How to See More in Observational Studies: Some New Quasi-Experimental Devices

Paul R. Rosenbaum

Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; email: rosenbaum@wharton.upenn.edu

Annu. Rev. Stat. Appl. 2015. 2:21–48

First published online as a Review in Advance on November 6, 2014

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-010814-020201

Keywords

differential effects, evidence factors, multiple control groups, sensitivity analysis, strengthening an instrumental variable

Abstract

In a well-conducted, slightly idealized, randomized experiment, the only explanation of an association between treatment and outcome is an effect caused by the treatment. However, this is not true in observational studies of treatment effects, in which treatment and outcomes may be associated because of some bias in the assignment of treatments to individuals. When added to the design of an observational study, quasi-experimental devices investigate empirically a particular rival explanation or counterclaim, often attempting to preempt anticipated counterclaims. This review has three parts: a discussion of the often misunderstood logic of quasi-experimental devices; a brief overview of the important work of Donald T. Campbell and his colleagues (excellent expositions of this work have been published elsewhere); and its main topic, descriptions and empirical examples of newer devices, including evidence factors, differential effects, and the computerized construction of quasi-experiments.

1. REDUCING AMBIGUITY IN OBSERVATIONAL STUDIES

1.1. Is It Possible to Avoid Some Mistaken Causal Inferences from Observational Studies?

In 2007, the UK Academy of Medical Sciences published a review of the literature on empirical studies of the causes of disease (Rutter et al. 2007). One section consisted of 23 case studies, famous examples of causal inferences currently believed to be valid (e.g., smoking and lung cancer) or invalid [e.g., hormone replacement therapy (HRT) and coronary artery disease]. The authors wrote the following (Rutter et al. 2007, pp. 68–69):

> The most quoted example of misleading conclusions from non-experimental data concerns the effects of HRT on coronary heart disease . . . Various case-control studies suggested that HRT protected against heart disease . . . , whereas a large-scale randomized controlled trial . . . plus other randomized controlled trials . . . suggested otherwise . . . [I]t was highly likely that there would be major health related lifestyle differences between the women choosing to use HRT and those not doing so, and this major selection effect alone should have made for caution before accepting the non-experimental data as showing causation.

Neel (2002) elaborates on the above reference to selection bias from "lifestyle differences" in "The marketing of menopause" as follows:

> Early films made by drug makers to educate doctors about hormone treatments are a blend of homespun medical wisdom and paternalism, playing on older women's fears . . . Estrogen's heyday started with a book for the masses, titled *Feminine Forever*. Author Robert Wilson was a Manhattan gynecologist with strong financial ties to the hormone makers . . . He won women over with scientific-sounding promises of youth and beauty and good sex, even though the FDA banned Wilson from certain research for making unsubstantiated claims.

To be specific: A woman concerned with the effects of aging might seek treatment with HRT, but she also might be more prone to exercise and diet to avoid obesity, and of course diet, exercise, and obesity are important to the risk of coronary artery disease. A woman who is receiving HRT is likely having checkups that include assessment of blood pressure and cholesterol, so it is less likely that she has untreated extreme hypertension, another important factor in coronary outcomes. And so on. In an experiment or clinical trial, these kinds of biases are avoided by randomly assigning treatments to individuals (Fisher 1935).

The observational studies and randomized experiments of HRT prompted a small literature asking the following question: "What went wrong in the observational studies?" The reader is referred to Hernán et al. (2008) for one interesting paper in this literature. I do not wish to add to this literature; however, one comment about one of the larger, more prominent observational studies serves as a prelude to the general topics discussed in the current article. The observational study of HRT by Grodstein et al. (1996) studied 59,337 women using age-adjusted incidence rates and a proportional hazards model to adjust for a number of measured covariates including age, body mass index (BMI), and smoking, among others. A proportional hazards model is a type of regression model. In the model used by Grodstein et al. (1996), the coefficient for treatment with both estrogen and progestin predicted lower risk among the treated when compared with no HRT. Grodstein et al. (1996, p. 453) lent a causal interpretation to this regression coefficient and concluded that "the addition of progestin to estrogen does not appear to attenuate the cardioprotective effects of hormone therapy in relatively young postmenopausal women." A subsequent

large randomized trial (Women's Health Initiative 2002) stopped this treatment early, concluding the following (p. 321): "Overall health risks exceeded benefits from use of combined estrogen plus progestin . . . . " Notably, the observational study lent a causal interpretation to a regression coefficient, but it did not perform any quantitative analysis with data to speak to the issue of unmeasured biases, namely, biases from unmeasured covariates not in this regression. Unlike the smoking and lung cancer studies (Cornfield et al. 1959; Rutter et al. 2007, p. 45), the observational study of HRT did not conduct any analysis of sensitivity to unmeasured biases. In addition, it did not use quasi-experimental devices in an effort to detect specific unmeasured biases if present or rule them out if absent. It is not now and never will be known whether quasi-experimental efforts would have succeeded, but the efforts were not made.

Fitting regressions is familiar. We learn about regression early in studying statistical methods. Most beginning students in a PhD program in statistics or biostatistics take a course covering various forms of regression during their first year of graduate study, so regression might account for a sixth of this first year of study. Thus, regression becomes an important topic in the PhD qualifying examination. Even students without graduate degrees in statistics or biostatistics are likely to encounter regression methods in simplified courses, that is, courses that may omit discussion of the mathematical theory that forms much of the basis of statistics. What we learn early and what we have done often feels familiar.

Fitting regressions is familiar. What is familiar can feel safe, but familiarity and safety are different things. There is nothing safe in fitting regressions to observational data and interpreting coefficients as effects caused by treatments: Grave errors are commonplace, perhaps typical. It does no good to append a claim that you have included in the regression all relevant covariates, a claim that there are no unmeasured confounders and that you could not be mistaken in making this claim. Who are you that you could not be mistaken?

1.2. The Inescapable Nature of Observational Evidence

Evidence linking smoking with disease was challenged by the tobacco industry. The response was a steady stream of stronger evidence intended to rebut that challenge. Ultimately, the evidence linking smoking and lung cancer was extremely strong. By contrast, the claim that HRT had unintended health benefits was congenial even to people who had no stake in the matter. Many, perhaps most, people who knew of the observational studies of HRT were surprised when the randomized trials found harm where the observational studies had found benefit.

