as an input, and this DAG rests on expert knowledge—no statistical approach can determine whether a given variable is a confounder, a mediator, or a collider. Only careful thought about the plausibility of causal relationships and temporality can help us determine a sound DAG (Cinelli et al., 2022; Wysocki et al., 2022).

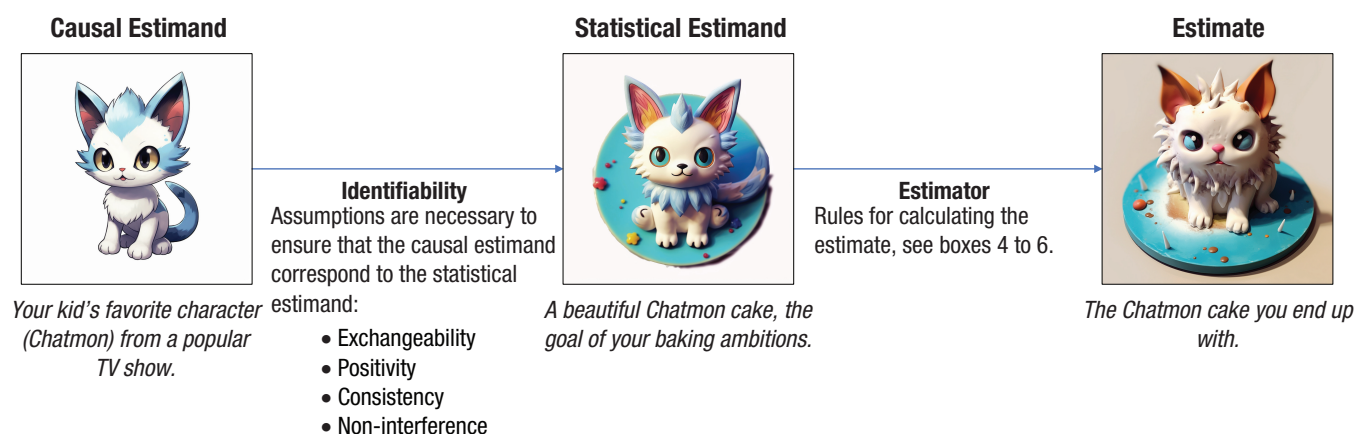


Fig. 2. A cooking metaphor for the causal estimation workflow.

probability) to experience it ($A = 1$), according to their values on the control variables C ; this can be written as $P(A = 1 | C)$. For a brief discussion of nonbinary actions, see Box 3. For example, $e(C)$ could be a model for AA attendance from controls (e.g., age, gender, family history of alcohol use disorders). In the next step, propensity-score-based methods use $e(C)$ to emulate a randomized controlled trial.

The second family of estimators is based on g-computation. The corresponding nuisance function can be denoted with $Q(A, C)$. The function takes both the action

A and the set of controls C as ingredients and returns the probability of the outcome itself, $P(Y = 1 | A, C)$. For example, $Q(A, C)$ could be a model predicting abstinence from both AA attendance and controls. At this point, one may wonder whether the estimation process is already finished. After all, $Q(A, C)$ is the type of regression model from which researchers routinely take the coefficients and interpret them as causal effects. However, for most models, this works in only the simplest (linear, additive) case (see also Box 1). Here, g-computation, which (unlike standard regression) has been specifically

Box 3. Causal Estimators for Nonbinary Actions

If the action is binary, the causal effect contrasts the potential outcome in one of the two hypothetical worlds (Y^1) to the potential outcome in the other world (Y^0). If the action is nonbinary, multiple hypothetical worlds can be contrasted—potentially even an infinite number of them. One classical way to evaluate them are dose-response curves in which causal effects across different values of a polynomial or continuous action are displayed. However, correctly identifying and estimating such effects can be challenging.

Considering the identification of causal effects, the assumptions described in the main text need to hold here as well—but for all possible values of A (McCaffrey et al., 2013). In particular, positivity can be a challenging when the action is continuous because this results in an infinite number of possible values for the action.

Considering the estimation of causal effects, the generalized propensity score is an extension proposed by Imai and van Dyk (2004) and Hirano and Imbens (2004) to deal with continuous exposure. The basic idea is to fit a propensity function (the equivalent of the propensity score) by modeling the continuous action on the controls needed to achieve the conditional exchangeability, using, for example, linear regression. Then, one can match, condition, stratify, or weight according to the propensity function. We refer readers to Zhao et al. (2020) for a comprehensive review of the generalized propensity score estimation and to the second section of Austin (2018) for a more accessible introduction. Note that stabilizing the weights in inverse probability weighting is much more important here to avoid infinite variance (Robins et al., 2000). Austin (2019) and Fong et al. (2018) suggested balance checking ways for a continuous action.

In g-computation, we directly model the hypothetical worlds: How could that work for continuous actions? When we are willing to assume a monotonous linear effect of the action, comparing one world where A is set to a with another world where A is set to $a + 1$ gives an estimate of the average causal effect. Without such an assumption, one must construct the dose-response curve by predicting the average individual hypothetical outcome in a sufficient number of hypothetical worlds (differing by their value of A) using substantively meaningful thresholds for the action.

developed for the task of causal inference (Robins, 1986), provides a much more general solution. It also works with nonlinear models and does not require coming up with clever coding schemes but instead requires an additional analysis step. In this additional step, $Q(A, C)$ is used to estimate individuals' outcomes in two (or more) hypothetical worlds—one in which they experience the action and one in which they do not. The contrast between their outcomes in those worlds then informs us about the causal effect.

Which of these two families is preferable—propensity-score-based methods in which the action-allocation process (e.g., AA attendance) is modeled or g-computation in which the outcome (e.g., abstinence) is modeled? A first rule of thumb would be to prefer propensity-score-based methods when the outcome is scarce (i.e., when almost no one in the sample is abstinent after 1 year) and g-computation when the action allocation is unbalanced (e.g., say an allocation ratio of 1 attendee for 5 nonattendees) to avoid a modeling issue. In general, g-computation is asymptotically more accurate (Tan, 2007). Nevertheless, each approach has its own strengths and pitfalls, which we discuss below.

Some sophisticated cakes require two different bases, and in that vein, a third family of estimation approaches combines both nuisance functions, which results in so-called doubly robust estimators. Here, we first model the action-allocation process and then make use of the resulting propensity scores when modeling the outcome. Such estimators have the desirable property that they result in a consistent estimate (i.e., theoretically unbiased in infinite samples) as long as one of the two nuisance models is correct; however, as we show below, this comes at a cost.

Propensity-score-based estimators and inverse probability weighting

The propensity score (introduced by Rosenbaum & Rubin, 1983) summarizes all observed controls into a single variable. It is a balancing score: Conditional on the correctly specified propensity score, the distribution of controls included in it is similar for individuals experiencing the action and individuals not experiencing the action. Therefore, it allows the emulation of a pseudorandomization situation to draw causal inferences. Once estimated, $e(C)$ can be used in four different ways.

First, adjustment means that the propensity score is included as a covariate—in the very same way one would usually include individual control variables as covariates. This approach relies on strong modeling assumptions (Vansteelandt & Daniel, 2014).

Second, for stratification, the sample is divided into subgroups (strata) based on their propensity score; in

the next step, the action's effect is estimated in every single subgroup, and those estimates are combined into an overall effect estimate. This can be done only with a finite number of subgroups, and thus, people with different scores will usually end up in the same stratum, which leads to residual confounding (Lunceford & Davidian, 2004).

Third, in matching, for each individual in the action group, we pick an individual not experiencing the action with a similar propensity score to include them in the control group. Simply comparing these two groups then yields an estimate of the action's effect. Some authors have argued against the usage of matching for reasons such as covariate balance, inefficiency, model dependence, and bias (King & Nielsen, 2019). However, matching remains a popular approach, with the central advantage that it results in a situation comparable with a randomized experiment with exchangeable groups. There are already excellent sources introducing psychologists to matching (Chan et al., 2022; Stuart, 2010), which is why we do not cover the topic in more depth.

This leaves us with, fourth, weighting; more specifically, inverse probability weighting (IPW; Robins et al., 2000). This approach appears to be less biased and more precise than matching according to simulation studies (Chatton et al., 2020; and references therein). In IPW, the idea is to generate a pseudosample in which the groups are exchangeable. Rather than actually picking individuals to be included in the groups (as is done in matching), here, one assigns weights to each individual, which determines how much they “contribute” to the analysis. Box 4 summarizes the IPW recipe.

The individual weights are determined as a function of $e(C)$, the propensity score. Weights can be calculated in different ways, which allows us to estimate effects for different target populations, including the entire population (ATE), the treated population (ATT), or the untreated population (ATU). Table 2 displays some of the weighting schemes, and the companion R notebook illustrates how they work in practice. These different weighting schemes render IPW the most flexible propensity-score-based approach. Once the weights have been computed, they can be stabilized. Stabilized weights preserve the sample size of the observed sample and avoid some estimation issues (e.g., variance inflation, Xu et al., 2010; or random violations of positivity, Robins et al., 2000). This is especially true for continuous actions (as described in Box 2). To evaluate the precision of the results, it might be helpful to calculate the so-called effective sample size, which is the size of an unweighted sample yielding the same precision as the weighted pseudosample (McCaffrey et al., 2004). In other words, it estimates the number of comparable individuals between the groups.

