Advance Access publication July 3, 2013

Political Analysis (2014) 22:169–182
doi:10.1093/pan/mpt006

A Selection Bias Approach to Sensitivity Analysis for Causal Effects

Matthew Blackwell
Department of Political Science, University of Rochester, 307 Harkness Hall, Rochester, NY 14627, NY
e-mail: m.blackwell@rochester.edu (corresponding author)

Edited by R. Michael Alvarez

The estimation of causal effects has a revered place in all fields of empirical political science, but a large volume of methodological and applied work ignores a fundamental fact: most people are skeptical of estimated causal effects. In particular, researchers are often worried about the assumption of no omitted variables or no unmeasured confounders. This article combines two approaches to sensitivity analysis to provide researchers with a tool to investigate how specific violations of no omitted variables alter their estimates. This approach can help researchers determine which narratives imply weaker results and which actually strengthen their claims. This gives researchers and critics a reasoned and quantitative approach to assessing the plausibility of causal effects. To demonstrate the approach, I present applications to three causal inference estimation strategies: regression, matching, and weighting.

1 Introduction

Scientific progress marches to the drumbeat of criticism and skepticism. Although the social sciences marshal empirical evidence for interpretations and hypotheses about the world, an academic's first (healthy!) instinct is usually to counterattack with an alternative account. This reinterpretation of empirical results demands a response—how would this alternative story affect the results? Often, the response is verbal and ad hoc, but there is room for improvement. A crucial, if rare, exercise is a formal sensitivity analysis that weighs these alternative accounts against the empirical evidence. As Rosenbaum (2002) puts it, the goal of a formal sensitivity analysis is "to give quantitative expression to the magnitude of uncertainties about bias." This article presents a broad methodology for evaluating the sensitivity of causal effect estimation to specific critiques of bias.

The estimation of causal effects in particular has a revered place in all fields of empirical political science. We are deeply interested in how institutions, policies, strategies, and beliefs affect political life. And although there has been a rapid growth in attention to the careful identification of causal effects, methodological and applied analyses in causal inference often overlook a fundamental fact: many scholars are skeptical of identification in observational studies. Most causal inferences require an assumption of ignorability or no omitted variables that requires treated units be comparable to control units, possibly conditional on a set of observed covariates. Of course, such an assumption is rarely justified by the study design alone.[1]

Author's note: The methods used in this article are available as an open-source R package, causalsens, on the Comprehensive R Archive Network (CRAN) and the author's web site. The replication archive for this article is available at the Political Analysis Dataverse as Blackwell (2013b). Many thanks to Steve Ansolabehere, Adam Glynn, Gary King, Jamie Robins, Maya Sen, and two anonymous reviewers for helpful comments and discussions. All remaining errors are my own.

[1] There has been a steady increase in attention to causal inference without ignorability assumptions. Extending the usual bounding approach by Manski (1990), political scientists have added additional assumptions to generate bounds, point estimates, or hypothesis tests for causal effects. See Mebane and Poast (2013) and Glynn and Quinn (2011) for examples of this approach.

In this article, I combine the approaches of two sensitivity analysis traditions. First, in the spirit of Brumback et al. (2004), Robins (1999), and Heckman et al. (1998), I introduce the confounding function, which quantifies the extent of unmeasured confounding. This approach is useful because it avoids the process of imagining the presence of specific (uni- or multivariate) omitted variables. Instead, researchers directly vary the selection bias inherent in the treatment assignment. I also extend this method by showing its applicability to nonweighting approaches to causal inference. Second, I combine the confounding function approach with that of Imbens (2003) to ground the sensitivity analysis in an easily interpretable framework.

One advantage of this approach is that, once we calculate the confounding function and a propensity score, the sensitivity analysis only requires an adjustment to the dependent variable. Whatever causal inference method a researcher uses in his or her main analysis (regression, matching, weighting) applies to this adjusted dependent variable. This approach even applies to marginal structural models with time-varying treatments (Blackwell 2013a). Thus, this approach is widely applicable with a minimal burden to applied researchers. In addition, the approach allows researchers to evaluate narratives about the sensitivity of their effects. They can answer questions of the following form: what would happen if the treated units are inherently better off than the control units? This approach allows for possible increases and decreases in the effect due to deviations from ignorability. As with attenuation due to measurement error, scholars want to know when their biases are in a "safe" direction as much as when they are not.

This article proceeds as follows. Section 2 reviews the foundations of causal inference. Section 3 lays out the approach to sensitivity analysis and provides a convenient reparameterization in terms of variance explained. Section 4 demonstrates the method in three distinct areas, each of which has a different estimation strategy: regression, matching, and weighting. Section 5 concludes with thoughts for future work.

2 A Review of Causal Inference

Let A_i be a dichtomous action or a treatment taken by unit i and Y_i be the outcome for that unit. It is common to refer to those units with $A_i = 1$ as *treated* and those with $A_i = 0$ as *control*. The goal will be to estimate the effect of A_i on Y_i. Following Rubin (1978), we can conceptualize causal effects as contrasts between various *potential outcomes*. Let $Y_i(1)$ denote the outcome if i were treated and $Y_i(0)$ denote the outcome if i received the control. The individual causal effect for unit i would be the difference between these two states of the world:

$$\tau_i = Y_i(1) - Y_i(0). \tag{1}$$

Without strong assumptions, these individual causal effects are inestimable because units live in, at most, one of the two states of the world. That is, we observe a unit's outcome under control or we observe a unit's outcome under treatment, but rarely both. This is often called the fundamental problem of causal inference.

Although individual causal effects are generally beyond reach, there are other causal quantities that are estimable with weaker assumptions. For instance, a common quantity is the average treatment effect, or ATE, which is simply the average of the individual effects:

$$\tau = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)], \tag{2}$$

where the expectation is over units. The ATE measures what would happen, on average, if all units were treated versus if all units were withheld treatment. Another common approach is to estimate the average effect of the treatment among the treated units, or ATT:

$$\tau_{\text{att}} = E[Y_i(1) - Y_i(0)|A_i = 1] = E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1]. \tag{3}$$

This quantity is attractive because it requires slightly weaker assumptions on how the treatment is assigned.