There is, however, no basis for surprise. The absence of randomization means that associations between treatments and outcomes in observational studies are inescapably ambiguous. The quantitative degree and qualitative nature of the ambiguity varies from study to study, but the presence of a measure of ambiguity is inescapable. In the analysis of an observational study, candor demands that the quantitative degree of ambiguity be reported, and, as conventional confidence intervals and tests do not do this, some form of sensitivity analysis is required (e.g., Cornfield et al. 1959; Rosenbaum & Rubin 1983; Rosenbaum 1988, 1991, 2002, 2007, 2010a; Robins et al. 1999; Gastwirth 1992; Manski & Nagin 1998; Imbens 2003; Diprete & Gangl 2004; Hosman et al. 2010; Liu et al. 2013; Stuart & Hanna 2013). In contrast, the goal in study design is to appraise ambiguities from earlier studies and then set out to reduce them in the current study.

1.3. What Are Quasi-Experimental Devices?

After adjustments have been made for measured covariates in an observational study, an association between treatment and outcome is ambiguous: An association may be an effect caused

by the treatment, or it may be an unmeasured bias in the way treatments were assigned. Quasi-experimental devices enlarge the set of considered associations with the intention of reducing this ambiguity. The oldest, most basic, and most familiar quasi-experimental devices include pretreatment measures of the outcome and multiple control groups (see Section 2). Devices are typically selected to address particular rival hypotheses or counterclaims, and a device is successful if it eliminates a particularly plausible rival hypothesis or counterclaim. It is virtually inconceivable that quasi-experimental devices could eliminate all possible counterclaims, but setting aside such unrealistic goals, practical devices can achieve practical progress.

Sensitivity analyses in observational studies measure the degree of ambiguity that is present, whereas quasi-experimental devices attempt to reduce the ambiguity. Although these are quite distinct tasks, they are not unrelated. If the most plausible sources of large biases can be anticipated, as the selection biases in the HRT studies were, then quasi-experimental devices may shed light on those particular biases. In observational studies, it is never possible to be sure that there are no small biases in treatment assignment, but a sensitivity analysis may reveal that no pattern of small or moderate biases could explain away the ostensible effect of the treatment. Sensitivity analysis can speak to unmeasured biases of any possible form, but not of any magnitude, whereas quasi-experimental devices are more likely to reveal large biases of anticipated form and to miss small biases of unanticipated form. In an observational study, it is not a small achievement to have rendered implausible the anticipated sources of large biases and to have demonstrated insensitivity to small and moderate biases, even though there remains the logical possibility of large unanticipated or implausible biases.

1.4. Reducing Ambiguity, Review of Traditional Methods, and Review of Recent Developments

Section 1.5 discusses the logic of reducing ambiguity without eliminating it; this logic is often misunderstood, with utopian ambitions blocking practical progress. When replicating an observational study, one hopes to remove or vary a particular source of unmeasured bias that made previous studies ambiguous (see Section 1.6). Quasi-experimental devices often set out to achieve the same objective within a single study; such devices effectively conduct several studies with different ambiguities within a single investigation. Section 2 briefly reviews traditional quasi-experimental devices and the work of Donald T. Campbell and colleagues. This review is brief, however, because several excellent expositions of these ideas have already been published (Campbell & Stanley 1963, Cook & Campbell 1979, Shadish et al. 2002). Three recent developments are reviewed in Sections 3, 4, and 5: evidence factors, differential effects to control generic unmeasured biases, and computerized constructions of quasi-experiments.

1.5. The Often Misunderstood Logic of Quasi-Experimental Devices

Quasi-experimental devices reduce ambiguity but do not eliminate it, making a study with effective quasi-experimental devices open to fewer credible challenges or counterclaims, not closed to credible challenges. In any conventional way of speaking, reducing ambiguity sounds like progress, perhaps less progress than we might want, but progress nonetheless. This ambiguity is not due to a limited sample size; it does not vanish as the sample size of the current study increases without bound (see Cochran 1965). Perhaps the ambiguity of one study may be reduced by the results of a different sort of study. For instance, Rutter et al. (2007, p. 55) note that the evidence from observational studies of smoking and cancer in humans "was much strengthened by animal studies that showed the carcinogenic effect of the tars involved in cigarette smoking."

The statistical literature can be misread to say that only the elimination of ambiguity, not its reduction, is acceptable. Such a misreading might result in skepticism about quasi-experimental devices that reduce but do not eliminate ambiguity. Although this misreading of some theoretical statements is a natural one, the statements were invariably made about some abstract problem without reference to observational studies, and it is never clear that the author of the misread statement meant to say anything about observational studies. For instance, Basu's (1983, Vol. 4, p. 2) entry on "Identifiability" in the *Encyclopedia of Statistical Sciences* says the following: "Before any inferential procedure can be developed, one needs to assert that the unknown parameters are identifiable." The entry then defines identifiable to mean that different states of the world, or different values of a parameter θ in a parameter space Θ, yield probability distributions for observable data that are themselves different. One could misread this statement as saying that we learn nothing about θ unless there is identification, nothing unless there is a consistent test for each value of θ, nothing unless there is a consistent estimate of θ. More careful than most, Basu is clearly aware that we often learn about nonidentified situations. Immediately after defining the term "identifiable," he observes that a problem may be partially identified: A many-to-one function, say $f(\theta)$, may be identified although θ is not (see also Manski 1995). In nonpathological situations, to be able to consistently estimate $f(\theta)$ is to be able to say something about θ, even if there is no consistent estimate of θ. For instance, if $\mathcal{C}$ is a 95% confidence set for $f(\theta)$, then the preimage $f^{-1}(\mathcal{C})$ of $\mathcal{C}$ under f is a 95% confidence set for θ. However, $\mathcal{C}$ may shrink to a point as the sample size increases, although $f^{-1}(\mathcal{C})$ does not. We might say that a study design for inference about θ is inconclusive if $f^{-1}(\mathcal{C})$ converges with increasing sample size to a set containing at least two elements from Θ and informative if $f^{-1}(\mathcal{C})$ converges to a proper subset of Θ. Speaking in this way would mean admitting that the collection of inconclusive but informative study designs could be very large, with some study designs being much better than others. Such an admission would in turn allow us to speak of preferring a greater reduction in ambiguity even when ambiguity cannot be eliminated.