Next, to obtain an estimate of the causal effect, a weighted regression (called “marginal structural model”

Box 4. Inverse Probability Weighting Recipe

Ingredients: Action A, outcome Y, and controls C.

Important: For an unbiased estimate, the controls C must be sufficient to achieve conditional exchangeability.

Step 1: Model the action allocation process as $e(C)$, a function of the controls. For example, this could be a logistic regression predicting A from C, using all individuals in the sample.

Step 2: Use $e(C)$ to compute the individual weights ω as defined in Table 2.

Step 3: Fit a regression with Y as the dependent variable and A as the sole independent variable, weighted using ω .

Possible modifications: Control variables may be included in this step to remove residual confounding. Note that this changes the targeted theoretical estimand.

Step 4: The coefficient of A is the estimate of the estimand defined by the weighting scheme (Step 2) and the weighted model (see possible modifications in Step 3).

Examples of R packages implementing this estimator: propensity, PSweight, WeightIt

[MSM]) modeling the outcome of interest is fitted. If the identifiability assumptions are met, this is in fact a model of the potential outcomes. The coefficient of the action in the MSM corresponds to a specific contrast of the potential outcomes, defined by the type of regression (Schnitzer et al., 2020). For example, if we run a linear regression for a binary outcome,⁵ the coefficient will give us the risk difference (“Attending AA increases the risk of abstinence by 30 percentage points”), a log-linear regression will give us the risk ratio (“Attending AA increases the risk of abstinence by a factor of 1.75”), and a logistic regression gives us an odds ratio (“Attending AA increases the odds of abstinence by a factor of 3.5”; all of these numbers reflect valid causal effects defined by different causal contrasts). To quantify the uncertainty of this estimate (e.g., to compute the standard error), one can use a so-called robust sandwich-type matrix or a bootstrap approach (for an introduction to bootstrapping, see Rousselet et al., 2021). According to recent simulation studies, bootstrapping seems more accurate (Austin, 2016, 2022) and yields valid inferences by considering

both uncertainties in the propensity score and in the MSM (Berk et al., 2013). A Bayesian approach is also possible (Spertus & Normand, 2018).

The goal of the weighting procedure is to balance the controls between the two groups (e.g., to make AA attendees and nonattendees comparable on the relevant third variables; West et al., 2014); whether such a balance has been achieved can be checked. Franklin et al. (2014) suggested 10 metrics for checking the balance of the pseudosample. Among them, the standardized mean difference has been reported as the most accurate (Ali et al., 2014); a value lower or equal to 10% is considered acceptable (Ali et al., 2015; for the formulas, see Austin & Stuart, 2015). The other metrics can also be used, but they need at least 1,000 individuals, according to Ali et al. (2014). There is no point in running statistical tests to compare individual control variables between the two action groups because the resulting p values are not informative (Imai et al., 2008). If the action groups remain unbalanced in the weighted (pseudo)sample on some controls, those variables can be included as predictors in the MSM. This might reduce potential residual

Table 2. Examples of Weighting Schemes and Their Targeted Population

Name	Weight if A = 1	Weight if A = 0	Target population
ATE (unstabilized)	$1/e(C)$	$1/[1 - e(C)]$	Whole sample
ATT (unstabilized)	1	$e(C)/[1 - e(C)]$	Treated
ATU (unstabilized)	$[1 - e(C)]/e(C)$	1	Untreated
Stabilized ATE	$P(A = 1)/e(C)$	$P(A = 0)/[1 - e(C)]$	Whole sample
Overlap	$1 - e(C)$	$e(C)$	Unclear ^a

Note: A = action, treatment; ATE = average treatment effect on the entire population; ATT = average treatment effect on the treated; ATU = average treatment effect on the untreated.

^aOverlap weights were suggested by F. Li et al. (2019) as a solution to extreme propensity scores, which we discuss in the section Iffy Identifiability.

Table 3. Hypothetical Data for the Example of Alcoholics Anonymous Attendance (AA) and Abstinence, Including Two Controls (Family History of Alcohol Abuse, Gender), $N = 200$

	Total		AA attendees ($a = 1$)		Nonattendees ($a = 0$)	
	Women	Men	Women	Men	Women	Men
Family history	30 (15% of sample)	70 (35%)	25, of which 20 (80%) abstinent	40, of which 25 (62.5%) abstinent	5, of which 3 (60%) abstinent	30, of which 8 (26.7%) abstinent
No family history	30 (15%)	70 (35%)	20, of which 16 (80%) abstinent	35, of which 24 (68.6%) abstinent	10, of which 5 (50%) abstinent	35, of which 14 (40%) abstinent

confounding, which is desirable with respect to exchangeability; at the same time, it may shift the theoretical estimand from marginal to conditional if the causal contrast is not collapsible and if the variables included in the MSM are effect modifiers (Robins et al., 2000), which may affect interpretability.

G-computation

G-computation has its roots in so-called stratification and standardization, the process of splitting up the sample into subgroups (strata), calculating the metric of interest in each group, and then reweighting the group-specific metrics to match, for example, the general population. This used to be a common approach to control for confounders in observational studies, dating back as far as the mid-19th century (Neison, 1844), before computationally more demanding methods took hold (for a historical perspective, see Keiding & Clayton, 2014). Robins (1986) extended the logic of standardization to the so-called g(eneral)-formula for estimating causal effects, which allows for incorporating time-dependent confounding within the potential-outcomes framework. It is thus suitable for longitudinal data, but here we consider the time-fixed setting to simplify explanations.

The idea behind the g-formula is to estimate the probability of the outcome (e.g., abstinence) under a hypothetical action (e.g., AA attendance or nonattendance)—in other words, we are trying to estimate the probability of the potential outcomes:

$$P(Y^a = 1) = \sum_c P(Y = 1 \mid A = a, C = c) P(C = c).$$

In words, the probability of the (potential) Outcome 1 under the action a , $P(Y^a = 1)$, equals the weighted sum⁶ of the outcome probabilities for the individuals experiencing a across each subgroup of controls, $P(Y = 1 \mid A = a, C = c)$. The weights are the respective probabilities of being a member of the particular subgroup c .

Consider a simple scenario with two controls with two levels: family history of alcohol abuse (yes/no) and gender (female/male). Cross-tabulating the outcome (abstinence) for these two controls for (a) the whole sample and separately for (b) AA attendees and (c) nonattendees (Table 3) gives us all the information we need to apply the g-formula. For the probability of abstinence among the attendees, for each of the four subgroups, we simply multiply the fraction abstinent (middle part of Table 3) with the fraction that the subgroup makes up in the whole sample (left part of Table 3) and then add up the numbers: $P(Y^1 = 1) = .80 \times .15 + .625 \times .35 + .80 \times .15 + .686 \times .35 \approx .70$. Repeating the same steps for the nonattendees (right part of Table 3) gives us the probability of abstinence among nonattendees, $P(Y^0 = 1) = .60 \times .15 + .267 \times .35 + .50 \times .15 + .40 \times .35 \approx .40$. Thus, under the identifiability assumptions spelled out above, attendance increases the probability of abstinence by 30 percentage points: from 40% to 70%.

The g-formula essentially allows us to place ourselves in counterfactual worlds in which everybody or nobody attended AA.⁷ It is nonparametric because it does not assume any functional form for the relationships between variables. In our simple scenario, this works well because we have only two controls with two levels, resulting in four subgroups. But things quickly get out of hand if we add more (categorical) controls—leading to an exponential increase in the number of subgroups (so-called curse of dimensionality)—and/or if we add continuous controls. Thus, for realistic scenarios in which more than just a few controls are necessary to achieve conditional exchangeability, we need g-computation, a model-based extension of the g-formula proposed by Robins (1986).⁸

G-computation (Box 4) is an attempt to emulate the time machine described at the beginning of this article. It aims to model two counterfactual worlds, one in which everybody who meets our inclusion criteria attends AA and one in which nobody does, and predict each individual's outcome in these worlds. The first step consists of fitting the nuisance function $Q(A, C)$ with AA attendance and all controls needed to achieve conditional exchangeability. Here, we use everybody's observed

Box 5. G-Computation Recipe

Ingredients: Action A, outcome Y, and controls C.

Important: For an unbiased estimate, the controls C must be sufficient to achieve conditional exchangeability.

Step 1: Model the outcome as $Q(A, C)$, a function of the controls. For example, this could be a logistic regression predicting Y from A and C, using all individuals in the sample.

Step 2: Duplicate the initial data set in two counterfactual data sets. In one of them, set $A = 1$; in the other one, set $A = 0$. All other variables keep their original values.

Step 3: Apply the function $Q(A, C)$ to predict each individual's outcome in the two counterfactual data sets; these are the model-implied potential outcomes Y^1 and Y^0 .

Step 4: Aggregate these potential outcomes (e.g., average across all individuals) and contrast them (e.g., by taking their difference) to arrive at an estimate of the estimand of interest.

