Causal Inference: Overview

Jennifer Hill, Steinhardt School of Culture, Education, and Human Development, New York University, New York, NY, USA
Elizabeth A Stuart, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

Abstract

This article discusses causal inference in statistics. It describes the theoretical framework and notation needed to formally define causal effects and the assumptions required to identify them nonparametrically. This involves definition of potential outcomes that represent the potential value of the outcome across different treatment exposures. Designs that allow researchers to satisfy or weaken these assumptions are briefly described. Then common parametric assumptions used to model effects and more current approaches that require weaker assumptions are discussed.

Introduction: Causal Inference as a Comparison of Potential Outcomes

Causal inference refers to an intellectual discipline that considers the assumptions, study designs, and estimation strategies that allow researchers to draw causal conclusions based on data. As detailed below, the term 'causal conclusion' used here refers to a conclusion regarding the effect of a causal variable (often referred to as the 'treatment' under a broad conception of the word) on some outcome(s) of interest. The dominant perspective on causal inference in statistics has philosophical underpinnings that rely on consideration of counterfactual states. In particular, it considers the outcomes that could manifest given exposure to each of a set of treatment conditions. Causal effects are defined as comparisons between these 'potential outcomes.' For instance, the causal effect of a drug on systolic blood pressure 1 month after the drug regime has begun (vs no exposure to the drug) would be defined as a comparison of systolic blood pressure that would be measured at this time given exposure to the drug with the systolic blood pressure that would be measured at the same point in time in the absence of exposure to the drug. The challenge for causal inference is that we are not generally able to observe both of these states: at the point in time when we are measuring the outcomes, each individual either has had drug exposure or has not.

Notation for Potential Outcomes and Causal Effects

There is a long history of philosophers, statisticians, and economists thinking about causality and how to establish causal relationships between variables. This philosophical discussion continues to this day, with many different perspectives still being expressed (see, e.g., Pearl, 2009; Dowd, 2011). Statisticians have generally focused on the investigation of the extent to which a given 'cause' leads to a change in some outcome(s) of interest. This is the framework detailed in this article. More general discussions of causal relationships, such as those investigated through structural equation models or path analysis, or those that attempt to determine a (or the) cause of an effect, typically require much stronger assumptions than do approaches that focus on estimating specific causal effects (Holland, 1986; VanderWeele, 2012).

To formalize assumptions and models in this framework, traditional statistical notation has been extended to explicitly define the 'potential outcomes' described above (Rubin, 1978). For instance, for an individual, i, and a binary treatment, Z_i, two outcomes can be defined that have the potential to be observed at a given time point, $Y_i(Z_i = 0) = Y_i(0)$ and $Y_i(Z_i = 1) = Y_i(1)$. In common language, these can be described as the outcome that would be observed if individual i had not received the treatment ($Z_i = 0$) and the outcome that would be observed if individual i had received the treatment ($Z_i = 1$), respectively. In the context of the example of the effect of a drug on blood pressure, $Y_i(0)$ would reflect the systolic blood pressure at a particular point in time for individual i observed in the absence of the drug and $Y_i(1)$ would reflect the systolic blood pressure observed *at that same point in time* for individual i after treatment with the drug.

One benefit of the potential outcome notation is that it allows for a simple and clear definition of the individual-level causal effect for observation i as a comparison between these two potential outcomes, for instance, $Y_i(1) - Y_i(0)$. (Nonlinear comparisons of potential outcomes, such as those used to define an odds ratio can also be considered; for simplicity this article focuses on the simple difference, in part because of complications that arise with noncollapsible functions of the potential outcomes; see, e.g., Greenland et al., 1999; Austin, 2007.) This formalization makes explicit the difficulty in inferring causality at the individual level given that we can never observe both $Y_i(1)$ and $Y_i(0)$ for the same individual, at any given time point; Holland (1986) refers to this as the 'fundamental problem of causal inference.' Using this potential outcomes framework to define causal effects also allows us to define those effects nonparametrically, and creates a useful distinction between structural assumptions and modeling assumptions (discussed below).

While it is nearly impossible to empirically identify individual-level causal effects without very strong assumptions or very specialized settings (for instance, in a laboratory where conditions can be precisely controlled), it is possible to make progress with respect to estimating average causal effects. The most common such estimand is the average of individual-level causal effects over a population or sample of interest, expressed generally as $E[Y(1) - Y(0)] = (1/N)\Sigma_i(Y_i(1) - Y_i(0))$. When this average is taken over the full population of interest, it is

referred to as the population average treatment effect. When this average is taken over the analysis sample it is referred to as the sample average treatments effect (SATE).