To sharpen the issue as it relates to quasi-experimental devices, consider a few trivial illustrations in schematic form. If there are three possible states of the world or values for a parameter θ, say $\Theta = \{\theta_1, \theta_2, \theta_3\}$, and if in large samples, the study design can rule out $\theta = \theta_3$ if false and $\theta \in \{\theta_1, \theta_2\}$ if false, then the design is informative, even if the data fail to provide any consistent estimate of θ or consistent tests of $H_0 : \theta = \theta_1$ and $H_0 : \theta = \theta_2$. Suppose there are two investigations or two study designs such that one could distinguish $\theta = \theta_3$ and $\theta \in \{\theta_1, \theta_2\}$ but could not distinguish between θ_1 and θ_2 and the other could distinguish θ_1 and θ_2 but could not eliminate θ_3 [i.e., the latter design could demonstrate the truth of the implication $(\theta = \theta_1) \vee (\theta = \theta_2) \Rightarrow (\theta = \theta_1)$ but not the truth of its premise $(\theta = \theta_1) \vee (\theta = \theta_2)$]. Taken together, these two investigations are conclusive, or identified, although each is individually inconclusive. Thus it cannot be the case that inconclusive data carry no information and are without value. Indeed, if these two investigations were the only possible investigations, then two inconclusive investigations would lead to a firm conclusion, whereas eliminating inconclusive investigations would also eliminate the possibility of a firm conclusion.

To continue the trivial illustration, suppose there were four rather than three possible states of the world or values of θ, say $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$, and there are three study designs, all of which are unable to distinguish θ_1 and θ_2. The first design can consistently distinguish among $\{\theta_1, \theta_2\}$, θ_3, and θ_4. The second design can consistently distinguish between $\{\theta_1, \theta_2, \theta_3\}$ and θ_4. The third design can consistently distinguish between $\{\theta_1, \theta_2\}$ and $\{\theta_3, \theta_4\}$. All three of these designs are inconclusive; design one is best in that it distinguishes the most, and designs two and three cannot be ordered to say that one distinguishes more than the other. However, designs two and three done together can do the work of design one if design one is impractical. With large samples, design two done twice, replicated exactly, is less informative than design two done once together with design three done

once. To recognize that inconclusive designs may produce valuable information and that study designs vary in the degree and manner in which they are inconclusive is to develop a preference for stronger study designs, designs that can settle more even if they cannot settle everything. Moreover, this recognition also means developing a preference for variety in inconclusive study designs, particularly when available designs cannot be ordered as more or less informative, as in the case of designs two and three above. There is nothing in statistical theory that opposes a reduction in ambiguity merely because it is not possible to eliminate ambiguity.

The issue of reducing ambiguity without eliminating it is not unique to observational studies. This issue also arises in fractional factorial experiments, which are ambiguous because certain effects are aliased with other effects. The art of designing such an experiment is to minimize worrisome ambiguities; for instance, one might attempt to minimize aliasing of main effects with each other or with two-factor interactions. The reader is referred to Hedayat et al. (1999) and Wu & Hamada (2011) for further discussion.

1.6. Goals of Replication and Their Relationship to Quasi-Experimental Devices: What Are We Seeking When We Seek More Evidence?

Replication of an observational study is intended to strengthen evidence by providing more of it, but what sort of additional evidence is needed? Additional evidence is valuable to the degree that it speaks to ambiguities in the available evidence. If the principal ambiguity in the available evidence stems from a limited sample size, then an exact replication of the initial study that increases the sample size might be of value. Many observational studies are large, however, so a limited sample size is not the main source of ambiguity (see Cochran 1965). When the principal source of ambiguity comes from unmeasured biases, a valuable replication tries to study the same treatment effect in the presence of different potential sources of unmeasured biases. If repeatedly varying the most plausible sources of bias does little to alter the ostensible effects of the treatment, then the evidence in favor of an actual treatment effect is gradually strengthened [as, for instance, in works by Susser (1987, p. 88), Rosenbaum (2001a), and Rutter et al. (2007, section 5.4.6)].

The central problem in observational studies—the problem that defines an observational study and distinguishes it from a randomized experiment—is that treatments were not randomly assigned. Given this, investigators often evince a remarkable lack of interest in how treatments were assigned. One goal in replication is to find a setting in which the same treatment is assigned by a different mechanism, one that is subject to different biases than those in the original study. As discussed in Section 1.5, this may reduce ambiguity without eliminating it. If similar results are observed for the same treatment with several different treatment assignment mechanisms, then explaining the observed results as a bias and not a treatment effect entails claiming that each of several different mechanisms yields the same biases. This claim is not impossible, but it becomes less plausible as the same treatment is studied in more and more settings thought to be susceptible to different biases.

To illustrate, many studies have claimed various health benefits result from eating fish. The typical study compares people who eat more or less fish. In many countries, fish is a comparatively expensive food, often favored by those seeking to eat a healthy diet, resulting in several potential sources of selection bias when comparing people who typically eat more or less fish. In Manhattan, the people regularly grabbing a burger in McDonald's form a different group of people from those regularly eating sushi, not in all cases, but in many cases. These different groups in Manhattan are doing other things differently besides eating burgers or sushi. If there are many studies of people who happen to eat more or less fish, then ask: Is the best replication to do another one? Or should one look for the same effect in a place where the selection bias is likely to be different? With these

issues in mind, Lund & Bønaa (1993) compared the health outcomes of the wives of fisherman in Norway with those of the wives of unskilled workers. Their thought was that wives of fisherman eat large quantities of fish but do so for reasons that are very different from the reasons people buy large quantities of sushi in Manhattan. To find similar effects of eating fish among groups that eat fish for different reasons is to gradually make selection bias less plausible as an explanation of the association between fish consumption and health outcomes.

Better replications of observational studies attempt to vary the likely sources of unmeasured biases in order to see whether the ostensible effect of the treatment comes and goes as one selection bias is replaced by another or whether the same treatment effect is seen as selection biases are varied. Quasi-experimental devices often attempt to achieve a similar kind of variation within a single study. This is true in various ways of multiple control groups, baseline measures, evidence factors, and differential effects.

2. TRADITIONAL QUASI-EXPERIMENTAL DESIGNS

2.1. The Distinctive Work of Donald T. Campbell

In 1957, Donald T. Campbell published "Factors relevant to the validity of experiments in social settings," which opens by asking the following questions (p. 297):

> What do we seek to control in experimental designs? What extraneous variables which would otherwise confound our interpretation of the experiment do we wish to rule out? The present paper attempts a specification of the major categories of such extraneous variables and employs these categories in evaluating the validity of standard designs . . .