2.1 *Assumptions and Estimation Strategies*

Without additional assumptions, the above causal quantities of interest are functions of unobservables. In order to estimate these causal effects, we need to make assumptions to connect the unobserved potential outcomes to the data.

Assumption 1 (Consistency). *Let* $a \in (0, 1)$ *be a treatment status. Then, for unit* i *with* $A_i = a$*, we assume* $Y_i(a) = Y_i$.

This assumption simply connects the potential outcomes to the observed outcomes. Namely, we assume that units that take an action will observe the potential outcomes for that action. Furthermore, the connection between potential and observed outcomes does not depend on any other variables. This forbids any spillover effects where the treatment assignment of one unit affects the outcome of another unit. The second assumption is the cornerstone of identification for most causal estimates.

Assumption 2 (Ignorability). *For a set of covariates,* X_i*, and treatment statuses,* $a \in (0, 1)$*,* $Y_i(a) \perp\!\!\!\perp A_i | X_i$.

Here, $B \perp\!\!\!\perp C|D$ means that B is independent of C, conditional on D (Dawid 1979). This assumption requires that the treatment status be independent of the potential outcomes, conditional on a set of covariates. When the treatment assignment is random, this assumption is satisfied trivially because everything will be independent of the assignment. In an observational study, however, the analyst's goal is to collect as many variables as possible to include in X_i to make Assumption 2 as plausible as possible. An unmeasured confounder that affects both the treatment status and the outcome would violate this assumption. Many sensitivity analysis methods, including Imbens (2003), imagine one such unmeasured confounder and vary its impact on the treatment assignment and the outcome to assess the sensitivity of effects. The present method instead directly models violations of ignorability, agnostic to the type or number of unmeasured confounders.

Three of the most common approaches to estimating the causal estimands τ and τ_{att} are regression, matching, and weighting. Under the above two assumptions, each of these can consistently estimate *some* causal parameter and there is a large literature comparing their relative advantages in different situations (see, e.g., Imbens 2004; Morgan and Winship 2007).[2] Below, I present results from the sensitivity analysis procedure applied to each.

2.2 *Previous Approaches to Sensitivity Analysis*

Formal sensitivity analyses has been a part of causal inference since at least Cornfield et al. (1959), with significant advances that focus largely on medical studies.[3] Rosenbaum (2002) presents a method based on the unobserved differences in treatment assignment probabilities. His sensitivity analysis framework then finds the most extreme inferences possible based on a specific unobserved difference. That is, Rosenbaum fixes a difference in treatment assignment and then calculates bounds on the significance level if those differences were maximally correlated with the outcome. To accomplish this, his model places constraints on the imagined unmeasured confounder: that it be between 0 and 1. This approach is very useful for general sensitivity, but less so when evaluating alternative stories. In general, the selection bias approach below can be used for both situations and it avoids placing restrictions on the unmeasured confounder.

Imbens (2003) uses a similar reparameterization as the present method, but still relies on a hypothesized unmeasured confounder and a larger parametric model to justify the reparameterization. The selection bias approach only requires a baseline model of the relationship between one

[2] And in some cases, these categories overlap in the sense that one method can be rewritten as a special case of another.
[3] It is important to note that simply varying the specification of a statistical model is not a good substitute for a formal sensitivity analysis. This might gauge how small perturbations to the statistical model may change estimates, but it does not provide any formal quantification of our uncertainty due to bias.

potential outcome and the covariates. Imai, Keele, and Yamamoto (2010) and Imai et al. (2011) provide an approach to sensitivity analysis similar in spirit to the selection bias approach, but targeted toward a specific causal parameter: the average causal mediation effect. These previous approaches to sensitivity analysis look for the minimum perturbations needed to overturn or significantly change results estimated under the standard assumptions. The approach here takes a different tack: it determines how specific violations of confounding alter the magnitude and direction of causal estimates.

3 A Selection Bias Approach to Sensitivity Analysis

3.1 *The Confounding Function*

One way to specify and describe sensitivity to unmeasured confounders is to vary the amount of confounding or selection bias that exists for a given causal estimate. At its core, confounding means that the potential outcomes vary by the treatment status. We can represent this confounding as a function of the observed covariates:

$$q(a, x) = E[Y_i(a)|A_i = a, X_i = x] - E[Y_i(a)|A_i = 1 - a, X_i = x]. \quad (4)$$

This function represents the confounding for treatment status a with covariates x. The ignorability assumption implies that $q = 0$ so that A_i and $Y_i(a)$ are (mean) independent, no matter the value of X_i. The confounding function directly models violations of ignorability: if $q(a, x)$ is positive, then units in group a have a higher mean potential outcome under a than those in group $1 - a$. Thus, q encodes the selection bias of the treatment assignment. For instance, suppose we have an observational study where the treatment is negative campaigns ($A_i = 1$) versus positive campaigns ($A_i = 0$) and the outcome is voter turnout. Then, $q(1, x) > 0$ implies that the observed negative campaigns have inherently higher turnout compared to the observed positive campaigns if those positive campaigns had in fact been negative instead. That is, there is a difference between the negative and positive campaigns beyond any causal effect.

At this point, q is completely unrestricted, but it is useful to use a simple parameterization to succinctly describe the selection bias and plot it against the estimated effect for that value of q. That is, we allow the confounding function to vary according to a single parameter, α:

$$q(a, x; \alpha) = \alpha. \quad (5)$$

Here, when $\alpha > 0$, the observed potential outcomes ($Y_i(1)$ for $A_i = 1$ and $Y_i(0)$ for $A_i = 0$) are on average higher than their counterfactuals at every level of X_i. If higher levels of Y_i are better, then the observed treatment allocation is preferred to an alternative where it is reversed. When $\alpha < 0$, the opposite is true—the observed treatment assignment is suboptimal.

We can alter our confounding function to change the type of sensitivity analysis we want to conduct. For instance, suppose the observed negative campaigns either have inherently higher or lower turnout, beyond the effect of campaign tone. Then, the confounding function varies by the treatment status,

$$q(a, x; \alpha) = \alpha(2a - 1), \quad (6)$$

so that when $\alpha > 0$, the treated group always has higher mean potential outcomes than the control group. Of course, when $\alpha < 0$, the control group is better off. A cornerstone of both parameterizations is that $q = 0$ when $\alpha = 0$, which corresponds to the standard ignorability assumption. In this case, the results of a typical matching or regression analysis will hold.