Examples of R packages implementing this estimator: `marginalEffects`, `RISCA`, `stdReg`.

characteristics, including their observed AA attendance. For example, $Q(A, C)$ could be a logistic regression. In the first step, we determine the coefficients of the predictor AA attendance and of the controls. In the second step, we create two hypothetical worlds—one in which everybody attends AA and one in which nobody attends AA. To do so, we simply copy the data twice and set the action variable to 1 (world of attendance) or 0 (world of nonattendance) for everybody, keeping their controls at the originally observed levels. We then use the coefficients from the first step to predict the (potential) outcomes in the two worlds. For each individual, we now have the individual's outcome probability for the scenario in which the individual attends and for the scenario in which the individuals does not attend.

We can take the difference between these probabilities to compute the individual-level causal effects, and we can calculate the ATE by averaging over individuals. Alternatively, we can first average the potential outcome probabilities and then compute a wider range of causal effects (e.g., the odds ratio). Other causal estimands are also easily computable from the predicted potential outcomes. Again, we usually do not want only a point estimate but also some way to quantify its uncertainty (e.g., to compute the standard error). Here, bootstrap approaches and the so-called delta method are frequently used (although a specific variance estimator also exists; Zou, 2009). A Bayesian approach is also possible (e.g., Keil, Daza, et al., 2018; for an applied example close to psychology, see also Rohrer et al., 2021), in which case, the posterior distribution of the parameter of interest provides for a straightforward quantification of uncertainty.

The recipe presented in Box 5 can be varied at multiple points. For example, instead of predicting both counterfactual worlds for all individuals in Step 3, we

may instead predict only the unobserved outcome (e.g., the outcome without action for those individuals who did in fact receive the action) and keep the observed outcomes untouched to improve accuracy (Westreich et al., 2015). In Step 1, we can also fit one nuisance model per action group, $Q(A = 1, C)$ and $Q(A = 0, C)$, for predicting the counterfactual outcome (Künzel et al., 2019). Fitting two nuisance models means that we do not have to explicitly model interactions between the action and controls; however, this approach is sensitive to data-set shift when predicting the potential outcomes: A nuisance model fitted only on one action group may have poor predictive performance when applied to another action group because they differ too much (Finlayson et al., 2021). In Step 4, to estimate the causal effect, we can also regress the counterfactual predictions on the action in an MSM (Snowden et al., 2011).

In contrast to propensity-score-based methods, g-computation does not require the assumption of balance between groups because it holds “by design” between the two counterfactual worlds. However, the flip side of this is that we can demonstrate balance only on measured controls when we use propensity-score-based methods. Such a demonstration can, in turn, convince both researchers and readers that bias because of measured controls has been removed. A similar trade-off arises for positivity. G-computation may be able to simply extrapolate over missing strata; propensity-score-based methods, in contrast, allow us to check for extreme propensity scores and thus notice positivity violations (or a lack thereof).

Doubly robust standardization

Both propensity-score-based methods and G-computation require the correct specification of their

Box 6. Doubly Robust Standardization Recipe

Ingredients: Action A, outcome Y, and controls C.

Important: For an unbiased estimate, the controls C must be sufficient to achieve conditional exchangeability.

Step 1: Model the action allocation process as $e(C)$, a function of the controls. For example, this could be a logistic regression modeling A from C, using all individuals in the sample.

Step 2: Use $e(C)$ to compute the individual weights ω as defined in Table 3.

Step 3: Model the outcome as a function of the controls, $Q(A,C)$; however, this time also weight the model by ω , as in inverse probability weighting.

Step 4: Duplicate the initial data set in two counterfactual data sets; set $A = 1$ in one of them and $A = 0$ in the other one.

Set 5: Apply the function $Q(A,C)$ to predict each individual's outcome in the two counterfactual data sets, Y^1 and Y^0 .

Step 6: Aggregate these potential outcomes (using the weighted mean) and contrast them to arrive at the estimate.

Examples of R packages implementing this estimator: `marginalEffects`, `RISCA`, `stdReg`.

Note that the user must provide the weighted model as an argument for the g-computation function and bootstrap the whole procedure.

respective nuisance model, $e(C)$ and $Q(A,C)$. This means that the models have to approximate the true data-generating process—either of the assignment of the action, $e(C)$, or of the outcome, $Q(A,C)$ —as closely as possible to result in valid inferences. However, because of the complexity of the real world, a correct specification is unlikely, and as a result, estimates can be biased (van der Laan & Rose, 2011, p. 9). Doubly robust estimators provide a partial solution to this problem by combining both nuisance models; they give us two shots to get things right: As long as one of the nuisance models is correctly specified, the resulting estimate does not suffer from misspecification bias.⁹ However, this doubly robust property comes at a cost: Although (systematic) bias may be reduced, variance increases in comparison with g-computation (Tan, 2007); thus, we face a bias-variance trade-off (Pargent et al., 2023).

There are different ways to combine the nuisance models $e(C)$ and $Q(A,C)$, resulting in various doubly robust estimators. Here, we focus on the one that we consider most intuitive: doubly robust standardization (DRS; Robins et al., 2007). Recall that IPW aims to balance the AA attenders and nonattenders on the controls so that the analysis emulates a randomized trial. If $e(C)$ is misspecified, this emulation fails, and some residual confounding remains. DRS tackles this residual confounding by adding a g-computation step after the IPW (Box 6). An alternative way to think about DRS is to consider that it is easier to model the counterfactual worlds with g-computation from a randomized trial (even if it is miss-emulated) rather than from scratch because some confounding has already been removed.

All variations of IPW and g-computation described above can be applied to DRS. Again, bootstrapping (for the whole process, i.e., for both the IPW and g-computation steps) or the delta method can be used to quantify the uncertainty in the resulting point estimate. When both nuisance models are misspecified, some doubly robust estimators are actually more biased than either IPW or g-computation (Kang & Schafer, 2007)—fortunately, DRS is not affected by this bias-amplification phenomenon (Chatton et al., 2022).

What Could Possibly Go Wrong?

Iffy identifiability

Any causal-estimation effort can succeed only if the statistical estimand actually corresponds to the causal estimand, and as explained earlier, this requires assumptions: exchangeability (an absence of confounding and collider bias, see Box 2), positivity (nonextreme probabilities of ending up in either action group) with respect to the controls included to achieve exchangeability, consistency (potential outcomes correspond to the observed outcomes), and noninterference (outcome of an individual is not affected by the outcome or action of another individual). Any of these assumptions can fail, leading to biased estimates. Conversely, inferences can be strengthened by trying to render these assumptions more plausible.

Recent reviews in social sciences suggest that the inclusion of controls to achieve exchangeability is often insufficiently justified (Bernierth & Aguinis, 2016; Kohler

et al., 2023), leaving a lot of room for improvement. Wysocki et al. (2022) suggested spelling out several plausible causal structures and selecting the controls as the minimal set blocking all backdoor paths. The selection of controls can also be achieved by data-driven procedures (for such an approach introduced in psychology, see Loh & Ren, 2023a)—however, such procedures in themselves are unaware of the underlying causal structure, and they thus need to be combined with existing domain knowledge to achieve exchangeability. As spelled out before, doubly robust estimators may offer advantages here because even if a confounder is missing in one nuisance model, groups remain conditionally exchangeable if it is present in the other nuisance model (Chatton et al., 2022). However, the (erroneous) inclusion of a mediator as a control in either $Q(A,C)$ or $e(C)$ withdraws the doubly robust property of DRS and increases the resulting bias compared with IPW or g-computation (Keil, Mooney, et al., 2018). Regardless of the estimation approach used, concerns regarding exchangeability call for robustness checks. So-called sensitivity analyses try to assess to which extent estimates may be biased because of unobserved confounding. Although these analyses provide no guarantees, they help gauge how worried one should be about the robustness of the results. X. Zhang et al. (2020) provided a review of modern statistical methods and suggested a specific order of steps to evaluate the impact of potential unmeasured confounders.

Although exchangeability always requires a leap of faith—one can never be completely certain that there is no unobserved confounding—positivity can be checked empirically. The classic approach here involves checking whether the estimated propensity scores include extreme values. We recommend using PoRT, a tree-based algorithm recently developed by Danelian et al. (2023), because it can be used with all estimators, does not require assumptions about the data-generating process, and clearly identifies the target population. If a violation of positivity is structural, the target population must be redefined—one cannot estimate the effect of the action in the subgroup that would never experience the action.

Random violations of positivity result in estimation issues that are especially harmful when using IPW because they result in extreme weights and, thus, oversized influence of individual observations on the results. They can be addressed in various manners. Several authors have proposed to trim propensity scores (i.e., to remove observations with extreme values) or to truncate them (i.e., to set all values exceeding a certain threshold to a fixed value). A recent simulation study suggests that a threshold of $5 / [\sqrt{n} \cdot \ln(n)]$ achieves the best performance (Gruber et al., 2022). However, such procedures can shift the target population and thus the theoretical estimand (Zhu et al., 2021). Alternatively, one may use

overlap weights (see Table 2) that target causal effects in the overlapping population, that is, in the part of the population in which positivity holds (F. Li et al., 2019). These weights down-weight individuals who are extremely unlikely or extremely likely to experience the action. Although these weights have strong statistical properties with respect to positivity, the target population once again changes and potentially becomes ill-defined, which may limit the external validity of the results.