Campbell's paper continues, saying the following: Here's a common problem, and here's a simple study design resistant to that problem, but, of course, vulnerable to another common problem. Here, however, is an improved yet simple design that is resistant to both previous problems but, of course, vulnerable to a third problem, and so on. In the same style, quasi-experimentation expanded to three important and influential books with collaborators Julian Stanley (Campbell & Stanley 1963), Thomas Cook (Cook & Campbell 1979), and William Shadish (Shadish et al. 2002). The reader is also referred to articles by Campbell & Boruch (1975), Meyer (1995), Stuart & Rubin (2008), West et al. (2008), Imbens & Wooldridge (2009), and Nagin & Weisburd (2013).

Campbell's work is distinctive in two ways. First, papers in many fields that conduct observational studies contain lists of potential biases with colorful names, such as "confounding by indication" or the "Will Rogers effect," but unlike most of these lists of problems, each problem on Campbell's list is accompanied by a solution. Second, when practical problems are clear and intractable, as is sometimes true in nonexperimental settings, investigators are often drawn to the use of subtle and complex methodology that succeeds in rendering these problems less clear. In contrast, in an observational study, a quasi-experimental device is intended to be a simple addition, the usefulness and limitations of which are evident upon casual inspection. These devices do something, not everything, and the intention is that both what these devices can do and what they fail to do are clear.

2.2. Can Quasi-Experimental Devices Reduce Ambiguity About Unmeasured Biases?

Two of the many devices prominent in the work of Campbell and colleagues are multiple control groups and baseline measures of the outcome. The following example is the simplest case. After

adjusting for measured pretreatment covariates, a difference in outcomes between treated and control groups may not be an effect caused by the treatment: It may instead be an unmeasured bias in the formation of treated and control groups. A change in outcomes from before treatment to after treatment may not be an effect caused by the treatment: It may instead reflect some process of change apart from the treatment, such as growth, maturation, or decay. The simplest quasi-experimental design, the triangle design, reduces ambiguity somewhat by looking at two ambiguous comparisons simultaneously: difference versus controls and change from baseline. The triangle design is so named because it makes three pairwise comparisons among three quantities: (*a*) measures for untreated controls, (*b*) pretreatment measures for treated subjects, and (*c*) posttreatment measures for treated subjects. This design is used when there is no natural anchor for time among controls. If controls are expected to change during the time from before treatment to after treatment in the treated group, it is common to replace a single control measure with two, one timed before, and the other timed after. This design is sometimes called a pretest–posttest control group design or a difference-in-differences design. Many variations on this theme exist; for instance, the reader is referred to work by Campbell & Stanley (1963, table 2, p. 40) or Rosenbaum (2001b).

Masjedi et al. (2000) asked whether the powerful drugs used to treat tuberculosis have the unintended side effect of causing genetic damage. As one cannot ethically give such drugs to healthy individuals or deny the drugs to individuals with tuberculosis, a randomized trial is not possible. Masjedi and colleagues (2000) looked at various measures of genetic damage, including the frequency of chromosome aberrations excluding gaps per 100 lymphocytes in blood samples. In particular, they compared 36 patients with tuberculosis before and after treatment and 36 untreated controls matched for age and gender. Two potential biases are obvious: (*a*) the patients had tuberculosis and the controls did not, and (*b*) there is no reason to think that the tuberculosis infection was static, and changes in the disease over time are confounded with the treatment. Thus, the triangle design is a small step up from a before–after study because it includes matched untreated controls, and it is also a small step up from a treated-versus-control matched pair study because it includes baseline measures for treated subjects. There is no distinction between before and after for the controls.

Figure 1 depicts the results of this study. **Figure 1*a*** shows the frequencies of chromosome aberrations for 36 patients before and after treatment, as well as those for 36 age- and gender-matched untreated controls. The statistical analysis takes account of this matching by focusing on matched pair differences. **Figure 1*b*** shows the 36 matched pair differences between each pair of two groups. A visual inspection of **Figure 1** shows that, in terms of the outcome of chromosome aberrations, tuberculosis patients and healthy controls looked fairly similar before the start of tuberculosis drugs, but they looked very different afterward. For instance, the median pair difference before treatment was 0.08, but it was 0.92 after treatment. Despite appearances, however, the tuberculosis-minus-control difference before treatment is statistically significant with a one-sided p-value of 0.023 in Maritz's (1979) randomization test using Huber's M-statistic. Logically, this significant difference cannot be an effect of tuberculosis drugs that have not yet been given, and it must indicate some form of bias, a bias that could easily also affect the comparison after treatment. There is, however, a quantitative dimension. A moderate bias in treatment assignment could explain the difference before treatment, but only a much larger bias could explain the much larger difference after treatment. Within-pair treatment assignment probabilities that deviated from that of randomization, $0.5 = 1/2$, in the interval $[0.4, 0.6]$ could yield a one-sided p-value as large as 0.12 in the comparison before treatment, but the largest one-sided p-value that a bias of that same magnitude could produce in the after-treatment comparison is 0.00021. Indeed, after treatment, to obtain a one-sided p-value above 0.05 without a treatment effect, the bias in

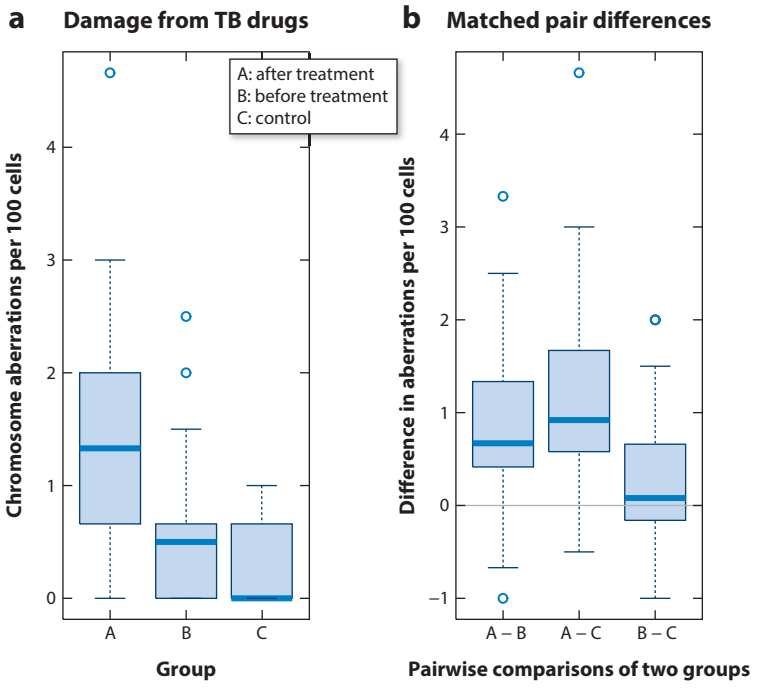

Figure 1

Genetic damage from tuberculosis drugs before treatment (B), after treatment (A), and for age- and gender-matched untreated controls (C). (*a*) Boxplots describing the frequency of chromosome aberrations excluding gaps per 100 cells. (*b*) Boxplots describing matched pair differences between groups: A − B = after − before (change from baseline), A − C = after − control (the difference between the treated and control groups), and B − C = before − control (a check for comparability before treatment).