3.2 *Implementation of the Sensitivity Analysis*

One benefit to the selection bias approach to sensitivity analysis is that the implementation is both straightforward and largely independent of the causal estimation strategy. In fact, once we have specified a confounding function, this approach only requires an estimate of the propensity score. With these in hand, we adjust the dependent variable and reestimate our original analysis on this

adjusted dependent variable. That is, we replace our observed outcome, Y_i, with the confounding-adjusted outcome,

$$Y_i^q = Y_i - q(A_i, X_i)\Pr[1 - A_i|X_i]. \tag{7}$$

Here, we are essentially subtracting the omitted variable bias from the outcome. To see how this adjustment works, it is instructive to look at a case without covariates:

$$E[Y_i(0)] = E[Y_i(0)|A_i = 0]\Pr[A_i = 0] + E[Y_i(0)|A_i = 1]\Pr[A_i = 1] \tag{8}$$

$$\begin{aligned} &= E[Y_i(0)|A_i = 0]\Pr[A_i = 0] + E[Y_i(0)|A_i = 1]\Pr[A_i = 1] \\ &\quad + E[Y_i(0)|A_i = 0]\Pr[A_i = 1] - E[Y_i(0)|A_i = 0]\Pr[A_i = 1] \end{aligned} \tag{9}$$

$$\begin{aligned} &= (\Pr[A_1 = 1] + \Pr[A_1 = 1])E[Y_i(0)|A_i = 0] \\ &\quad - (E[Y_i(0)|A_i = 0] - E[Y_i(0)|A_i = 1])\Pr[A_i = 1] \end{aligned} \tag{10}$$

$$= E[Y_i|A_i = 0] - q(0)\Pr[A_i = 1] \tag{11}$$

$$= E[Y_i^q|A_i = 0]. \tag{12}$$

Note that equation (11) follows from consistency, and the rest of these from the properties of conditional probability. None invoke ignorability. This analysis holds even if covariates are added.

With the adjustment in hand, researchers can run their original analysis model on this transformed outcome. Different estimands require slightly different adjustments. If the ATT is of interest, for instance, one need only adjust the control units:

$$Y_i^q = Y_i - (1 - A_i)q(0, X_i)\Pr[A_i = 1|X_i]. \tag{13}$$

Equations (8–12) show why and how this works: the mean of the adjusted outcome for controls equals the mean of the potential outcome under control. Brumback et al. (2004) shows that with the confounding-adjusted outcome, Y_i^q, a marginal structural model and inverse probability of treatment weighting can consistently estimate causal effects. We can be more general, though: any estimator that consistently estimates causal effects under mean unconfoundedness will consistently estimate causal effects with the confounding-adjusted outcome.

An intuitive reason for this result is that the adjustment ensures that mean ignorability holds: $E[Y(0)|X_i] = E[Y_i^q|A_i = 0, X_i]$. Thus, in this adjusted data, confounding no longer causes bias because it no longer exists. Thus, any consistent estimator for $E[Y(0)|X_i]$ that relies on unconfoundedness will have asymptotic bias when using Y_i, but be consistent when using Y_i^q. This allows regression and matching to recover the causal effect in the face of specified unmeasured confounding. This is crucial for our sensitivity analysis because we can vary q or a parameter of q and see the consistently estimated causal effect that q implies.

The confounding adjustment approach to sensitivity analysis has the attractive property of not requiring any change to the matching procedure or propensity score estimation. The only change to the estimation procedure when assuming $q \neq 0$ is an adjustment to the outcome. Thus, we only have to reestimate any function of the dependent variable, such as a regression model or difference in means. Any preprocessing steps remain fixed over various assumptions about q.

3.3 *The Choice of Confounding Function*

The parametric assumptions on the confounding function are crucial to the sensitivity analysis performed. This is because the selection bias approach can only detect sensitivities in the directions allowed by the confounding function. Take as an example the confounding function $q = \alpha(2a - 1)$, which tests against *one-sided bias*: $Y_i(1)$ is higher (lower) for the treatment group when $\alpha > 0$ ($\alpha < 0$). As the name implies, this function can detect sensitivity to one-sided selection bias, but it would fail to detect other deviations from ignorability. That is, it can only determine the bias resulting from the treatment group being on average better off or the control group being on average better off. The sensitivity analysis is rigid in this way because the confounding function

is not identified from the data, so that the causal model in the last section is only identified conditional on a specific choice of that function. The goal of the sensitivity analysis is not to choose the "correct" confounding function, since we have no way of evaluating this correctness. By its very nature, unmeasured confounding is unmeasured. Rather, the goal is to identify plausible deviations from ignorability and test sensitivity to those deviations. The main harm that results from the incorrect specification of the confounding function is that hidden biases remain hidden.

An alternative confounding function, $q = \alpha$, identifies sensitivity to what I call *alignment bias*. This type of bias is likely to occur when units select into treatment and control based on their predicted treatment effects. For instance, this might occur with observational studies of voter outreach: campaigns might already be targeting their turnout efforts toward individuals who they suspect will respond more positively to these messages. More generally, the crucial goal of choosing a confounding function is to find the most persuasive accounts of selection bias and tailor the confounding function to address those accounts. In this way, both the researcher and the critic have important roles to play in the design of sensitivity analyses.

3.4 *Reparameterization of the Confounding Function*

Although the q function is a useful and simple summary of confounding, it is helpful to augment our intuition about its magnitude. Currently, q reports mean differences in the potential outcomes, but it is difficult to know if these differences are large or small. Analysts need a good basis for comparison to judge the magnitude. In this section, I introduce an alternative parameterization of q that allows for researchers to compare the importance of the confounding relative to the importance of observed covariates. The key insight is that each confounding function implies a share of the potential outcome variance due to unmeasured confounding. This share provides intuition about the parameters of the confounding function. Combining the information about the variance explained with the direction of the selection bias helps assess how various departures from no unmeasured confounding will affect the estimates.