As previously discussed, g-computation can be less sensitive to random violations of positivity because it allows for extrapolation over the missing strata while still targeting the initial estimand (Léger et al., 2022). However, if not combined with a diagnostic tool such as PoRT, positivity violations can remain unnoticed when doing g-computation. DRS can also extrapolate in the missing strata, although extreme propensity scores remain an issue. And for any such extrapolation to succeed, $Q(A,C)$ needs to be specified correctly (Robins et al., 2007).

Finally, as already mentioned in the beginning, a lack of consistency suggests that the research question of interest must be redefined. And a lack of noninterference—in other words, interference—leads to its own estimands and methods. For example, Tchetgen Tchetgen et al. (2021) provided an extension of g-computation for causal effects on networks of connected units; for a discussion of the use of propensity scores in the presence of interference, see B. Zhang, Hudgens, & Halloran (2023).

Here in the real world

Model misspecification and machine learning.

Beyond nonidentifiability, applied researchers must deal with other sources of bias. For example, to avoid estimation bias, nuisance models must be specified correctly. This means that relevant interactions between predictors must be included, and functional forms need to be correct. This is again a step for which expert knowledge is helpful, but here, machine-learning approaches also hold some promise (Le Borgne et al., 2021; Pirracchio et al., 2015). There is no theoretical proof that one can simply combine machine learning with bootstrapping to arrive at valid inferences; therefore, some doubly robust estimators have been specifically designed to incorporate machine-learning approaches. Among these are the augmented-IPW (Glynn & Quinn, 2010), targeted maximum likelihood estimator (also known as targeted minimum loss-based estimator; van der Laan & Rose, 2011), the collaborative targeted maximum likelihood estimator (van der Laan & Gruber, 2010), and the double/debiased machine learning (Chernozhukov et al., 2018). These estimators should be viewed as complete frameworks for causal inference and are built specifically for the estimation problem at hand. Their implementation is far more complex than, for example, DRS, and requires knowledge about semiparametric

Box 7. Alternative Forms of Inverse Probability Weighting

Inverse probability weighting (IPW) can be used to balance the sample in various ways. The IPW recipe illustrated in Box 4 explains how to balance on the action and is sometimes called IPTW, where T means treatment.

The same recipe can be applied for dealing with missing data; here, the dependent variable is an indicator of missingness, and the controls are causes of missingness. This is known as IPMW, with M for “missing.” This approach can balance the sample on the complete cases and remove selection bias introduced by the missing data. See L. Li et al. (2013) for a review.

IPW can also be used when the sample does not fit the target population; this is known as IPSW, with S for “sampling.” This method balances the sample on the characteristics of the target population, for instance to account for the shift of the covariate distribution because of sampling. The model here includes an indicator of being eligible in the study population as the dependent variable and the confounders and effect modifiers as controls. For example, Colnet et al. (2024) discussed IPSW for generalizing the results of a randomized trials.

A last use of IPW is to take into account attrition because of dropout (sometimes called censoring, hence the name IPCW). Here, the dependent variable is continued participation, and variables from the past are used as covariates to obtain the weights. This allows one to effectively “inflate” underrepresented subjects. If the posited model is correctly specified, one can recover the associations that would have been observed if all subjects had stayed in the study (Huber, 2012).

All these approaches (IPTW, IPMW, IPSW, and IPCW) require the identifiability assumptions stated in the main text, that is, consistency, exchangeability, and positivity. However, these assumptions need to hold for the outcome and the set of controls used, resulting in different sets of assumptions when spelled out. For instance, IPSW requires as positivity assumption that all individuals had a nonextreme probability to be eligible in the study.

Finally, these approaches can be included in doubly robust standardization or even combined with each other by multiplying the resulting weights. The resulting pseudo-sample could thus be balanced in several aspects as long as all the identifiability assumptions are respected and the posited models are well specified. A notable exception is when control values are missing and the missingness and action are not independent; here, the IPMW procedure should be done first and the resulting pseudo-sample is used for computing any other weights (Ross et al., 2022). In practice, we often face a trade-off between balancing one side against another. Again, there is no free lunch.

estimation theory (Díaz, 2020). Returning to our cake comparison somewhat belatedly, these estimators belong to the realm of haute cuisine.

Missing data. One of the most common issues in practice is missing data. Multiple introductions to the different types of missing data and how to deal with them can be found in the literature (e.g., Hayes & Enders, 2023; for graphical representation of missing data problems, which highlight the causal nature of the resulting inferential problems, see Thoemmes & Mohan, 2015), so here we only briefly touch on the topic with a special focus on the estimators introduced earlier. The most “convenient” approach to missing data involves simply tossing away any incomplete observations, which results in so-called complete case analysis. In some types of models, this can result in unbiased results as long as the chance of being a complete case does not depend on the outcome after taking covariates into consideration (Hughes et al., 2019). But it is still generally discouraged because first, it inadvertently targets a complete-cases population that differs from the intended target population (decreased external validity), and second, even for this complete-cases population, effect estimates can be biased because the missingness can induce

new noncausal associations (decreased internal validity). Mathur (2023) provided sensitivity analyses to gauge how sensitive estimates from complete-cases analyses are in different situations.

Another way to handle missing values is multiple imputation, which uses observed variables to create multiple plausible imputations of the missing values. These imputed data sets are then analyzed, and the results are pooled across them. When the missingness of the outcome can be explained by controls that have been measured, multiple imputation can result in unbiased estimates. However, there are also scenarios in which multiple imputation may yield biased results, for example, in the presence of effect modification for propensity-score-based methods (Choi et al., 2019), and there may even be scenarios in which it performs worse than complete case analysis with adequate controls (Hughes et al., 2019). There is also a lack of literature on how to combine multiple imputation with g-computation, and multiple imputation can become particularly time-consuming when combined with bootstrapping or machine-learning approaches. Alternatively, one can fit $Q(A,C)$ on the complete cases only but then use it to predict the potential outcomes for all individuals (Bregier

et al., 2020; Westreich et al., 2015; for an extension in longitudinal settings, see Bartlett et al., 2023). Another approach involves weighting—here, weights are created that are inverse to the probability of missingness, and these are applied much in the same ways as IPW weights (for an overview of the alternative use of IPW, see Box 7). For DRS, all the approaches mentioned here can be employed or even combined.

Considering missing values on the controls, we suggest it is possible to add a missingness indicator among the controls or to apply specific schemes of multiple imputation (Blake et al., 2020; Leyrat et al., 2019; J. Zhang, Dashti, et al., 2023). Furthermore, some machine-learning approaches (e.g., random forest; Strobl et al., 2009) have in-built approaches to handle missing controls.

Measurement error. Another source of bias is measurement errors. How measurement error affects results depends on the underlying causal net, that is, on what causes the deviation between the true value and the observed value of a variable (Hernán & Cole, 2009; van Bork et al., 2022). But in almost all scenarios, measurement error will introduce bias. Thus, high-quality data are crucial to minimize the risk of such biases upfront. This is a particular concern for psychological constructs because reliability may often be modest; for example, failing to account for measurement error in confounding constructs can lead to high rates of mistaken conclusions (Westfall & Yarkoni, 2016).

Causal approaches to correct measurement errors have mainly been developed for propensity-score-based methods. First, considering measurement error in controls, we found that in the epidemiological literature, Rudolph and Stuart (2018) reviewed three ways to deal with an error-prone control for various measurement-error structures using existing sensitivity analyses; in the psychometrics literature, Hong et al. (2017) suggested a Bayesian approach. Blackwell et al. (2017) suggested a multiple-imputation-like approach, called “multiple overputation,” to handle multiple error-prone controls. In general, if the true values of the mismeasured controls are strongly correlated, this will reduce the bias of the estimated effects. However, if the measurement errors of the controls are correlated, this will actually render the bias worse (Hong et al., 2019). Second, considering measurement error in the action, this can be handled through an instrumental-variable procedure (Gustafson, 2007), a regression-calibration-based adjustment (Wu et al., 2019), or a two-step estimation process relying on validation data (Braun et al., 2017). Third, Shu and Yi (2019) discussed causal estimation with an error-prone continuous or binary outcome. In contrast, the literature on g-computation with measurement errors is much more scarce (Blette, 2021); Shu and Yi (2019) proposed a doubly robust estimator.

In psychology, latent variable modeling is the predominant approach to take into account measurement error, and there have been various efforts to explicitly apply it to causal inference. Structural equation modeling (SEM) in particular was originally developed for causal inference (Pearl, 2012), and there have been newer efforts to use SEM to estimate conditional and average effects, taking into account both latent controls and latent outcomes (Mayer et al., 2016). But other ways to combine latent variable modeling and causal inference have also been explored; for example, Lanza et al. (2016) combined IPW with latent class modeling to estimate the effects of depression on substance use (conceived as a latent class). Note that whether or not latent variable modeling “solves” the problem of measurement error crucially depends on whether the assumed measurement model is correct—for example, if a common factor model is mistakenly assumed, the bias that is introduced may sometimes be worse than the measurement bias that is supposed to be removed (Rhemtulla et al., 2020). There have been fairly recent efforts to think about measurement from the viewpoint of causality, both within psychology (van Bork et al., 2022) and within epidemiology (Hernán & Cole, 2009; VanderWeele, 2022), highlighting how this is an area of active conceptual development.