treatment assignment would need to permit the assignment probability to move away from 1/2 to range over the much longer interval of [0.125, 0.875], and a bias of this magnitude would suffice to explain away the association between heavy smoking and lung cancer in Hammond's (1964) study (see Rosenbaum 2002, section 4.3.2).

In summary, the simplest quasi-experimental device succeeded in demonstrating first that unmeasured biases were indeed present and were not trivially small in magnitude and second that only vastly larger biases could explain away the much larger difference in chromosome aberrations seen after the start of treatment with tuberculosis drugs. In short, we can see unmeasured biases in available data, and we can see that these biases are not trivially small and could be mistaken for an effect. We can also see, however, that the observable biases are not large enough to explain the ostensible effect of tuberculosis drugs, although inevitably very large biases could conceivably explain this effect. In brief, we learn several unambiguous facts about unmeasured biases in this study by collecting a little more data beyond treatment–control pairs and performing a little more analysis beyond a test that assumes unmeasured biases are absent. The triangle design leaves open many questions, but it does answer a few that would otherwise remain unanswered.

The final paragraph of this subsection, which may be skipped, gives some of the technical details about the sensitivity analysis described above. The sensitivity analysis again used the Huber–Maritz M-statistic, as above, but it allowed treatment assignment probabilities within pairs to range over the interval $[1/(1+\Gamma), \Gamma/(1+\Gamma)]$, so one person in a pair might be Γ times more likely than

the other to receive treatment because of a bias in treatment assignment (see Rosenbaum 2007). What is the smallest Γ that could produce a p-value above 0.05? The comparison before treatment yields a maximum p-value of 0.052 at $\Gamma = 1.2$ and a p-value of 0.120 at $\Gamma = 1.5$, whereas the comparison after treatment gives a maximum p-value of 0.053 at $\Gamma = 7$. Using McNemar's test, Hammond's (1964) study of smoking and lung cancer yields a maximum p-value of 0.10 at $\Gamma = 6$ (see Rosenbaum 2002, section 4.3.2). The calculation for chromosome aberrations used the `senmv` function in the `sensitivitymv` package in `R` and inner trimming (`method="i"`). Theory shows that inner trimming increases design sensitivity for many symmetric unimodal distributions of errors such as the normal, logistic, and double-exponential error distributions (see Rosenbaum 2013a, section 3). This example is examined from a different perspective related to equivalence testing in an article by Rosenbaum (2008, section 3).

2.3. Distinguishing Effects, Trends, and Regression Toward and Away From the Mean

The example in Section 2.2 had a single baseline measure of the outcome. As Campbell and colleagues (Campbell & Stanley 1963, Cook & Campbell 1979, Shadish et al. 2002) often observed, a single baseline measure of the outcome suffices to recognize a pretreatment difference in level of outcome between treated and control subjects, but more ambiguity may be removed by using a sequence of baseline measures of the outcome. In the example in Section 2.2, two or more baseline measures of chromosome aberrations in the treated group at a sequence of times prior to treatment could distinguish a linear trend in bias over time from a treatment effect that is absent prior to treatment.

More generally, a temporal sequence of multiple baseline measures can also help in recognizing regression toward or away from the mean. Regression toward the mean can occur when the treatment tends to be given to individuals with high levels of a given outcome prior to treatment. In such cases, we might expect to see a decline from baseline in the treated group in the absence of a treatment effect. Conversely, the so-called horse-race effect (Peto 1981) can occur when the treatment tends to be given to individuals whose outcome levels are rising most quickly prior to treatment. In those cases, we might expect to see a greater posttreatment rise from baseline in the treated group in the absence of a treatment effect. The horse-race effect is so named because of the phenomenon in which the horse that is the frontrunner at the middle of the race may well be even further ahead near the end of the race simply because it is running fastest.

3. EVIDENCE FACTORS

3.1. What Are Evidence Factors?

Section 1.6 argued that a good replication is a new test, statistically independent of previous tests, of the same treatment effect in a context that is likely to have different selection biases. Three elements are in play here: statistical independence or near independence of the new test, the same treatment, and biases that are expected to be different from those seen previously. Evidence factors produce these three elements within a single observational study (Rosenbaum 2010b).

Evidence factors should be distinguished from the common, perhaps unfortunate, practice of analyzing the same data several times using slightly different methods. Similar to evidence factors, repeated analyses concern the same treatment. Unlike evidence factors, however, repeated analyses are often both highly dependent and affected by the same unmeasured biases. Consider the case of matched treatment–control pairs, in which the matching has controlled the observed covariates. A Wilcoxon signed rank test and a paired t-test applied to such pairs do not constitute

two evidence factors: These tests are highly dependent, so the information they provide is highly redundant, and the validity of both requires that matching for observed covariates has effectively produced a randomized paired experiment, which is both a key assumption and a doubtful one. In an observational study, the mostly likely reason that the Wilcoxon test will falsely reject a true null hypothesis of no effect—namely unmeasured biases in treatment assignment—is also the most likely reason that the paired t-test will falsely reject a true null hypothesis. Thus, one learns little by doing both tests, and one risks a false sense of security by seeing similar results from two analyses. Repeated analyses of the same matched pairs using different statistics do have a role in sensitivity analysis, namely, in appraising the magnitude of unmeasured bias needed to explain away the ostensible treatment effect. In this case, however, the analysis must explicitly adjust for multiple testing in a highly dependent context (Rosenbaum 2012a,b), and the repeated analyses constitute one test, not independent replicates subject to different biases.