In the spirit of Imbens (2003), I reparameterize the q function in terms of the proportion of variance explained by selection bias. To see how this works, first define the proportion of potential outcome variance due to X and A under function q as

$$R^2_q(X_i, A_i) = 1 - \frac{\text{var}[Y_i(0)|X_i, A_i, q]}{\text{var}[Y_i(0)]}. \tag{14}$$

Here, I use $Y_i(0)$ instead of Y_i because, under ignorability, A_i should explain none of the variance in the potential outcomes. And, unless the confounding function varies by treatment status, using $Y_i(0)$ will have the same results as $Y_i(1)$. Compare this value to the variance explained simply by X_i:

$$R^2_q(X_i) = 1 - \frac{\text{var}[Y_i(0)|X_i, q]}{\text{var}[Y_i(0)]}. \tag{15}$$

With these two values in hand, we can calculate the portion of the unexplained variance in $Y_i(0)$ due to A_i alone:

$$R^2_q(A_i) = \frac{R^2_q(X_i, A_i) - R^2_q(X_i)}{1 - R^2_q(X_i)}. \tag{16}$$

This partial R^2 is the amount of the unexplained variance in the potential outcomes that is due to selection. One can compare this to the partial R^2 values for individual covariates.[4] Thus, there is some basis of comparison for the magnitude of the confounding. Note that $R^2_q(A_i)$ will be 0 when $q = 0$ because, in that case, $Y_i(0) \perp\!\!\!\perp A_i|X_i$, so that A_i will not affect the distribution of the potential outcomes.

[4] If Y is binary, there are methods for partial R^2 values based on a latent-index model. See Imbens (2003) for more details.

It is straightforward to show that $R^2_q(A_i)$ can be rewritten as

$$R^2_q(A_i) = 1 - \frac{\text{var}[Y_i(0)|X_i, A_i, q]}{\text{var}[Y_i(0)|X_i, q]}. \tag{17}$$

There is a simple way to calculate this value for the case where the confounding function is the constant function $q = \alpha(2a - 1)$. For a continuous outcome, this q implies $Y_i(0) = X_i\beta + \alpha A_i + \epsilon_i$. When ignorability holds, α must be 0 and the only difference between $Y_i(1)$ and $Y_i(0)$ is the treatment effect. In addition, let ϵ'_i be the error from the restricted model, $Y_i(0) = X_i\beta + \epsilon'_i$, so that $\epsilon'_i = \alpha A_i + \epsilon_i$. Note that the confounding function summarizes all of the selection bias, so that whereas ϵ'_i clearly depends on A_i, ϵ_i is independent of A_i, conditional on X_i. Under this model, we can write

$$R^2_q(A_i) = 1 - \frac{\text{var}[\epsilon_i]}{\text{var}[\epsilon'_i]} \tag{18}$$

$$= \frac{\text{var}[\epsilon'_i] - \text{var}[\epsilon_i]}{\text{var}[\epsilon'_i]} \tag{19}$$

$$= \frac{\text{var}[\alpha A_i + \epsilon_i] - \text{var}[\epsilon_i]}{\text{var}[\epsilon'_i]} \tag{20}$$

$$= \frac{\alpha^2 \text{var}[A_i]}{\text{var}[\epsilon'_i]}. \tag{21}$$

Obviously, different q functions would lead to slightly different functional forms here. This $R^2_q(A_i)$ value, though, only shows the magnitude of the selection bias, not the direction. To show this, simply combine α and $R^2_q(A_i)$:

$$R^2_\alpha(A_i) = \text{sgn}(\alpha) R^2_q(A_i). \tag{22}$$

This reparameterization depends on the model of $Y_i(0)$ conditional on X_i: $R^2_\alpha(A_i)$ represents the effect of selection compared to this baseline model.[5] And since there are no restrictions on $Y_i(1)$, the reparameterization also places no restrictions on the treatment effect.

Up to this point, these variances have been hypothetical; an analyst never observes $Y_i(0)$ for any individual unit. By consistency, though, $Y_i = Y_i(0)$ for units with $A_i = 0$. Further, under the assumption that the q function is correct, $E[Y_i(0)] = E[Y^q_i]$. Thus, a regression of Y^q_i on and X_i among those with $A_i = 0$ recovers an estimate of $\text{var}[\epsilon'_i]$. Every value of q implies a calculable variance of the potential outcomes that is due to unmeasured confounding. Furthermore, the variance explained by each covariate provides a baseline to gauge how serious confounding is. For instance, if a researcher shows that confounding would have to explain double the variance explained by the most influential covariate to overturn her result, she would have a rather robust result.

A useful way to show the results of this sensitivity analysis is to simply plot the directional $R^2_\alpha(A_i)$ on the x axis and the implied treatment effects and their confidence intervals on the y axis. A bootstrap approach is useful, though time-consuming, for calculating these confidence intervals. Brumback et al. (2004) also suggest the possibility of using a sandwich estimator for standard errors in this setting.

4 Illustrations

I now turn to providing three examples of this method in practice. Each of these examples uses a different estimation strategy, and yet the selection bias approach to sensitivity analysis works in each case.

[5] One can use an alternative scaling such as dividing by the standard deviation of $Y_i(0)$ to eliminate this baseline model and still retain comparability.

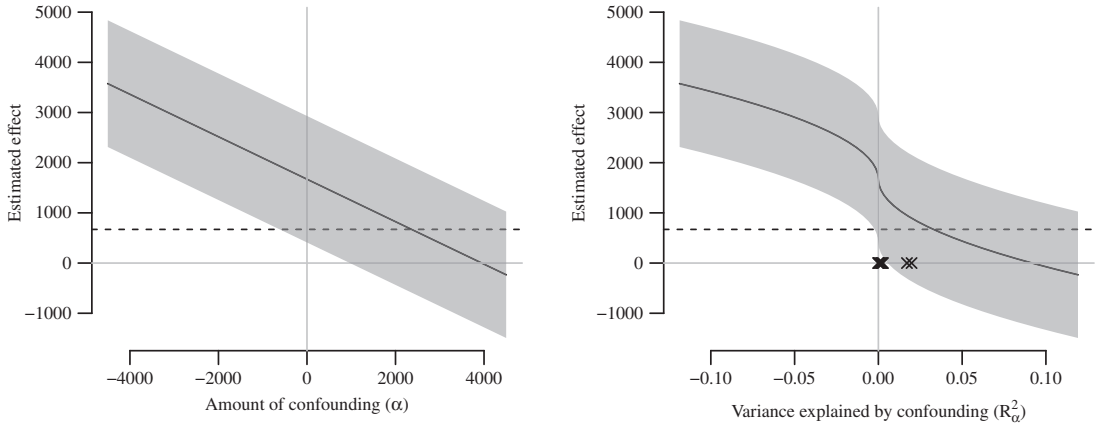

Fig. 1 Sensitivity analysis of the LaLonde (1986) data on the effect of a job-training program. The left panel plots the effect as a function of the raw confounding—that is, in the units of the dependent variable. The right panel shows the same effects as a function of the direction of confounding multiplied by the proportion of unexplained variance explained by the confounding. The × symbols are the partial R^2 for the covariates.