Quantifying the magnitude of our errors. Data and models are, of course, never perfect, and thus, some bias is inevitable. Here, the epidemiological framework of quantitative bias analysis is helpful (for an introduction and best practices, see Lash et al., 2014), which tries to gauge the direction and magnitude of one’s errors. Another helpful framework is the so-called target-trial framework, which spells out an idealized experiment that, in turn, can guide both study planning and data analysis. Bulbulia (2023) provided an illustration of this framework in psychology, trying to answer the question of whether religious service attendance reduces anxiety. Many other sources of bias and ways to avoid them are summarized in Wulff et al. (2023).

Outlook: Other Cakes to Bake

The estimators we have introduced are quite versatile and can be extended in various ways. For example, our focus has been on internal validity (correctness of the results for the targeted population); however, one could also be interested in questions of external validity. This may involve questions about generalizability (is the effect estimate valid for a broader population?) and transportability (can we draw conclusions about the causal effect in different settings or for different populations?; see also Deffner et al., 2022). All the estimators

we presented here can be used to address such research questions if they are applied within a transportability framework (Lesko et al., 2017).

Furthermore, here we were interested in the occurrence of the outcome at a specified time point, such as abstinence after 1 year. But we may also be interested in the outcome's occurrence in time; for example, we may ask whether AA attendance has an effect on the timing of a relapse. Chatton et al. (2022) discussed the particularity of the estimands in this context and proposed an extension of the estimators presented here.

Our focus was on marginal effects that average over groups of people. But sometimes other estimands may be more relevant—for example, one may be interested in heterogeneous causal effects (Bryan et al., 2021) or may want to disentangle indirect and direct effects in the context of mediation analysis. Pösch (2021) illustrated the use of g-computation in this context.

Finally, going beyond cross-sectional data, in a longitudinal setting, both the action and confounders may vary over time. One common issue in this context is confounder-action feedback. For example, imagine we had monthly data spanning 1 year and were interested in the effect of attending AA every month, as opposed to never, on abstinence at the end of the year. Attending AA in a given month may affect subsequent social isolation, and social isolation may, in turn, affect both subsequent AA attendance and abstinence. Thus, social isolation is both an outcome of the action and a confounder, which leaves us in a bad spot: If we statistically adjust for it, we may accidentally induce collider bias; if we do not statistically adjust for it, we are stuck with confounding bias. Traditional methods fail to handle such confounder-action feedback, and so we need the longitudinal extension of the estimators presented here (Hernán & Robins, 2020; for an introduction to g-computation in a longitudinal context for psychologists, see Loh & Ren, 2023c. Two recent articles targeting the psychological community introduced g-estimation (another estimator from epidemiological literature), which is another valid approach for this specific setting (Loh & Ren, 2023b, 2023d).

Conclusion

In this article, we have provided recipes for causal estimators in the presence of time-fixed confounding. A companion R notebook illustrating the implementation of these estimators is available at github.com/ArthurChatton/CausalCookbook. We focused on estimators commonly used in epidemiological literature rather than in psychology to bridge the gap between these disciplines and to broaden psychologists' causal-inference toolbox. Epidemiology is, of course, not the only field

with a strong focus on causal inference. For example, methods from economics can be another valuable addition; in particular, those estimators that do not rely on conditional exchangeability but make other assumptions about the underlying causal net that may sometimes be more palatable (Grosz et al., 2024; Kim & Steiner, 2016).

All estimators, like cake recipes, require good ingredients—no statistical method can overcome poor data. And a lot of effort may be wasted if one sets out to bake the wrong cake—no statistical method can overcome poor research questions. Finally, in causal inference (and elsewhere), there is no free cake: Different approaches make different trade-offs with respect to bias and variance but also with respect to the underlying assumptions. The availability of a large set of estimators—based on different assumptions but targeting the same or at least related estimands—is crucial to improve evidence from observational studies through triangulation (Munafò & Davey Smith, 2018).

Transparency

Action Editor: David A. Sbarra

Editor: David A. Sbarra

Author Contributions

Arthur Chatton: Conceptualization; Visualization; Writing – original draft; Writing – review & editing.

Julia M. Rohrer: Conceptualization; Visualization; Writing – review & editing.

Declaration of Conflicting Interests


The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

A. Chatton is supported by an Institute for Data Valorization (IVADO) postdoctoral fellowship (2022-7820036733).

ORCID iDs

Arthur Chatton  <https://orcid.org/0000-0002-0018-5899>

Julia M. Rohrer  <https://orcid.org/0000-0001-8564-4523>

Acknowledgments

We thank Saloni Dattani, Stefan C. Schmukle, C. Kronenberg, and R. Furasoli for their helpful feedback on a previous version of this article. Figure 2 was made possible by Midjourney; thereby, no cakes were harmed during the making of this article (albeit some were eaten). The idea behind Figure 2 and the cooking metaphor came from a Twitter meme likely initiated by Simon Grund. This article is loosely based on the second chapter of A. Chatton's PhD thesis (www.theses.fr/2021NANT4062, in French). A preprint of this article was published on <https://osf.io/preprints/psyarxiv/k2gzp>.

Notes

1. In contrast, model identification in the context of structural equation modeling focuses on whether it is possible to calculate

unique parameter values from the observed data. A parallel between the two concepts is that without identification, there is no unique “solution.” In a structural equation model that is not identified, multiple combinations of parameters are compatible with the observed data. In causal inference, without causal identification, different causal effects are compatible with the observed data; for example, if there is no causal identification because of an unobserved confounder, a positive association between the variables of interest can be compatible with a positive causal effect but also with no effect at all or even a negative effect (a phenomenon known as Simpson/reversal paradox; Messick & Van de Geer, 1981).

2. The estimators we introduce in this article can be applied to randomized experiments to deal with such problems and also to generally increase the precision of effect estimates by using the information provided by covariates.

3. Some causal estimators may need a different set of identifiability assumptions. For instance, instrumental-variable-based methods replace the assumption of conditional exchangeability with other assumptions, most notably that the instrumental variable is related to the outcome only via the action of interest (Grosz et al., 2024; Kim & Steiner, 2016).

4. Here, it may seem logical to include instruments (see Box 2) because they predict the action; however, this can yield random positivity violations and thus increase bias and variance (Pearl, 2011).

5. Such models are commonly used in economics under the label “linear probability model”; see Gomila (2021) for an introduction for psychologists.

6. To simplify notation, we assume C includes only categorical controls; in practice, C may include continuous covariates in which case the equations would be rewritten as integrals.

7. It is thus closely related to Pearl’s (2009) do-operator, which is a means to express that we place ourselves in such worlds. In Pearl’s framework, the g -formula is usually referred to as “truncated factorization formula.”

8. In the time-fixed setting, g -computation is also referred to as “regression standardization,” “ g -standardization,” or “S(ingle)-learner.” In the time-varying setting, the term “parametric g -formula” is used almost exclusively.

9. In practice, chances to correctly specify even just one of the models may be close to zero. But even then, doubly robust estimators allow one to use machine learning to relax parametric assumptions about the data-generating mechanism (discussed below).

References

- Ali, M. S., Groenwold, R. H. H., Belitser, S. V., Pestman, W. R., Hoes, A. W., Roes, K. C. B., de Boer, A., & Klungel, O. H. (2015). Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: A systematic review. *Journal of Clinical Epidemiology*, 68(2), 122–131. <https://doi.org/10.1016/j.jclinepi.2014.08.011>
- Ali, M. S., Groenwold, R. H. H., Pestman, W. R., Belitser, S. V., Roes, K. C. B., Hoes, A. W., de Boer, A., & Klungel, O. H. (2014). Propensity score balance measures in pharmacoepidemiology: A simulation study. *Pharmacoepidemiology and Drug Safety*, 23(8), 802–811. <https://doi.org/10.1002/pds.3574>
- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3), 399–424. <https://doi.org/10.1080/00273171.2011.568786>
- Austin, P. C. (2016). Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine*, 35(30), 5642–5655. <https://doi.org/10.1002/sim.7084>
- Austin, P. C. (2018). Assessing the performance of the generalized propensity score for estimating the effect of quantitative or continuous exposures on binary outcomes. *Statistics in Medicine*, 37(11), 1874–1894. <https://doi.org/10.1002/sim.7615>
- Austin, P. C. (2019). Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures. *Statistical Methods in Medical Research*, 28(5), 1365–1377. <https://doi.org/10.1177/0962280218756159>
- Austin, P. C. (2022). Bootstrap vs asymptotic variance estimation when using propensity score weighting with continuous and binary outcomes. *Statistics in Medicine*, 41(22), 4426–4443. <https://doi.org/10.1002/sim.9519>
- Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28), 3661–3679. <https://doi.org/10.1002/sim.6607>
- Bartlett, J. W., Parra, C. O., Granger, E., Keogh, R. H., van Zwet, E. W., & Daniel, R. M. (2023). *G-formula for causal inference via multiple imputation*. arXiv. <https://doi.org/10.48550/arXiv.2301.12026>
- Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2), 802–837. <https://doi.org/10.1214/12-AOS1077>
- Bernerth, J. B., & Aguinis, H. (2016). A critical review and best-practice recommendations for control variable usage. *Personnel Psychology*, 69(1), 229–283. <https://doi.org/10.1111/peps.12103>
- Blackwell, M., Honaker, J., & King, G. (2017). A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods & Research*, 46(3), 303–341. <https://doi.org/10.1177/0049124115585360>
- Blake, H. A., Leyrat, C., Mansfield, K. E., Seaman, S., Tomlinson, L., Carpenter, J., & Williamson, E. (2020). Propensity scores using missingness pattern information: A practical guide. *Statistics in Medicine*, 39(11), 1641–1657. <https://doi.org/10.1002/sim.8503>
- Blette, B. (2021). *Causal inference for error-prone exposures* [Doctoral dissertation, University of North Carolina at Chapel Hill]. Carolina Digital Repository. <https://doi.org/10.17615/b63q-er23>
- Braun, D., Gorfine, M., Parmigiani, G., Arvold, N. D., Dominici, F., & Zigler, C. (2017). Propensity scores with misclassified treatment assignment: A likelihood-based adjustment. *Biostatistics*, 18(4), 695–710. <https://doi.org/10.1093/biostatistics/kxx014>
- Breger, T. L., Edwards, J. K., Cole, S. R., Westreich, D., Pence, B. W., & Adimora, A. A. (2020). Two-stage g -computation: Evaluating treatment and intervention impacts in observational cohorts when exposure information is partly missing. *Epidemiology*, 31(5), 695–703. <https://doi.org/10.1097/EDE.0000000000001233>

- Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, 5(8), 980–989. <https://doi.org/10.1038/s41562-021-01143-3>
- Bulbulia, J. A. (2023). A workflow for causal inference in cross-cultural psychology. *Religion, Brain & Behavior*, 13(3), 291–306. <https://doi.org/10.1080/2153599X.2022.2070245>
- Chan, G. C. K., Lim, C., Sun, T., Stjepanovic, D., Connor, J., Hall, W., & Leung, J. (2022). Causal inference with observational data in addiction research. *Addiction*, 117(10), 2736–2744. <https://doi.org/10.1111/add.15972>
- Chatton, A., Le Borgne, F., Leyrat, C., & Foucher, Y. (2022). G-computation and doubly robust standardisation for continuous-time data: A comparison with inverse probability weighting. *Statistical Methods in Medical Research*, 31(4), 706–718. <https://doi.org/10.1177/09622802211047345>
- Chatton, A., Le Borgne, F., Leyrat, C., Gillaizeau, F., Rousseau, C., Barbin, L., Laplaud, D., Léger, M., Giraudeau, B., & Foucher, Y. (2020). G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: A comparative simulation study. *Scientific Reports*, 10(1), Article 9219. <https://doi.org/10.1038/s41598-020-65917-x>
- 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), C1–C68. <https://doi.org/10.1111/ectj.12097>
- Choi, J., Dekkers, O. M., & le Cessie, S. (2019). A comparison of different methods to handle missing data in the context of propensity score analysis. *European Journal of Epidemiology*, 34(1), 23–36. <https://doi.org/10.1007/s10654-018-0447-z>
- Cinelli, C., Forney, A., & Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research*. Advance online publication. <https://doi.org/10.1177/00491241221099552>
- Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., & Yang, S. (2024). Causal inference methods for combining randomized trials and observational studies: A review. *Statistical Science*, 39(1), 165–191.
- Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press.
- Danelian, G., Foucher, Y., Léger, M., Le Borgne, F., & Chatton, A. (2023). Identification of in-sample positivity violations using regression trees: The PoRT algorithm. *Journal of Causal Inference*, 11(1), Article 20220032. <https://doi.org/10.1515/jci-2022-0032>
- Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2–21. <https://doi.org/10.1016/j.socscimed.2017.12.005>
- Deffner, D., Rohrer, J. M., & McElreath, R. (2022). A causal framework for cross-cultural generalizability. *Advances in Methods and Practices in Psychological Science*, 5(3). <https://doi.org/10.1177/25152459221106366>
- Díaz, I. (2020). Machine learning in the estimation of causal effects: Targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 21(2), 353–358. <https://doi.org/10.1093/biostatistics/kxz042>
- D’Onofrio, B. M., Sjölander, A., Lahey, B. B., Lichtenstein, P., & Öberg, A. S. (2020). Accounting for confounding in observational studies. *Annual Review of Clinical Psychology*, 16, 25–48. <https://doi.org/10.1146/annurev-clinpsy-032816-045030>
- Elwert, F. (2013). Graphical causal models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 245–273). Springer. https://doi.org/10.1007/978-94-007-6094-3_13
- Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., & Saria, S. (2021). The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3), 283–286. <https://doi.org/10.1056/NEJMc2104626>
- Fong, C., Hazlett, C., & Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1), 156–177. <https://doi.org/10.1214/17-AOAS1101>
- Foster, E. M. (2010). Causal inference and developmental psychology. *Developmental Psychology*, 46(6), 1454–1480. <https://doi.org/10.1037/a0020204>
- Franklin, J. M., Rassen, J. A., Ackermann, D., Bartels, D. B., & Schneeweiss, S. (2014). Metrics for covariate balance in cohort studies of causal effects. *Statistics in Medicine*, 33(10), 1685–1699. <https://doi.org/10.1002/sim.6058>
- Geiger, D., Verma, T., & Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks*, 20(5), 507–534. <https://doi.org/10.1002/net.3230200504>
- Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1), 36–56. <https://doi.org/10.1093/pan/mpp036>
- Gomila, R. (2021). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, 150(4), 700–709. <https://doi.org/10.1037/xge0000920>
- Greifer, N., & Stuart, E. A. (2021). *Choosing the estimand when matching or weighting in observational studies*. arXiv. <http://arxiv.org/abs/2106.10577>
- Grosz, M. P., Ayaita, A., Arslan, R. C., Buecker, S., Ebert, T., Hünermund, P., Müller, S., Rieger, S., Zapko-Willmes, A., & Rohrer, J. M. (2024). Natural experiments: Missed opportunities for causal inference in psychology. *Advances in Methods and Practices in Psychological Science*, 7(1). <https://doi.org/10.1177/25152459231218610>
- Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, 15(5), 1243–1255. <https://doi.org/10.1177/1745691620921521>
- Gruber, S., Phillips, R. V., Lee, H., & van der Laan, M. J. (2022). Data-adaptive selection of the propensity score truncation level for inverse-probability-weighted and targeted maximum likelihood estimators of marginal point treatment effects. *American Journal of Epidemiology*, 191(9), 1640–1651. <https://doi.org/10.1093/aje/kwac087>
- Gustafson, P. (2007). Measurement error modelling with an approximate instrumental variable. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(5), 797–815. <https://doi.org/10.1111/j.1467-9868.2007.00611.x>

- Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, 15(3), 234–249. <https://doi.org/10.1037/a0019623>
- Hayes, T., & Enders, C. K. (2023). Maximum likelihood and multiple imputation missing data handling: How they work, and how to make them work in practice. In H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Data analysis and research publication* (2nd ed., Vol. 3, pp. 27–51). American Psychological Association. <https://doi.org/10.1037/0000320-002>
- 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., & Cole, S. R. (2009). Invited commentary: Causal diagrams and measurement bias. *American Journal of Epidemiology*, 170(8), 959–962; discussion 963–964. <https://doi.org/10.1093/aje/kwp293>
- 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. (2020). *Causal inference: What if?* Chapman & Hall/CRC.
- Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. In A. Gelman & X.-L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 73–84). John Wiley & Sons. <https://doi.org/10.1002/0470090456.ch7>
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960. <https://doi.org/10.1080/01621459.1986.10478354>
- Hong, H., Aaby, D. A., Siddique, J., & Stuart, E. A. (2019). Propensity score-based estimators with multiple error-prone covariates. *American Journal of Epidemiology*, 188(1), 222–230. <https://doi.org/10.1093/aje/kwy210>
- Hong, H., Rudolph, K. E., & Stuart, E. A. (2017). Bayesian approach for addressing differential covariate measurement error in propensity score methods. *Psychometrika*, 82(4), 1078–1096. <https://doi.org/10.1007/s11336-016-9533-x>
- Huber, M. (2012). Identification of average treatment effects in social experiments under alternative forms of attrition. *Journal of Educational and Behavioral Statistics*, 37(3), 443–474. <https://doi.org/10.3102/1076998611411917>
- Hughes, R. A., Heron, J., Sterne, J. A. C., & Tilling, K. (2019). Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. *International Journal of Epidemiology*, 48(4), 1294–1304. <https://doi.org/10.1093/ije/dyz032>
- Hünemund, P., & Louw, B. (2023). On the nuisance of control variables in causal regression analysis. *Organizational Research Methods*. Advance online publication. <https://doi.org/10.1177/10944281231219274>
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4), 309–334. <https://doi.org/10.1037/a0020761>
- Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 171(2), 481–502. <https://doi.org/10.1111/j.1467-985X.2007.00527.x>
- Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467), 854–866. <https://doi.org/10.1198/016214504000001187>
- Jamison, J. C. (2019). The entry of randomized assignment into the social sciences. *Journal of Causal Inference*, 7(1). <https://doi.org/10.1515/jci-2017-0025>
- Kahan, B. C., Jairath, V., Doré, C. J., & Morris, T. P. (2014). The risks and rewards of covariate adjustment in randomized trials: An assessment of 12 outcomes from 8 studies. *Trials*, 15(1), Article 139. <https://doi.org/10.1186/1745-6215-15-139>
- Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4), 523–539. <https://doi.org/10.1214/07-STS227>
- Keele, L., Stevenson, R. T., & Elwert, F. (2020). The causal interpretation of estimated associations in regression models. *Political Science Research and Methods*, 8(1), 1–13. <https://doi.org/10.1017/psrm.2019.31>
- Keiding, N., & Clayton, D. (2014). Standardization and control for confounding in observational studies: A historical perspective. *Statistical Science*, 29(4), 529–558. <https://doi.org/10.1214/13-STS453>
- Keil, A. P., Daza, E. J., Engel, S. M., Buckley, J. P., & Edwards, J. K. (2018). A Bayesian approach to the g-formula. *Statistical Methods in Medical Research*, 27(10), 3183–3204. <https://doi.org/10.1177/0962280217694665>
- Keil, A. P., Mooney, S. J., Jonsson Funk, M., Cole, S. R., Edwards, J. K., & Westreich, D. (2018). Resolving an apparent paradox in doubly robust estimators. *American Journal of Epidemiology*, 187(4), 891–892. <https://doi.org/10.1093/aje/kwx385>
- Kim, Y., & Steiner, P. (2016). Quasi-experimental designs for causal inference. *Educational Psychologist*, 51(3–4), 395–405. <https://doi.org/10.1080/00461520.2016.1207177>
- King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, 27(4), 434–454. <https://doi.org/10.1017/pan.2019.11>
- Kohler, U., Class, F., & Sawert, T. (2023). Control variable selection in applied quantitative sociology: A critical review. *European Sociological Review*, 40(1), 173–186. <https://doi.org/10.1093/esr/jcac078>
- Kunicki, Z. J., Smith, M. L., & Murray, E. J. (2023). A primer on structural equation model diagrams and directed acyclic graphs: When and how to use each in psychological and epidemiological research. *Advances in Methods and Practices in Psychological Science*, 6(2). <https://doi.org/10.1177/25152459231156085>
- Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences, USA*, 116(10), 4156–4165. <https://doi.org/10.1073/pnas.1804597116>
- Lanza, S. T., Moore, J. E., & Butera, N. M. (2013). Drawing causal inferences using propensity scores: A practical guide for community psychologists. *American Journal of Community Psychology*, 52(3–4), 380–392. <https://doi.org/10.1007/s10464-013-9604-4>