In his review of observational studies, Cochran (1965, p. 236) wrote the following: "The planner of an observational study should always ask himself the question, 'How would the study be conducted if it were possible to do it by controlled experimentation?'." Randomized experiments exist that involve more than one randomization [see Brien & Bailey (2006) and the references therein]. For example, one might first randomly assign treatment doses to matched pairs, then randomly pick one person in each pair to receive the treatment at the assigned dose and the other (the control) to receive zero dose (e.g., **Figure 2**, Design 1). Alternatively, one might randomly assign pairs of doses to matched pairs, then randomly pick one person from each pair and encourage that person to accept the higher dose. As a third alternative, N subjects can be randomly split into two groups of size $N/3$ and $2N/3$, the smaller of which is given the standard medical treatment. Then, the larger group is split again at random into two groups of size $N/3$ that are encouraged to accept either the standard or new medical treatments. The reader is referred to Zelen (1979) and subsequent literature for clinical trial designs with this sort of randomization and various additions to investigate the consequences of refusal to participate in a randomized trial. Design 2 of **Figure 2** shows a similar matched randomized design. Following Cochran's (1965) advice, we may consider analogous observational studies. If a given experiment has assigned treatments via two randomizations, then in an analogous observational study either or both randomizations

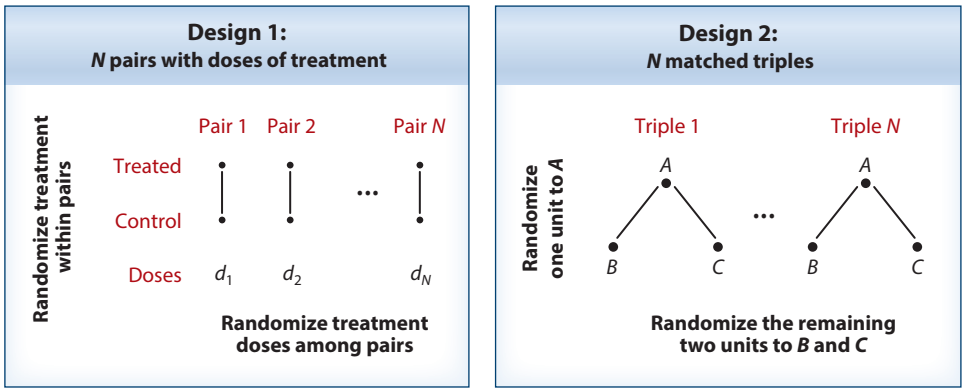

Figure 2

Two simple designs, each employing two randomizations. In analogous observational studies, either randomization could be biased. The two randomizations may be affected by different potential biases, however, and they produce two evidence factors.

might be biased by nonrandom assignment. Depending on the context, the selection biases that operate in these two assignments might be different. In such a context, seeing similar or different results in two analyses that use the two assignments differently may shed light on the kinds of selection biases that would need to be present to explain away the results as something other than an effect of the treatment.

Because evidence factors provide independent p-values with respect to testing the null hypothesis of no treatment effect, these independent p-values may be combined using, for example, Fisher's product of independent p-values or the truncated product proposed by Zaykin et al. (2002). This is true for both randomization tests and sensitivity analyses, the latter of which may allow for different magnitudes and types of bias affecting the two factors. For instance, one might conclude that the null hypothesis of no treatment effect would be rejected no matter how large the biases affecting either factor are, provided that the biases affecting the complementary factor are not very large. Hsu et al. (2013) argue that in sensitivity analyses, the truncated product of independent p-values is a better, more powerful method than Fisher's method because these analyses often produce p-value distributions that are much larger than the uniform distribution.

The simplest type of evidence factor is based on the notion that one group of matched pairs is subject to certain unmeasured biases, whereas a second nonoverlapping group of matched pairs is subject to different unmeasured biases; Section 3.4.1 includes one example of this kind. In this simplest case, outcomes in the pairs are conditionally independent given covariates because the groups are nonoverlapping and both the matching that formed the pairs and the division of the pairs into nonoverlapping groups used covariates but not outcomes. More commonly, each of two evidence factors is computed using all of the data, but these factors are independent under the null hypothesis of no treatment effect. Section 3.3 presents an example based on a study by Silber et al. (2007) with discussion by Cannistra (2007). Section 3.2 provides some technical intuition about how two tests based on the same data can be independent.

3.2. How Can Two Statistics Computed from the Same Data Be Independent?

There is a literature concerned with independent tests computed from the same data; the reader is referred to, for instance, Terpstra (1952), Savage (1957), Dwass (1960), Alam (1974), Mudholkar & McDermott (1989), Marden (1992), and Randles & Hogg (1971). There are also several concepts of nearly independent test statistics [see Randles & Hogg (1971) and Wolfe (1973), among others]. Some of this work draws inspiration or techniques from one of Renyi's ideas, which is attractively presented by Resnick (1999, section 4.3.1). Although the present review is not the place to discuss in detail either the technical issues or their integration with sensitivity analyses in observational studies (Rosenbaum 2010b, 2011), it is useful to exhibit the simplest nontrivial case to show how the technical issues develop.

Suppose that we have three individuals, one of whom will be picked at random for group A, after which one of the two remaining will be picked at random for group B, and the last will end up in group C (**Figure 2**, Design 2). The null hypothesis of no effect is true, so the randomization labels people as group members but does nothing to them, and each of the three individuals will exhibit the same three (conveniently untied) responses, say responses y_1, y_2, and y_3 for individuals 1, 2, and 3, no matter how the randomization comes out. Define rk_A to be the rank, 1, 2, or 3, of the response of the individual randomly assigned to group A. Taking out the individual assigned to group A, define rk_B to be the rank, 1 or 2, of the individual assigned to group B among the two individuals not assigned to group A. For instance, if $y_1 < y_3 < y_2$ and individual 3 is assigned to group A, then $\mathsf{rk}_A = 2$. Then, if individual 2 is subsequently assigned to group B, $\mathsf{rk}_B = 2$. A very special case of Renyi's theorem (Resnick 1999, section 4.3.1) says that these sequential

ranks, rk_A and rk_B, are independent. Terpstra (1952), Dwass (1960), and Alam (1974) build independent hypothesis tests from considerations of this sort, and Marden (1992) extends this reasoning to more complex contrasts among groups. Savage (1957) and Randles & Hogg (1971) take a different approach, for example, exploiting the independence of ranks and order statistics.