4.1 *Regression Illustration: Job-Training Program*

To get a sense for how the confounding function works in a well-studied case, I first look at a job-training program first analyzed by LaLonde (1986) and subsequently by many authors, especially on the topic of matching estimators. The goal of this experiment was to evaluate the effectiveness of a job-training program on subsequent wages. In the experiment, the estimated effect of the program is \$1794, with a standard error of \$633. Imbens (2003) applies his sensitivity analysis approach to the LaLonde data to see how much variation in Y_i and A_i an unmeasured confounder would have to explain in order to change the estimated effect by \$1000.

I run the above analysis on the experimental data from LaLonde (1986), using a regression to control for observed covariates. For this analysis, I choose the confounding function $q = \alpha(2a - 1)$, which assumes that treated units are either better off ($\alpha > 0$) or worse off ($\alpha < 0$) in terms of earnings. In this case, we are probably most interested in this one-sided deviation from ignorability since people that enroll in job-training programs are likely to have higher levels of inherent motivation and ability than those who choose not to enroll. Alternatively, if the job-training program was tailored specifically to the treated group, alignment bias might be more plausible. However, since this was a broad program meant to help as many people as possible, this type of tailoring might be less of a concern. In general, though, it is crucial to consider these types of concerns when choosing a confounding function.

Figure 1 show the results of this analysis, with several notable features. First, I plot the results as a function of both α (left panel) and $R^2_\alpha(A_i)$ (right panel) to show the difference between the two. As expected, the estimated effect is a linear function of α and a nonlinear function of the variance explained. The reparameterization here is quite helpful. Without any more information it is difficult to assess how large the various values of α are relative to (1) the distribution of the dependent variable and (2) the relative impacts of other variables. With the $R^2_\alpha(A_i)$ approach, it is straightforward to plot the covariate partial R^2 values (× on Fig. 1) and there is immediate comparability on both of these dimensions.

Second, the right panel demonstrates that the selection bias approach maintains the major results of Imbens (2003). Namely, this sensitivity analysis finds that selection accounting for roughly 3%–3.5% of the unexplained variance in $Y_i(0)$ would decrease the point estimate by \$1000 (the horizontal dashed line in Fig. 1). On a similar note, Imbens (2003) finds that a single confounder explaining 10%–20% of the variance in treatment assignment would have to explain 2%–4% of the outcome variance in order to change the estimated treatment effect by \$1000.

Obviously, the above confounding function only has one parameter compared to the two parameters of the Imbens approach. Each value of the confounding function, though, implies some combination of the Imbens parameters. To see this, imagine there is an unmeasured confounder, U_i. The Imbens approach allows the relationship between Y_i and U_i to vary independently of the relationship between A_i and U_i. The confounding function moves these relationships together: both get stronger or both get weaker as α changes. Seen in this light, the confounding approach is conservative as a direct replacement to the Imbens approach, since it never allows for more "robust" combinations of the Imbens parameters, where one relationship is fixed and the other allowed to vary. This is why both approaches will generally come to the same conclusion.

The selection bias analysis does, however, provide more information than the Imbens (2003) approach, showing that the positive benefits of the job-training program disappear when $\alpha > 0$, or when treated units tend to have higher incomes. In fact, looking at confidence intervals and statistical significance, these results are quite sensitive: confounding of less than 0.05% in this direction would make the treatment effect insignificant at typical levels.[6] Thus, this method provides both the severity and the direction of the confounding needed to overturn the observed results, giving researchers a broader and more comprehensive picture of the sensitivity of the results.

4.2 *Matching Illustration: Female Judges and Panel Effects*

In the literature on matching, there has been a vigorous debate over the specific quantity of interest under investigation. It is well known that matching procedures that keep all treated units and only some control units identify the ATT. Unless treatment effects are constant, the ATT is not, in general, equal to the ATE. Thus, in any matching analysis, a researcher has a choice in interpretation: assume constant effects and estimate the ATE or assume no constant effects and estimate the ATT. Crucially, the estimator for these two scenarios is exactly the same, so that these differences are matters of interpretation and assumption, not matters of procedure. It appears that the constant effects assumption and therefore the choice of estimand makes little difference for any causal inferences. These equivalences break down, though, in the face of unmeasured confounding.

To demonstrate how the choice of estimand can affect the sensitivity of estimates, I apply the above methods to the analysis of Boyd, Epstein, and Martin (2010). Their analysis investigates the effect of mixed-gender appellate judge panels in the U.S. Court of Appeals on the vote of male judges on those panels. In particular, they seek to estimate the effect of having at least one woman colleague on a panel on the votes of the male panel members. Thus, in this case, $A_i = 1$ corresponds to male judges on appellate panels with at least one female member and $A_i = 0$ are male judges on appellate panels with all men. The dependent variable, Y_i, is whether or not the male judge voted in a liberal direction on a sex discrimination case. To uncover these effects, Boyd, Epstein, and Martin (2010) perform nearest-neighbor matching on the propensity score (Rosenbaum and Rubin 1983; Ho et al. 2006) after matching exactly on Court of Appeals circuit and decision year of the case. In their matching analysis, they keep all treated units and match them to (multiple) control units with replacement. This procedure identifies the ATT and, under the assumption of constant treatment effects, also identifies the ATE.