There is also sometimes interest in causal effects for particular subgroups. As a simple example, researchers may be interested in the SATE for males as distinct from the SATE for females. A slightly more complex example is sometimes referred to as the 'average treatment effect on the treated' (ATT). This is the average treatment effect for those in the treatment group, and is defined as $E[Y(1) - Y(0)|Z = 1] = (1/N_t)\Sigma_i(Y_i(1) - Y_i(0)|Z_i = 1)$. (An analogous quantity called the sample average treatment effect on the control (SATC) can also be defined.) Researchers are sometimes interested in the ATT in settings where there is no thought of giving the treatment to the full population, and rather interest is in the effect of the treatment on the people who actually received it. This might be the relevant quantity, for instance, in studies of the effects of adolescent drug use, where there is no thought of 'imposing' drug use on all adolescents. The sample average treatment effect on the treated (SATT) in such a study would compare the average outcomes for drug users with the potential outcomes for the same group had they not used drugs. This could inform the potential effect of programs that help to prevent or terminate drug use in teenagers. The SATE, on the other hand, would represent the difference between the average potential outcomes if all adolescents in the study were heavy drug users versus the potential outcomes that would manifest if none of the adolescents in the study were heavy drug users. This would not be the relevant scientific quantity given that some portion of that sample would likely never choose to use drugs so the effect for these people is rather meaningless. See Imai et al. (2008) and Harder et al. (2010) for more discussion of possible estimands.

To illustrate some of these concepts, consider the hypothetical data in Table 1. These data reflect 14 participants in a study that aimed to estimate the effect of a job training program on hourly wages. The table presents an omniscient view of the data that the researcher could never see in entirety because the Y(0) cells for those in the treatment group and the Y(1) cells for those in the control group are complete. The column for Y denotes the outcome that is observed for each person: Y = Treat*Y(1) + (1 − Treat)*Y(0). Seeing this complete data set with all of the potential outcomes observed is of course not possible in the real world, but, for illustration, it allows us to see the causal effects for each individual and, by extension, to calculate several different sample estimands. The treatment effects vary in this example depending on whether the participant had received a high school education (HS). The subgroups without high school education (HS = 0) are all defined to have treatment effects of 4. On the other hand, those with a high school education would experience no gain from the job training (for each, Y(0) = Y(1)). Since the distribution of those with high school education varies across the treatment groups, the average treatment effects vary as well. The SATT is equal to 2.67 and the SATC is equal to 1.

Table 1 Displays data from a hypothetical job training study. Data on high school graduation status (HS), age, and hourly wage (Y) could be observed by the researcher. The full set of potential outcomes could not be observed. They are displayed here to aid in conceptualization of individual-level causal effects as well as several sample average estimands

Person	*Treat (Z)*	*HS*	*Age*	*Y(0)*	*Y(1)*	*Y*
1	1	0	25	12	16	16
2	1	0	22	11	15	15
3	1	0	21	10	14	14
4	1	0	23	11	15	15
5	1	1	22	17	17	17
6	1	1	24	19	19	19
7	0	0	20	10	14	10
8	0	0	24	12	16	12
9	0	1	22	17	17	17
10	0	1	22	16	16	16
11	0	1	26	19	19	19
12	0	1	20	18	18	18
13	0	1	28	19	19	19
14	0	1	21	17	17	17

Structural Assumptions Required for Nonparametric Identification

If we could observe all of the potential outcomes for our sample, then causal inference would be easy. However, as discussed above, the fundamental problem of causal inference reflects the fact that at best we can observe only half of the potential outcomes, and thus need to rely on assumptions about those missing potential outcomes. In some sense this is what distinguishes causal inference from standard statistical inference, in that it (generally) requires untestable assumptions regarding the unobserved potential outcomes. This section discusses the structural assumptions needed to estimate causal effects, followed by a brief discussion of possible study designs and additional parametric assumptions.