Renyi's theorem uses sequential ranks. If the fixed observations $y_1 < y_3 < y_2$ were randomized to groups so that y_A is the observation for the individual randomly picked for group A, and so on, then strict independence is lost. That is, randomization alone does not make y_A independent of y_B; for instance, at most one of y_A and y_B equals y_3. Randomization does, however, make the sign of $y_A - (y_B + y_C)/2$ uncorrelated with the sign of $y_B - y_C$. With some care for technical details, the comparison of group A to the union of groups B and C is nearly independent of the comparison of groups B and C, in the sense that certain kinds of p-values are stochastically larger than the uniform distribution on the unit square and may validly be combined despite dependence using Fisher's product of p-values. The reader is referred to Rosenbaum (2011) for discussion of the omitted technical details. Renyi's theorem is the inspiration for certain types of independent tests, but it is not necessary to restrict attention to ranks to get the needed sorts of independence, at least approximately. Wolfe (1973) provides an elegant discussion of related considerations.

In some observational studies, the process that selects people for group A is very different from the process that divides those remaining into groups B and C (**Figure 2**, Design 2). For instance, consider the case in which the people in group A were not offered the treatment. The people in groups B and C were both offered the treatment, but those in group B declined it, and those in group C accepted it. The B-versus-C comparison is affected by self-selection biases, namely, the preferences of the individuals themselves. The comparison of group A versus the union of B and C is not affected by self-selection biases because self-selection did not enter into the formation of group A. However, the decision by someone else to not offer the treatment to people in group A may be biased by a different process. In an alternative example, the people in group A did not apply for treatment, whereas the people in groups B and C did apply, but the people in group B were turned down. In this context, self-selection operates at the first but not the second stage. The two comparisons may both be biased, but the biases are likely to be different, so with care the two comparisons may be structured to be nearly independent under the null hypothesis of no treatment effect. Design 2 in **Figure 2** is the simplest version of the evidence-factor design discussed in Rosenbaum (2011), and the example in that paper is included in the `sensitivitymv` package in `R`. A related, though structurally different, example along these lines is discussed briefly in Section 3.4.2.

3.3. Example of Evidence Factors: Intensity of Chemotherapy for Ovarian Cancer

3.3.1. Background: medical and gynecological oncologists. Does more intensive use of chemotherapy benefit women with ovarian cancer? Or is such use harmful, producing additional chemotherapy-related toxicity without benefit for survival? In this context, an increase in toxicity might be judged to be a more than acceptable trade-off if survival increased, but it would be disappointing if the increase in toxicity conferred no compensatory benefits. In the absence of a randomized trial, there is an obvious problem with comparing patients who happened to receive more or less chemotherapy: A patient might well receive more chemotherapy because her cancer required it, or she might receive less because she was unable to tolerate the resulting toxicity. A direct comparison of patients who received more or less chemotherapy could thus be strongly biased by the very reasons individual patients received different amounts of treatment. Silber et al. (2007) avoided this problem by considering two types of physicians who provide chemotherapy

for ovarian cancer: medical oncologists (MOs) and gynecological oncologists (GOs). The MOs were initially trained as oncologists and provided chemotherapy for cancers of all kinds, whereas the GOs were initially trained as gynecologists and received additional training in gynecological oncology. The GOs often perform surgery for ovarian cancer and may also provide chemotherapy. Silber et al. (2007) guessed correctly that MOs would use chemotherapy more intensively than GOs, both as initial treatment and in subsequent years if the cancer recurred. There is no obvious or certain reason to seek chemotherapy from an MO versus a GO, and many women will be unaware of the distinction. Silber et al. (2007) examined data from all 344 Medicare patients with ovarian cancer at sites included in the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute from 1991 to 2001 who received chemotherapy from a GO and matched each one to a patient with a similar diagnosis who received chemotherapy from an MO, creating 344 matched pair MO-minus-GO differences. Their matching controlled for 36 pretreatment covariates, including cancer stage; tumor grade; surgeon qualifications; other health problems such diabetes, chronic obstructive pulmonary disease, hypertension, and congestive heart failure; year of diagnosis; age; race; and SEER site. They compared three outcomes: survival, weeks with chemotherapy administration, and weeks with chemotherapy-related toxicity. The reader is referred to table 1 in Silber et al. (2007) for precise definitions of chemotherapy and toxicity, as well as their table 4 for detailed comparisons of these outcomes, including confidence intervals. Kaplan–Meier survival curves for patients of MOs and GOs were virtually identical, not significantly different, and crossing repeatedly; median survival was 2.98 years in the MO group and 3.04 years in the GO group, and five-year survival rates were 34.2% in the MO group and 35.1% in the GO group. These similarities were true despite more intensive use of chemotherapy in the MO group, both in the first year after diagnosis (year 1, indicative of initial treatment) and in years 2 through 5 (presumably treatment of recurrence). On average, in the first five years after diagnosis, MO patients received 16.5 weeks of chemotherapy, whereas GO patients received 12.1 weeks, where $16.5/12.1 = 1.36$. In the absence of longer survival in the group receiving more chemotherapy, it is of interest to ask about differences in chemotherapy-related toxicity. On average, in the first five years after diagnosis, MO patients experienced 16.2 weeks with chemotherapy-related toxicity, whereas GO patients experienced 8.9 weeks. On the surface, then, it appears that MO patients received more weeks of chemotherapy with more toxicity but without benefit for survival.

3.3.2. Two evidence factors affected by different unmeasured biases. This is an observational study, not a randomized experiment, so it may be reasonable to have doubts about whether the surface appearance described above actually reflects effects caused by the differing doses of chemotherapy used by MOs and GOs. In particular, Cannistra (2007) (an MO) raised several issues to consider. Despite being matched for cancer stage, Cannistra (2007) wondered whether the MOs were treating patients with more residual cancer (cancer not removed by surgery). He also wondered whether MOs might give smaller doses in more frequent administrations, causing them to be recorded as having given more weeks of chemotherapy when, in fact, the chemical doses were the same. He also wondered whether MOs might be more diligent in recording toxicity, resulting in their patients seeming to have more weeks with toxicity when actual toxicity was the same. In other words, Cannistra (2007) wondered whether MOs had produced equal survival in sicker patients using the same chemotherapy intensity with equivalent toxicity despite the appearance of equally sick patients receiving more chemotherapy with greater toxicity. Blank & Curtin (2007) (two GOs) expressed some skepticism about the issues raised by Cannistra (2007). Obviously, the observed data cannot entirely put to rest concerns about issues that were not measured, but perhaps these data can be of some assistance in doing so.