Boyd, Epstein, and Martin (2010) choose to interpret their results as the ATE, implicitly assuming constant effects, but this assumption has strong implications for the sensitivity of their results to violations of ignorability. Figure 2 shows the results of the above sensitivity analysis for the two different parameters, the ATT and the ATE, and two different confounding functions. The first tests against one-sided bias and has $q = \alpha(2a - 1)$, so that $Y_i(1)$ is higher (lower) for the treatment group when $\alpha > 0$ ($\alpha < 0$). Suppose that male judges on panels with women are more likely to be liberal due to selection—say, because senior status judges in more liberal circuits with more women are more likely to choose to sit on sex discrimination cases.

The other confounding function tests against alignment bias and has $q = \alpha$, so that the observed arrangement of mixed-gender panels produces more liberal voting (i.e., higher $Y_i(0)$ and $Y_i(1)$) than

[6] This result is consistent with Keele (2010), who performs a sensitivity analysis in the tradition of Rosenbaum (2002).

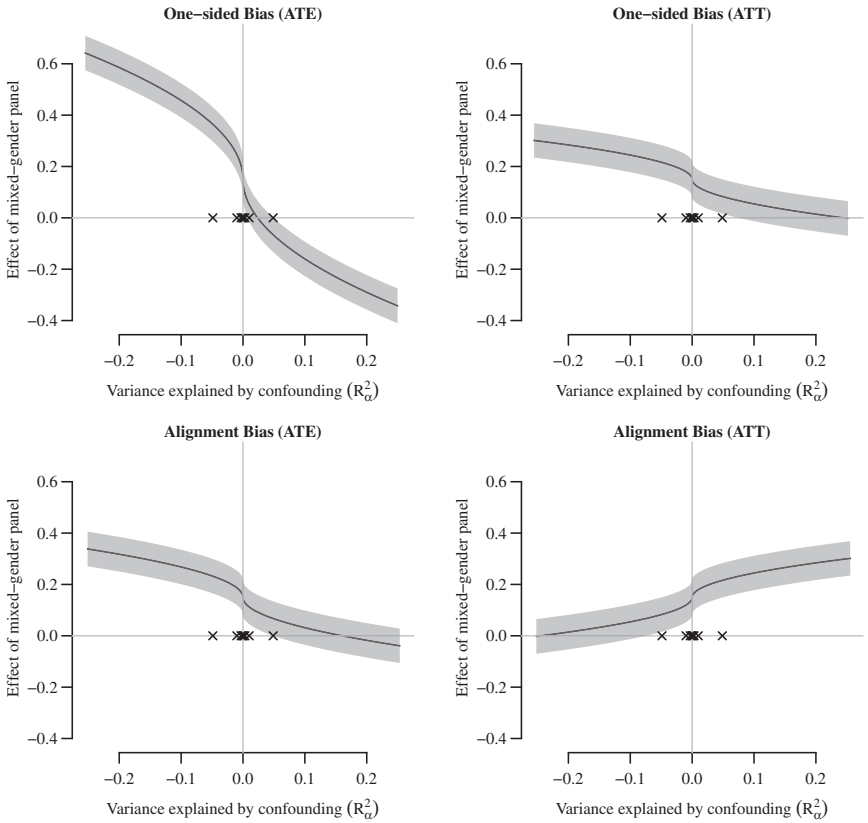

Fig. 2 Sensitivity analysis of the Boyd, Epstein, and Martin (2010) data on the effect of mixed-gender panels in the U.S. Court of Appeals. The left panels plot the sensitivity for the ATE, and the right panels plot the sensitivity for the ATT. One-sided bias occurs when panels without women are more likely to be conservative, and alignment bias occurs when the observed gender on panels produces the most liberal outcomes compared to the reverse. The choice of confounding function and estimand may lead to dramatically different sensitivities. The × symbols are the partial R^2 for the covariates.

if the arrangement was reversed. Reversing the treatment here would put the observed males on same-sex panels on mixed-gender panels instead and vice versa. In this case, the treated units are aligned (or misaligned if $\alpha < 0$) with higher values of the outcome. This might occur if male judges on mixed-gender panels would have been more conservative with an all-male panel than those on all-male panels in the data. On the other hand, those control judges would not have been as liberal as the treated units are observed to be. This could be because judges that are more susceptible to influence by female panel members are more likely to sit on panels with women. That is, the treatment effect might be higher for judges that sit on panels with women. Although this might be less plausible in the case of judges and votes, this alignment bias could be very important in studies where the treatment is thought to help the units under study.

Under one-sided bias, the effect of the ATE becomes statistically insignificant, with confounding explaining just 0.5% of the unexplained variance. Thus, the results for the ATE to this type of ignorability violation are very sensitive. And yet if the ATT is the parameter of interest, the estimates are much less sensitive: the confounding would have to explain 10% of the variance to overturn the statistical significance. More interesting are the results for alignment bias, where the ATE is slightly less sensitive and the ATT sensitivity actually reverses. Positive alignment bias implies a decline in the ATE toward zero, but an increase in the ATT. This switch is a result of the assumptions involved—the ATT requires only ignorability among the control units, whereas the ATE requires ignorability over all units. Investigating sensitivity for the ATT, one only has to check the control units, which can push the qualitative results of the sensitivity analysis far afield, especially if the violations of ignorability imply differential treatment effects as they do under

alignment bias.[7] What is important here is that although the choice of assumption and parameter may leave the main estimates unchanged, they have strong consequences for the broader implications and sensitivities of causal effects.

It is important to note that in this case we are not choosing between two confounding functions, but rather, we are investigating how the estimated effect varies due to these two types of ignorability violations. In order to keep the presentation and interpretation simple, this approach fixes one type of bias at zero, while allowing the other to vary. In principle, both types of bias might be present at the same time, which might amplify or dampen the estimated biases. Detecting more complicated biases would require more complicated confounding functions.

4.3 *Weighting Example: Dynamic Causal Inference and the Effect of Negativity on Turnout*

A core question in the study of American politics is what inspires or discourages citizens to turn out to vote. Many scholars focus on the question of how a campaign, and specifically the tone of campaign advertising, can affect electoral participation.[8] Observational studies of turnout rely on summaries of the overall campaign advertising tone and its effect on the percent turnout, controlling for various aspects of the candidates and the campaign itself (Ansolabehere et al. 1994; Finkel and Geer 1998; Ansolabehere, Iyengar, and Simon 1999; Brooks 2006). This approach, however, ignores the issues of dynamic causal inference (Robins 1999; Robins, Hernán, and Brumback 2000; Blackwell 2013a) that lead to serious biases that matching and regression cannot solve. In this illustration, I analyze new data to show that the above framework adapts easily into the dynamic setting.