Ignorability

Unconditional Ignorability

What assumption would allow us to identify average causal effects? To make the issues concrete, consider each piece of the SATE. Estimating E[Y(1)] for the sample is challenging because we only observe Y(1) for those observations that were assigned to receive Z = 1. If those observations assigned to Z = 0 have a different mean of Y(1) than those who were assigned to Z = 1, then an estimate of the mean of the treated outcomes, E[Y(1)|Z = 1], will be a biased estimate of E[Y(1)]. Similarly, estimating E[Y(0)] is challenging because we only observe Y(0) for those observations that were assigned to receive Z = 0. If those observations assigned to Z = 1 have a different mean of Y(0) than those who were assigned to Z = 0, then an estimate of the mean of the control outcomes, E[Y(0)|Z = 0], will be a biased estimate of E[Y(0)]. The most straightforward set of assumptions that ensures that a simple difference in means can provide an unbiased estimate of the average treatment effect relies on mean independence:

$$E[Y(0)] = E[Y(0)|Z = 0]$$

$$E[Y(1)] = E[Y(1)|Z = 1]$$

Designs and estimation procedures discussed in a later section highlight the distinction between assumptions and estimation. However, it is worth mentioning at this point that a completely randomized experiment would automatically satisfy this mean independence assumption. In fact, it will satisfy an even stronger assumption, that the distributions of the potential outcomes are independent of the binary outcome Z: $Y(0), Y(1) \perp Z$. Because of the connection to this classic design in causal inference, this assumption, sometimes referred to as the ignorability assumption, is also referred to as the randomization assumption.

The other estimands discussed above require slightly weaker assumptions. For instance, the effect of the treatment on the treated can be decomposed into two pieces, $E[Y(1)|Z = 1]$ and $E[Y(0)|Z = 1]$. Since we observe $Y(1)$ for the treated ($E[Y(1)|Z = 1] = E[Y|Z = 1]$), it is straightforward to estimate $E[Y(1)|Z = 1]$. However, it is difficult to estimate $E[Y(0)|Z = 1]$, since we do not observe $Y(0)$ for any treated individuals. It only becomes straightforward if $E[Y(0)|Z = 1] = E[Y(0)|Z = 0]$. This would be a consequence of randomization of the treatment Z as well since that design ensures that $Y(0) \perp Z$. Similarly, $Y(1) \perp Z$ would allow us to estimate the effect of the treatment on the controls. While these are weaker assumptions than are necessary for identification of the average treatment effect, it is difficult to find situations in which one assumption is clearly more plausible than the other.

Ignorability

Unless a randomized experiment has been performed, the above version of the ignorability assumption is typically not plausible because it implies that there are no systematic differences between the treatment and control groups with respect to the potential outcomes. Therefore, many causal inference methods rely on the more general version of the assumption above that conditions on observed covariates, X, formalized as $Y(0), Y(1) \perp Z|X$. In words, this assumption requires that the treatment and control group members that have the same values of X are not systematically different with regard to their potential outcomes. Some authors refer to this formulation as 'conditional ignorability,' however, this nomenclature is redundant since the standard formulation of the assumption allows for conditioning on covariates.

A simple example in which we would know that the ignorability assumption holds would be a randomized block experiment in which the blocks are defined by covariates, X. Observational studies differ in a crucial way from randomized block experiments, however, because no explicit randomization occurs; thus, we must take it on faith that the ignorability assumption holds.

How does ignorability help? When the ignorability assumption holds it implies the following properties:

$$E[E[Y|Z = 0, X]] = E[E[Y(0)|Z = 1, X]] = E[E[Y(0)|X]] = E[Y(0)]$$

and

$$E[E[Y|Z = 1, X]] = E[E[Y(1)|Z = 1, X]] = E[E[Y(1)|X]] = E[Y(1)]$$

This means that if ignorability holds (conditional on X) and we can appropriately condition on X, then all we have to do is average these conditional means over the distribution of X and we can recover the marginal means of the potential outcomes. If X is not low-dimensional, however, this conditioning can prove difficult in practice, which has led to use of some of parametric, semiparametric, and sophisticated nonparametric approaches discussed below or in other articles.