- Lanza, S. T., Schuler, M. S., & Bray, B. C. (2016). Latent class analysis with causal inference: The effect of adolescent depression on young adult substance use profile. In W. Wiedermann & A. von Eye (Eds.), *Statistics and causality* (pp. 385–404). John Wiley & Sons. <https://doi.org/10.1002/9781118947074.ch16>
- Lash, T. L., Fox, M. P., MacLehose, R. F., Maldonado, G., McCandless, L. C., & Greenland, S. (2014). Good practices for quantitative bias analysis. *International Journal of Epidemiology*, 43(6), 1969–1985. <https://doi.org/10.1093/ije/dyu149>
- Le Borgne, F., Chatton, A., Léger, M., Lenain, R., & Foucher, Y. (2021). G-computation and machine learning for estimating the causal effects of binary exposure statuses on binary outcomes. *Scientific Reports*, 11(1), Article 1435. <https://doi.org/10.1038/s41598-021-81110-0>
- Léger, M., Chatton, A., Le Borgne, F., Pirracchio, R., Lasocki, S., & Foucher, Y. (2022). Causal inference in case of near-violation of positivity: Comparison of methods. *Biometrical Journal*, 64(8), 1389–1403. <https://doi.org/10.1002/bimj.202000323>
- Lesko, C. R., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G., & Cole, S. R. (2017). Generalizing study results: A potential outcomes perspective. *Epidemiology*, 28(4), 553–561. <https://doi.org/10.1097/EDE.0000000000000664>
- Leyrat, C., Seaman, S. R., White, I. R., Douglas, I., Smeeth, L., Kim, J., Resche-Rigon, M., Carpenter, J. R., & Williamson, E. J. (2019). Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical Methods in Medical Research*, 28(1), 3–19. <https://doi.org/10.1177/0962280217713032>
- Li, F., Thomas, L. E., & Li, F. (2019). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, 188(1), 250–257. <https://doi.org/10.1093/aje/kwy201>
- Li, L., Shen, C., Li, X., & Robins, J. M. (2013). On weighting approaches for missing data. *Statistical Methods in Medical Research*, 22(1), 14–30. <https://doi.org/10.1177/0962280211403597>
- Loh, W. W., & Ren, D. (2022). Estimating social influence in a social network using potential outcomes. *Psychological Methods*, 27(5), 841–855. <https://doi.org/10.1037/met0000356>
- Loh, W. W., & Ren, D. (2023a). Data-driven covariate selection for confounding adjustment by focusing on the stability of the effect estimator. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000564>
- Loh, W. W., & Ren, D. (2023b). Estimating time-varying treatment effects in longitudinal studies. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000574>
- Loh, W. W., & Ren, D. (2023c). *G-formula: What it is, why it matters, and how to implement it in lavaan*. PsyArXiv. <https://doi.org/10.31234/osf.io/m37uc>
- Loh, W. W., & Ren, D. (2023d). A tutorial on causal inference in longitudinal data with time-varying confounding using g-estimation. *Advances in Methods and Practices in Psychological Science*, 6(3). <https://doi.org/10.1177/25152459231174029>
- Lucas, R. E. (2023). Why the cross-lagged panel model is almost never the right choice. *Advances in Methods and Practices in Psychological Science*, 6(1). <https://doi.org/10.1177/25152459231158378>
- 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>
- Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review*, 86(3), 532–565. <https://doi.org/10.1177/00031224211004187>
- Mathur, M. B. (2023). The M-value: A simple sensitivity analysis for bias due to missing data in treatment effect estimates. *American Journal of Epidemiology*, 192(4), 61–620. <https://doi.org/10.1093/aje/kwac207>
- Mayer, A., Dietzfelbinger, L., Rosseel, Y., & Steyer, R. (2016). The EffectLiteR approach for analyzing average and conditional effects. *Multivariate Behavioral Research*, 51(2–3), 374–391. <https://doi.org/10.1080/00273171.2016.1151334>
- McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19), 3388–3414. <https://doi.org/10.1002/sim.5753>
- McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4), 403–425. <https://doi.org/10.1037/1082-989X.9.4.403>
- Messick, D. M., & Van de Geer, J. P. (1981). A reversal paradox. *Psychological Bulletin*, 90(3), 582–593. <https://doi.org/10.1037/0033-2909.90.3.582>
- Morris, T. P., Walker, A. S., Williamson, E. J., & White, I. R. (2022). Planning a method for covariate adjustment in individually randomised trials: A practical guide. *Trials*, 23(1), Article 328. <https://doi.org/10.1186/s13063-022-06097-z>
- Munafò, M. R., & Davey Smith, G. (2018). Robust research needs many lines of evidence. *Nature*, 553(7689), 399–401. <https://doi.org/10.1038/d41586-018-01023-3>
- Neison, F. G. P. (1844). On a method recently proposed for conducting inquiries into the comparative sanatory condition of various districts, with illustrations, derived from numerous places in Great Britain at the period of the last census. *Journal of the Statistical Society of London*, 7(1), 40–68. <https://doi.org/10.2307/2337745>
- Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles. *Annals of Agricultural Sciences*, 5, 465–480.
- Nguyen, T. Q., Schmid, I., & Stuart, E. A. (2020). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000299>
- Pargent, F., Schoedel, R., & Stachl, C. (2023). Best practices in supervised machine learning: A tutorial for psychologists. *Advances in Methods and Practices in Psychological Science*, 6(3). <https://doi.org/10.1177/25152459231162559>

- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4), 669–688. <https://doi.org/10.2307/2337329>
- Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge University Press.
- Pearl, J. (2011). Invited commentary: Understanding bias amplification. *American Journal of Epidemiology*, 174(11), 1223–1227. <https://doi.org/10.1093/aje/kwr352>
- Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 68–91). The Guilford Press.
- Pirracchio, R., Petersen, M. L., & van der Laan, M. (2015). Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology*, 181(2), 108–119. <https://doi.org/10.1093/aje/kwu253>
- Pösch, K. (2021). Testing complex social theories with causal mediation analysis and g-computation: Toward a better way to do causal structural equation modeling. *Sociological Methods & Research*, 50(3), 1376–1406. <https://doi.org/10.1177/0049124119826159>
- Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, 25(1), 30–45. <https://doi.org/10.1037/met0000220>
- Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9), 1393–1512. [https://doi.org/10.1016/0270-0255\(86\)90088-6](https://doi.org/10.1016/0270-0255(86)90088-6)
- Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 550–560. <https://doi.org/10.1097/00001648-200009000-00011>
- Robins, J. M., Sued, M., Lei-Gomez, Q., & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when “inverse probability” weights are highly variable. *Statistical Science*, 22(4), 544–559. <https://doi.org/10.1214/07-STS227D>
- Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27–42. <https://doi.org/10.1177/2515245917745629>
- Rohrer, J. M., Hünemund, P., Arslan, R. C., & Elson, M. (2022). That's a lot to process! Pitfalls of popular path models. *Advances in Methods and Practices in Psychological Science*, 5(2). <https://doi.org/10.1177/25152459221095827>
- Rohrer, J. M., Keller, T., & Elwert, F. (2021). Proximity can induce diverse friendships: A large randomized classroom experiment. *PLOS ONE*, 16(8), Article e0255097. <https://doi.org/10.1371/journal.pone.0255097>
- Rohrer, J. M., & Murayama, K. (2023). These are not the effects you are looking for: Causality and the within-/between-person distinction in longitudinal data analysis. *Advances in Methods and Practices in Psychological Science*, 6(1). <https://doi.org/10.1177/25152459221140842>
- 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>
- Ross, R. K., Breskin, A., Breger, T. L., & Westreich, D. (2022). Reflection on modern methods: Combining weights for confounding and missing data. *International Journal of Epidemiology*, 51(2), 679–684. <https://doi.org/10.1093/ije/dyab205>
- Rousseelet, G. A., Pernet, C. R., & Wilcox, R. R. (2021). The percentile bootstrap: A primer with step-by-step instructions in R. *Advances in Methods and Practices in Psychological Science*, 4(1). <https://doi.org/10.1177/2515245920911881>
- 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>
- Rudolph, K. E., & Stuart, E. A. (2018). Using sensitivity analyses for unobserved confounding to address covariate measurement error in propensity score methods. *American Journal of Epidemiology*, 187(3), 604–613. <https://doi.org/10.1093/aje/kwx248>
- Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, 13(4), 279–313. <https://doi.org/10.1037/a0014268>
- Schnitzer, M. E., Platt, R. W., & Durand, M. (2020). A tutorial on dealing with time-varying eligibility for treatment: Comparing the risk of major bleeding with direct-acting oral anticoagulants vs warfarin. *Statistics in Medicine*, 39(29), 4538–4550. <https://doi.org/10.1002/sim.8715>
- Senn, S. (2013). Seven myths of randomisation in clinical trials. *Statistics in Medicine*, 32(9), 1439–1450. <https://doi.org/10.1002/sim.5713>
- Shu, D., & Yi, G. Y. (2019). Causal inference with measurement error in outcomes: Bias analysis and estimation methods. *Statistical Methods in Medical Research*, 28(7), 2049–2068. <https://doi.org/10.1177/0962280217743777>
- Snowden, J. M., Rose, S., & Mortimer, K. M. (2011). Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7), 731–738. <https://doi.org/10.1093/aje/kwq472>
- Spertus, J. V., & Normand, S.-L. T. (2018). Bayesian propensity scores for high-dimensional causal inference: A comparison of drug-eluting to bare-metal coronary stents. *Biometrical Journal*, 60(4), 721–733. <https://doi.org/10.1002/bimj.201700305>
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4), 323–348. <https://doi.org/10.1037/a0016973>
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1), 1–21. <https://doi.org/10.1214/09-STS313>
- Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science*, 22(4), 560–568. <https://doi.org/10.1214/07-STS227A>
- Tchetgen Tchetgen, E. J., Fulcher, I. R., & Shpitser, I. (2021). Auto-g-computation of causal effects on a network. *Journal of the American Statistical Association*, 116(534), 833–844. <https://doi.org/10.1080/01621459.2020.1811098>

- Textor, J., Hardt, J., & Knüppel, S. (2011). DAGitty: A graphical tool for analyzing causal diagrams. *Epidemiology*, 22(5), 745. <https://doi.org/10.1097/EDE.0b013e318225c2be>
- Thoemmes, F., & Mohan, K. (2015). Graphical representation of missing data problems. *Structural Equation Modeling*, 22(4), 631–642. <https://doi.org/10.1080/10705511.2014.937378>
- Thoemmes, F., & Ong, A. D. (2016). A primer on inverse probability of treatment weighting and marginal structural models. *Emerging Adulthood*, 4(1), 40–59. <https://doi.org/10.1177/2167696815621645>
- Thoemmes, F., Rosseel, Y., & Textor, J. (2018). Local fit evaluation of structural equation models using graphical criteria. *Psychological Methods*, 23(1), 27–41. <https://doi.org/10.1037/met0000147>
- Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, 46(1), 90–118. <https://doi.org/10.1080/00273171.2011.540475>
- Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, 46(3), 514–543. <https://doi.org/10.1080/00273171.2011.569395>
- van Bork, R., Rhemtulla, M., Sijtsma, K., & Borsboom, D. (2022). A causal theory of error scores. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000521>
- van der Laan, M. J., & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1), Article 17. <https://doi.org/10.2202/1557-4679.1181>
- van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data*. Springer.
- VanderWeele, T. J. (2022). Constructed measures and causal inference: Towards a new model of measurement for psychosocial constructs. *Epidemiology*, 33(1), 141–151. <https://doi.org/10.1097/EDE.0000000000001434>
- Vansteelandt, S., & Daniel, R. M. (2014). On regression adjustment for the propensity score. *Statistics in Medicine*, 33(23), 4053–4072. <https://doi.org/10.1002/sim.6207>
- West, S. G., Cham, H., Thoemmes, F., Renneberg, B., Schulze, J., & Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and application in clinical treatment outcome research. *Journal of Consulting and Clinical Psychology*, 82(5), 906–919. <https://doi.org/10.1037/a0036387>
- West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods*, 15(1), 18–37. <https://doi.org/10.1037/a0015917>
- Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLOS ONE*, 11(3), Article e0152719. <https://doi.org/10.1371/journal.pone.0152719>
- Westreich, D. (2020). *Epidemiology by design: A causal approach to the health sciences*. Oxford University Press.
- Westreich, D., Edwards, J. K., Cole, S. R., Platt, R. W., Mumford, S. L., & Schisterman, E. F. (2015). Imputation approaches for potential outcomes in causal inference. *International Journal of Epidemiology*, 44(5), 1731–1737. <https://doi.org/10.1093/ije/dyv135>
- Westreich, D., & Greenland, S. (2013). The Table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, 177(4), 292–298. <https://doi.org/10.1093/aje/kws412>
- Whitcomb, B. W., & Naimi, A. I. (2021). Defining, quantifying, and interpreting “noncollapsibility” in epidemiologic studies of measures of “effect.” *American Journal of Epidemiology*, 190(5), 697–700. <https://doi.org/10.1093/aje/kwaa267>
- Wu, X., Braun, D., Kioumourtzoglou, M.-A., Choirat, C., Di, Q., & Dominici, F. (2019). Causal inference in the context of an error prone exposure: Air pollution and mortality. *The Annals of Applied Statistics*, 13(1), 520–547. <https://doi.org/10.1214/18-AOAS1206>
- Wulff, J. N., Sajons, G. B., Pogrebn, G., Lonati, S., Bastardo, N., Banks, G. C., & Antonakis, J. (2023). Common methodological mistakes. *The Leadership Quarterly*, 34(1), Article 101677. <https://doi.org/10.1016/j.leaqua.2023.101677>
- Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science*, 5(2). <https://doi.org/10.1177/25152459221095823>
- Xu, S., Ross, C., Raebel, M. A., Shetterly, S., Blanchette, C., & Smith, D. (2010). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health*, 13(2), 273–277. <https://doi.org/10.1111/j.1524-4733.2009.00671.x>
- Ye, Y., & Kaskutas, L. A. (2009). Using propensity scores to adjust for selection bias when assessing the effectiveness of Alcoholics Anonymous in observational studies. *Drug and Alcohol Dependence*, 104(1–2), 56–64. <https://doi.org/10.1016/j.drugalcdep.2009.03.018>
- Zhang, B., Hudgens, M. G., & Halloran, M. E. (2023). Propensity score in the face of interference: Discussion of Rosenbaum and Rubin (1983). *Observational Studies*, 9(1), 125–131. <https://doi.org/10.1353/obs.2023.0013>
- Zhang, J., Dashti, S. G., Carlin, J. B., Lee, K. J., & Moreno-Betancur, M. (2023). Should multiple imputation be stratified by exposure group when estimating causal effects via outcome regression in observational studies? *BMC Medical Research Methodology*, 23(1), Article 42. <https://doi.org/10.1186/s12874-023-01843-6>
- Zhang, X., Stamey, J. D., & Mathur, M. B. (2020). Assessing the impact of unmeasured confounders for credible and reliable real-world evidence. *Pharmacoepidemiology and Drug Safety*, 29(10), 1219–1227. <https://doi.org/10.1002/pds.5117>
- Zhao, S., van Dyk, D. A., & Imai, K. (2020). Propensity score-based methods for causal inference in observational studies with non-binary treatments. *Statistical Methods in Medical Research*, 29(3), 709–727. <https://doi.org/10.1177/0962280219888745>
- Zhu, Y., Hubbard, R. A., Chubak, J., Roy, J., & Mitra, N. (2021). Core concepts in pharmacoepidemiology: Violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches. *Pharmacoepidemiology and Drug Safety*, 30(11), 1471–1485. <https://doi.org/10.1002/pds.5338>
- Zou, G. Y. (2009). Assessment of risks by predicting counterfactuals. *Statistics in Medicine*, 28(30), 3761–3781. <https://doi.org/10.1002/sim.3751>