To investigate both the effect of negativity on turnout and the sensitivity of this effect, I use data on 176 U.S. Senate and Gubernatorial campaigns from 2000 to 2006. I use a marginal structural model (MSM), combined with inverse probability of treatment weighting (IPTW), to estimate the effect of late-campaign Democratic negativity (i.e., negativity during October and November) on the turnout in the election, conditional on a set of baseline variables.[9] In general, it is acceptable to include these baseline covariates in a regression model of a dynamic treatment on an outcome, but including dynamic confounders can lead to posttreatment bias (Blackwell 2013a). Of course, omitting these confounders ignores their effect on subsequent treatment decisions and can lead to omitted-variable bias. In this example, the percent undecided in a race may be influenced by past negativity if negative ads tend to activate partisan feelings and may also affect the decision for candidates to go negative in the future. This variable is likely also correlated with the final turnout in the election. A variable like this, that both affects and is affected by the treatment, is called a *time-varying confounder*.

Since the addition of time-varying confounders to a marginal structural model would induce bias, I instead remove the effect of these variables by weighting. As shown by Robins, Hernán, and Brumback (2000), weighting by the inverse of the propensity score for the entire treatment history as a function of time-varying confounders will remove the omitted-variable bias of these confounders without introducing posttreatment bias. This result, though, only holds under the assumption of sequential ignorability, the generalization of the ignorability assumption to the dynamic case. Fortunately, the sensitivity analysis approach works even in this case by applying the confounding function to each time period.

Let A_{it} be the treatment in a given period, $\underline{A}_{it} = (A_{i1}, \ldots, A_{it})$ be the treatment history up to time t, and $\underline{A}_i = \underline{A}_{iT}$ be the entire treatment history. Let a, $\underline{a}_t$, and $\underline{a}$ be a representative value of these

[7] A Rosenbaum (2002) style sensitivity analysis indicates that the results become insignificant when $\Gamma > 1.5$. This is a moderate level of sensitivity for social science research (Keele 2010). Of course, this approach lacks any evidence of direction.

[8] Lau, Sigelman, and Rovner (2007) provide a meta-analysis of studies attempting to pinpoint the effects of negative advertising on various political outcomes, including turnout.

[9] The baseline variables here include support in polls for the Democratic candidate after the primary, percent undecided after the primary, whether the Democratic candidate was the incumbent, the *Congressional Quarterly* rating of seat competitiveness, office, campaign length in weeks, and fixed effects for election cycle.

variables, and define similar variables and values for X. This notation helps generalize Assumption 2 to dynamic situations.

Assumption 3 (Sequential Ignorability). *For every treatment history* $\underline{a}$ *and time-period* t, $Y_i(\underline{a}) \perp\!\!\!\perp A_{it} | \underline{X}_{it}, \underline{A}_{it-1}$.

This assumption states that, conditional on the treatment and covariate histories up to t, the treatment status in period t is independent of the potential outcomes. In this setting, the confounding function becomes

$$\begin{aligned} q_t(\underline{a}, \underline{x}_t) = E[Y(\underline{a}) | A_t = a_t, \underline{A}_{t-1} = \underline{a}_{t-1}, \underline{X}_t = \underline{x}_t] \\ - E[Y(\underline{a}) | A_t = 1 - a_t, \underline{A}_{t-1} = \underline{a}_{t-1}, \underline{X}_t = \underline{x}_t]. \end{aligned} \tag{23}$$

Here, q_t represents how the treated and control units differ in some period t, when they share the same treatment and covariate histories up to t. Again, when sequential ignorability holds, then $q_t = 0$. One can write this time-varying confounding function in terms of a single parameter,

$$q_t(\underline{a}, \underline{x}_t; \alpha) = \alpha(2a_t - 1), \tag{24}$$

which implies that when $\alpha > 0$, negative campaigns tend to have higher turnouts than positive campaigns. This might capture some underlying attention or enthusiasm for the race that is not captured in the baseline or time-varying covariates. Brumback et al. (2004) show that for a given confounding function, an adjusted outcome can eliminate the bias due to confounding, just as in the single-shot case. With a time-varying treatment, the adjusted outcome becomes

$$Y_i^{\alpha} = Y_i - \sum_{t=0}^{T} q_t(\underline{A}_i, \underline{X}_{it}; \alpha) \cdot \Pr(A_t = 1 - A_{it} | \underline{A}_{it-1}, \underline{X}_{it}). \tag{25}$$

This is simply the time-varying generalization of equation (7). This adjustment subtracts the sum of the assumed confounding of a treatment history multiplied by the probability of reaching that treatment history. Conveniently, the last term of equation (25) is a function of the time-varying propensity score used in the IPTW estimation.

To calculate the weights, I model A_{it} as a function of past Democrat negativity, Democratic support in the polls at time t, percent undecided at time t, and past Republican negativity in the race.[10] In the weighted MSM, I allow the effect of negativity to vary by the incumbency status of the Democratic candidate and find that an additional week of negative advertising late in the campaign leads to roughly a two percentage-point increase in turnout for Democratic incumbents and no effect for Democratic nonincumbents.[11] The effect for incumbents is statistically significant, and yet one might worry that incumbents going negative is an indication of a more interesting race because of challenger quality or incumbent weakness not captured by polling. Figure 3 shows how deviations from sequential ignorability affect these estimates. The x-axis again is the amount of unexplained variance explained by the confounding. In fact, these results are quite *insensitive*: this confounding would have to explain close to half of the unexplained variance in order to overturn these results. This value is so high partially because the confounding compounds over time, so that even small values of α end up explaining quite large amounts of the variance. Thus, this sensitivity analysis procedure can help support results even in situations fairly far away from the typical regression or matching situations researchers face.

5 Discussion and Conclusion

Following Robins (1999) and Brumback et al. (2004), this article proposes a method of sensitivity analysis that tests specific deviations from ignorability to see how these deviations affect estimates.