Common Support

Satisfying ignorability ensures comparability (with regard to potential outcomes) across treatment and control groups for observations with the same values of the covariates. However, this is only helpful if there is *common support* across treatment groups with respect to the confounders. That is, there must be observations from both treatment groups for each neighborhood of the covariate space where we want to make inferences. If, for example, no control units exist for a given neighborhood of the covariate space where treated units exist, then, in essence, no 'empirical counterfactuals' exist for those treatment units and we cannot make causal inferences about them without further assumptions. For this reason, causal researchers often require an assumption of overlap or 'common support' when attempting to identify causal effects, typically formalized as $0 < Pr(Z|X) < 1$. In a randomized experiment, this is operationalized by the fact that all individuals in the experiment have a positive probability of receiving either treatment condition; individuals who could not receive one of the treatments would be deemed ineligible for the study.

In practice, if researchers determine that this assumption is not satisfied, they might decide to perform inference only for observations in portions of the covariate space where both treated and control units exist. This may lead to a different causal estimand than originally intended and results must be interpreted accordingly (see, e.g., Crump et al., 2006). This trade-off in what is being estimated and how well it can be estimated is often of particular relevance in studies that use propensity score methods, discussed further in other articles on observational studies.

The determination of common support depends on the inferential goal. For instance, if the goal is to estimate the effect of the treatment on the treated, this would require that control observations exist in every neighborhood of the covariate space where treatment observations exist, but not vice versa. If there are parts of the covariate space where there are controls but no treated subjects, it does not matter for nonparametric identification of the ATT since the goal is not to make inferences about the controls anyway. This lack of overlap could influence parametric estimation, however.

Common support is the only standard causal inference assumption that is empirically verifiable. However, the usefulness of this property is hampered by two issues. The first is that researchers may condition on a large number of covariates in an attempt to satisfy ignorability and it becomes increasingly difficult to verify overlap as the dimensionality of the covariate space grows. A common approach to this problem is simply to examine overlap with regard to the propensity score (*see* Observational Studies: Overview). However, this requires correct specification of the propensity score model. The second issue is that even if the included covariates satisfy ignorability, they may represent a superset of the covariates

actually needed to satisfy ignorability. Researchers forcing common support with respect to all included covariates may make overly conservative judgments about which observations need to be discarded from an analysis. Alternatives exist that aim to satisfy *common causal support*, $0 < \Pr(Z|W) < 1$, which requires overlap only with respect to the subset of covariates needed to satisfy ignorability, W (Hill and Yu-Sung, 2013).

Stable Unit Treatment Value Assumption

The potential outcome notation introduced above has an assumption embedded in it. The potential outcomes for individual i under treatment and control conditions, respectively, were defined as a function of the treatment exposure for individual i alone, $Y_i(Z_i = 1) = Y_i(1)$ and $Y_i(Z_i = 0) = Y_i(0)$. This ignores the possibility that this outcome could be influenced by the treatments received by others. A more general formulation, $Y_i(Z)$, could define potential outcomes based on the full vector of treatments received by everyone in the study, Z. Even with a modest sample size, such a formulation would increase the complexity of the research questions and subsequent analysis beyond a reasonable limit, not to mention the impossibility of ever achieving common support. For instance, in a study with a binary treatment and 10 people this formulation would lead to $2^{10} = 1024$ different potential outcomes for each person. The assumption that each person's potential outcomes are only defined based on his or her own treatment assignment can be formalized by the following expression: $Y_i(Z) = Y_i(Z')$, if and only if $Z_i = Z'_i$. This assumption is often discussed colloquially as the 'no interference assumption' to reflect the requirement that one individual's potential outcomes cannot be changed by another person's treatment. However, implicit in the stable unit treatment value assumption (SUTVA) is also the assumption that the treatment and control conditions mean the same thing for all units. In other words, there is only one 'version' of the treatment condition and one 'version' of the control condition (see Hernan and VanderWeele, 2011 for more thorough discussion of this assumption and its implications).

The most common strategy for avoiding concerns about SUTVA violations is through study designs that assign treatments at the group level. For example, many studies of educational interventions randomize whole schools to treatment conditions rather than randomizing individuals within schools, due to concerns that the treatment received by one individual may affect the outcomes of other individuals in the same school. This interference is less of a concern when considering schools as the units.

Temporality

Implicit in our understanding of causality is the notion that our causal variable (or treatment) occurs temporally prior to the outcome. Moreover, when conditioning on covariates it is crucial that they represent events that occur or characteristics measured prior to the causal (treatment) variable, otherwise they are implicitly outcomes and need to be conceived as such. Rosenbaum (1984) presents a discussion of this issue as do other authors (such as Gelman and Hill, 2007). Conditioning on variables that may be affected by the treatment (e.g., mediators) brings in additional complexities because they require consideration of the potential values of those post-treatment variables. A framework for thinking about post-treatment variables, called principal stratification, is discussed further in the article with that name. Other approaches to exploring causal pathways posttreatment are discussed in the article on Mediation.

Design Considerations

Causal researchers have carefully considered the types of research designs that facilitate satisfying the required assumptions. Generally, randomized experiments are seen as the strongest design for estimating causal effects since ignorability is assured through the randomization. Randomization also obviates the need to fit models that condition on pretreatment covariates and makes inferences from such models (when used, for instance, to increase efficiency) more robust to departures from the parametric assumptions. However, randomized experiments are not always possible, and can have their own complications that make inferences challenging, such as noncompliance and missing data (Barnard et al., 2003). A final concern with randomized experiments is that they often enroll subjects that are not representative of the population of interest, potentially making them less useful for estimating population treatment effects (Imai et al., 2008).

A number of strong nonexperimental study designs exist for settings when randomized experiments are infeasible or do not answer the question of interest. A common theme across nonexperimental studies is the need to design them with as much (or more) care as is used for randomized experiments (Rosenbaum, 1999; Rubin, 2008). Some nonexperimental designs are motivated by the desire to avoid or make more plausible the assumption of ignorability of the treatment assignment. These designs include regression discontinuity, interrupted time series, fixed effects, and instrumental variables approaches. They are discussed in more detail in other articles on Observational Studies.

Other nonexperimental study designs are motivated by the desire to avoid or weaken parametric assumptions by creating as much balance and overlap across treatment and control groups as possible. Matching and stratification (with or without propensity scores) are the most prominent designs in this genre. This type of 'preprocessing' (see Ho et al., 2007) can allow either for nonparametric inference on the newly constructed treatment and control groups (for instance, by comparing mean outcomes across matched treatment and control groups) or for more robust inference using standard parametric models (for instance, by fitting a regression on matched treatment and control groups). Treatment effect estimation is more robust to deviations from the parametric models when the treatment and control groups have common support and are well balanced (that is, the covariate distributions are similar across groups).

Nonexperimental studies of all types are ideally designed in ways that limit their sensitivity to unobserved confounding (violations of ignorability). Rosenbaum (2004) formalizes this

in a concept called 'design sensitivity' and Zubizarreta et al. (2013) provide one example of implementing these ideas in practice, such as comparing extreme values of a continuous treatment variable, matching on a large set of observed characteristics, and tying the statistical test used to the hypothesized pattern of effects. Gelman and Hill (2007) and Shadish et al. (2002) provide more information on nonexperimental study designs, and cite relevant articles on observational studies.

Parametric Assumptions

The focus of this article thus far has been on identification of causal effects through nonparametric approaches (comparisons of means or subgroup-specific means). It can be difficult to maintain a completely nonparametric approach in situations where it is necessary or beneficial to condition on covariates. For instance, if ignorability requires conditioning on covariates, then we must find a way to model the requisite conditional expectations. A straightforward approach to such conditioning is by fitting a linear regression model to the observed data. In particular, if one assumes that $Y(0) = X\beta + \varepsilon_0$ and $Y(1) = X\beta + \tau + \varepsilon_1$, then regressing observed Y on X and Z would be a reasonable approach for estimating τ. However, we rarely believe that such simple additive and linear models hold in practice, and it can be difficult to diagnose deviations from these models when X is high-dimensional or when there is lack of common support across treatment and control groups.

As discussed above, some designs (notably matching and stratification) are motivated by the desire to avoid these types of stringent parametric assumptions. However, an alternative (or complementary) approach is to simply use a semiparametric or nonparametric strategy when estimating causal effects. Examples of semiparametric estimation including regression models that incorporate weights based on the propensity score could also be considered to be a part of this paradigm (Hahn, 1998; Rosenbaum 1987; Hernan et al., 2001; Kurth et al., 2006). A more liberal interpretation might also frame weighting strategies as observational 'designs' because the goal is to create a 'pseudo-population' within which treatment and control groups are balanced. Nonparametric regression methods that can accommodate large numbers of covariates are typically Bayesian or computationally driven, or both (examples can be found in Hill, 2011 and Karabatsos and Walker, 2012).