[10] I estimated separate weights for incumbents and nonincumbents, with a subset of these variables for either chosen on the basis of which produced the best balance.

[11] A candidate goes negative in a given week if more than 10% of his or her ads mention the opponent.

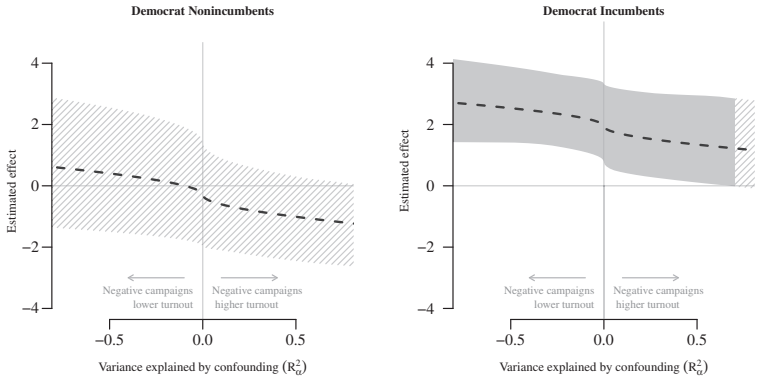

Fig. 3 Sensitivity analysis of the effect of Democratic negativity on turnout in Senate and Gubernatorial elections. Confidence intervals are bootstrapped to account for variation in the weighting model.

This approach is *critique-based*—if one gives an alternative story to the estimated effect, this sensitivity analysis can investigate and respond to that exact story. In addition, I introduce a convenient reparameterization of the confounding function and show how the method works with the three main approaches to causal inference: regression, matching, and weighting. Further, this approach fits easily into the dynamic causal inference framework and can provide insight into how the chosen estimand affects the sensitivity of its estimates.

As with all methods, there are limitations to this approach to sensitivity analysis. First and foremost, it relies on a "selection on the observables" assumption at its core, so that it is incompatible with certain other approaches to causal inference such as instrumental variables. It may be the case, though, that an instrument could provide evidence for the amount of unmeasured confounding. Future research should investigate how these approaches could interact. Second, this approach requires an estimate of the propensity score, which may or may not be part of an analyst's estimation strategy. If it is not, then this requires additional modeling that may be difficult, depending on the empirical problem. Last, demonstrating that a result is insensitive to a specific confounding function over a specific set of parameters does not imply that the estimated effect is truly causal. There could always be confounding that is greater in magnitude than the sensitivity analysis has assumed.

There are many avenues for progress on sensitivity analyses for causal inference. To ease exposition, this article has focused on rather simple functional forms for the confounding function, but the framework itself does not impose these limits. A covariate might affect the degree of confounding in either one-sided bias ($q = \alpha(2a - 1)x$) or alignment bias ($q = \alpha x$). These alternative forms only modify the confounding function and leave the rest of the calculations and intuitions unchanged. Future work should explore how and when these more complex selection biases might affect inferences in the social sciences. Furthermore, the relationship between the estimand and its sensitivity are raised here, but only briefly. The full implications of these results could provide guidance to individuals deciding between different causal quantities of interest.

References

Ansolabehere, Stephen, Shanto Iyengar, and Adam Simon. 1999. Replicating experiments using aggregate and survey data: The case of negative advertising and turnout. *American Political Science Review* 93(4):901–9.

Ansolabehere, Stephen, Shanto Iyengar, Adam Simon, and Nicholas Valentino. 1994. Does attack advertising demobilize the electorate? *American Political Science Review* 88(4):829–38.

Blackwell, Matthew. 2013a. A framework for dynamic causal inference in political science. *American Journal of Political Science* 57(2):504–20.

———. 2013b. Replication data for: A selection bias approach to sensitivity analysis for causal effects. Dataverse Network, hdl:1902.1/21131.

Boyd, Christina L., Lee Epstein, and Andrew D. Martin. 2010. Untangling the causal effects of sex on judging. *American Journal of Political Science* 54(2):389–411.

Brooks, Deborah Jordan. 2006. The resilient voter: Moving toward closure in the debate over negative campaigning and turnout. *Journal of Politics* 68(3):684–96.

Brumback, Babette A., Miguel A. Hernán, Sebastien J. P. A. Haneuse, and James M. Robins. 2004. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine* 23(5):749–67.

Cornfield, Jerome, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, and Ernst L. Wynder. 1959. Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* 22:173–203.

Dawid, A. Phillip. 1979. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)* 41(1):1–31.

Finkel, Steven E., and John G. Geer. 1998. A spot check: Casting doubt on the demobilizing effect of attack advertising. *American Journal of Political Science* 42(2):573–95.

Glynn, Adam N., and Kevin M. Quinn. 2011. Why process matters for causal inference. *Political Analysis* 19(3):273–86.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. Characterizing selection bias using experimental data. *Econometrica* 66(5):1017–98.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2006. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15(3):199.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review* 105(4):765–89.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science* 25(1):51–71.

Imbens, Guido W. 2003. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* 93(2):126–32.

———. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86(1):4–29.

Keele, Luke. 2010. An overview of rbounds: An R package for Rosenbaum bounds sensitivity analysis with matched data. Unpublished manuscript.

LaLonde, Robert J. 1986. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76(4):604–20.

Lau, Richard R., Lee Sigelman, and Ivy Brown Rovner. 2007. The effects of negative political campaigns: A meta-analytic reassessment. *Journal of Politics* 69(4):1176–209.

Manski, Charles F. 1990. Nonparametric bounds on treatment effects. *American Economic Review* 80(2):319–23.

Mebane, Walter R., and Paul Poast. 2013. Causal Inference without ignorability: Identification with nonrandom assignment and missing treatment data. *Political Analysis* 21(2):233–51.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and causal inference. Methods and principles for social research*. Cambridge: Cambridge University Press.

Robins, James M. 1999. Association, causation, and marginal structural models. *Synthese* 121(1/2):151–79.

Robins, James M., Miguel A. Hernán, and Babette A. Brumback. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* 11(5):550–60.

Rosenbaum, Paul R. 2002. *Observational studies*. 2nd ed. New York: Springer.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Rubin, Donald B. 1978. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* 6(1):34–58.