Some causal inference approaches target the structural assumptions without considering the parametric assumptions. For instance, classic fixed effects models are often used in causal inference settings as a way of capturing group-level confounders that are not directly measured (*see* Hierarchical Models: Random and Fixed Effects). However, these models are traditionally embedded in a classical linear regression framework with all the parametric assumptions inherent in that framework. Moreover there is no guarantee that such strategies will reduce rather than amplify bias.

Conclusions

This article has discussed the potential outcomes framework for defining and estimating causal effects, in particular determining the effects of treatments on outcomes. It has highlighted the key structural and parametric assumptions commonly made, and the implications of those assumptions in experimental and nonexperimental settings. For more information on the details of estimating causal effects please see the other associated articles in this volume.

See also: Experimental Design: Bayesian Designs; Experimental Design: Compliance; Experimental Design: Large-Scale Social Experimentation; Experimental Design: Overview; Hierarchical Models: Random and Fixed Effects; Instrumental Variables in Statistics and Econometrics; Latent Structure and Causal Variables; Longitudinal Causal Inference; Mediation, Statistical; Observational Studies: Overview; Semiparametric Models; Time Series: General.

Bibliography

Austin, Peter C., 2007. The performance of different propensity score methods for estimating marginal odds ratios. Statistical Medicine 26, 3078–3094.

Barnard, J., Frangakis, C.E., Hill, J.L., Rubin, D.B., 2003. Principal stratification approach to broken randomized experiments: a case study of school choice vouchers in New York city. Journal of the American Statistical Association 98 (462), 299–323.

Crump, R.K., Hotz, J.V., Imbens, G.W., Mitnik, O.A., 2006. Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand. NBER Technical Working Paper No. 330.

Dowd, Bryan E., 2011. Separated at birth: statisticians, social scientists, and causality in health services research. Health Services Research 46 (2), 397–420.

Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge.

Greenland, S., Robins, J.M., Pearl, J., 1999. Confounding and collapsibility in causal inference. Statistical Science 14 (1), 29–46.

Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66, 315–322.

Harder, V.S., Stuart, E.A., Anthony, J., 2010. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychological Methods 15 (3), 234–249.

Hernan, M.A., VanderWeele, T.J., 2011. Compound treatments and transportability of causal inference. Epidemiology 22 (3), 368–377.

Hernan, M., Brumback, B., Robins, J., 2001. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. Journal of the American Statistical Association 96 (454), 440–448.

Hill, Jennifer L., 2011. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics 20, 217–240.

Holland, P.W., 1986. Statistics and causal inference. Journal of the American Statistical Association 81 (396), 945–960.

Imai, K., King, G., Stuart, E.A., 2008. Misunderstandings between experimentalists and observationalists about causal inference. Journal of the Royal Statistical Society, Series A 171, 481–502.

Karabatsos, G., Walker, S.G., 2012. A Bayesian nonparametric causal model. Journal of Statistical Planning and Inference 142, 925–993.

Kurth, T., Walker, A.M., Glynn, R.J., Chan, K.A., Gaziano, J.M., Berger, K., Robins, J.M., 2006. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of non-uniform effect. American Journal of Epidemiology 163 (3), 262–270.

Pearl, J., 2009. Causality: Models, Reasoning, and Inference, second ed. Cambridge University Press, New York.

Rosenbaum, P.R., 1999. Choice as an alternative to control in observational studies. Statistical Science 14 (3), 259–304.

Rosenbaum, Paul R., 1987. Model-based direct adjustment. Journal of the American Statistical Association 82, 387–394.

Rosenbaum, Paul R., 2004. Design sensitivity in observational studies. Biometrika 91 (1), 153–164.

Rubin, Donald, 1978. Bayesian inference for causal effects: the role of randomization. The Annals of Statistics 6, 34–58.

Rubin, Donald B., 2008. For objective causal inference, design trumps analysis. The Annals of Applied Statistics 2 (3), 808–840.

Shadish, William R., 2002. Revisiting field experimentation: field notes for the future. Psychological Methods 7, 3–18.

VanderWeele, Tyler J., 2012. Invited commentary: structural equation models and epidemiologic analysis. American Journal of Epidemiology 176 (7), 608–612.

Zubizarreta, J.R., Cerda, M., Rosenbaum, P.R., 2013. Effect of the 2010 Chilean earthquake on posttraumatic stress: reducing sensitivity to unmeasured bias through study design. Epidemiology 24 (1), 79–87.