This observational study has matched pairs, MO versus GO, and in each pair there are two doses of chemotherapy, the dose given by the MO and that given by the GO. The MO-minus-GO dose difference in weeks of chemotherapy tended to be positive, but it varied from pair to pair, and in some pairs the GO gave more weeks of chemotherapy. Both patients in each pair had the same clinical stage and tumor grade. The structure of this study has a limited resemblance to that of the randomized experiment mentioned in Section 3.1 and depicted in **Figure 2**, Design 1, with random assignment of dose differences to pairs and random assignments to MO or GO within pairs. However, the resemblance is limited, in part because randomization was not used at either step. In addition, both MOs and GOs gave nonzero doses of chemotherapy, although MOs often gave more, so the role played by a dose in **Figure 2**, Design 1 is played here by the MO-minus-GO difference in doses of chemotherapy. This difference is typically but not always positive. If both steps had been randomized, there would have been two randomization tests of no effect on toxicity, one looking at MO-minus-GO pair differences in toxicity, and the other correlating the pair difference in chemotherapy with the pair difference in toxicity. Moreover, these tests would have been independent if they had been done using appropriate rank tests (Rosenbaum 2010b, section 5.1), and they would have been nearly independent using some other tests (Rosenbaum 2011). In the observational study, the process that led one woman to treatment by an MO and another to treatment by a GO might be biased by aspects of the way in which patients are referred to particular physicians. The process that led the MO and GO to use different levels of chemotherapy for two paired patients with the same clinical stage and tumor grade might be biased by considerations that these physicians can see in their patients that we cannot see in the SEER–Medicare data. Nonetheless, whatever biases are present in referral patterns do not seem to be obviously connected to the biases in unrecorded patient characteristics, so it seems worthwhile to examine both comparisons.

Figure 3 plots the MO-minus-GO pair difference in weeks with toxicity against the MO-minus-GO pair difference in weeks with chemotherapy. The first comparison or factor refers to the

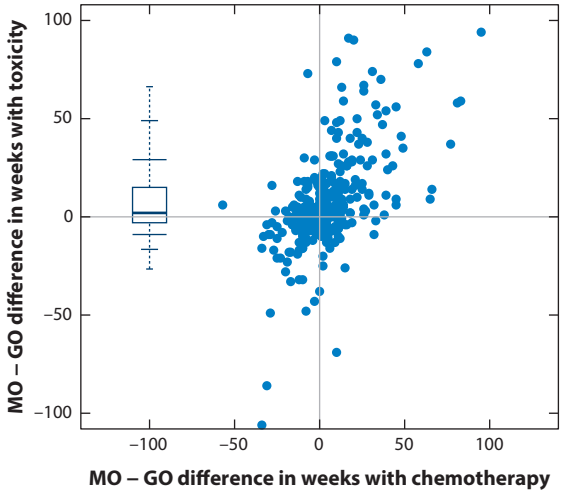

Figure 3

Toxicity and chemotherapy in years 1 to 5. MO − GO matched pair differences in weeks with chemotherapy-related toxicity plotted against matched pair differences in weeks with chemotherapy for 344 pairs of patients with ovarian cancer. The boxplot shows the usual median and quartiles, as well as three more of Tukey's letter values, specifically the 1/8th, 1/16th, and 1/32nd quantiles.

Table 1 Relationship between the MO – GO differences in weeks of chemotherapy and weeks of chemotherapy-related toxicity[a]

	Which patient had more toxicity?		
Which patient had more chemotherapy?	**GO**	**MO**	**Total**
GO	26	6	32
MO	6	71	77
Total	32	77	109

[a]Data represent the 109 pairs in which the absolute difference in chemotherapy was at least equal to the median absolute difference of 9 weeks, and the absolute difference in toxicity was at least equal to the median absolute difference of 8 weeks. The sample odds ratio in this table is 47.8.

boxplot and simply notes that in ostensibly similar pairs of patients, the MO-minus-GO difference in toxicity tends to be positive ($p = 2.6 \times 10^{-16}$ from Wilcoxon's signed rank test). The second comparison or factor notes that the difference in toxicity tends to increase with the difference in chemotherapy (Kendall's correlation $= 0.39$; $p = 2.2 \times 10^{-16}$). In the imagined twice-randomized experiment, the signed rank test and Kendall's test would be independent if the null hypothesis were true, so they would provide separate pieces of information. In many pairs, the MO and GO gave similar weeks of chemotherapy and had similar weeks of toxicity (note the dense mass at the center of **Figure 3**), perhaps because both MO and GO were following the guidance provided by clinical trials. However, in the 109 pairs in which there was a substantial difference in both chemotherapy and toxicity (see **Table 1**), more chemotherapy was strongly associated with more toxicity.

More important than p-values under the implausible assumption of random assignment is the way the two factors relate to unmeasured biases. Cannistra (2007) suggested that the tilt toward more toxicity among MO patients may reflect a greater diligence of MOs in recording toxicity. That is, the first factor, the boxplot, may reflect a particular bias. Notably, in the second factor in **Figure 3**, the scatterplot, the greater diligence of MOs at recording toxicity is most noticeable in pairs in which the MO gave more weeks of chemotherapy than did the GO, whereas in pairs in which the GO gave more chemotherapy, the greater diligence of the MO is lost, and the GO suddenly appears to acquire diligence in recording toxicity. Perhaps more plausible, in light of **Figure 3**, is that differing doses of chemotherapy, not diligent paperwork, are producing the association between weeks with chemotherapy and weeks with chemotherapy-related toxicity.

3.3.3. Sensitivity analyses for the two factors. These comparisons also have a quantitative side. For an unobserved covariate to explain the first factor in terms of a departure from random assignment within pairs—i.e., if the unobserved covariate is to explain the greater frequency of toxicity among MO patients—it would have to alter the fair treatment assignment probabilities of 1/2 to at least 1/3 to 2/3 or 2-to-1 odds ($\Gamma = 2$). Doing this would require an unobserved covariate that could triple the odds of greater toxicity and cause a fivefold increase in the odds of treatment by an MO. Thus, it would take a moderately large bias to spuriously produce the boxplot in **Figure 3**. The pattern in **Table 1** is even more insensitive to unmeasured bias: Producing **Table 1** would take a shift from random assignment or 1-to-1 odds to at least 15-to-1 odds ($\Gamma = 15$). Moreover, these two factors are two separate pieces of information: Either factor could be affected by infinitely large biases ($\Gamma \to \infty$) without invalidating the other. In effect, this says that treatment assignment could be deterministic between pairs without invalidating the comparison within pairs or deterministic within pairs without invalidating the comparison between pairs.

These evidence factors are separate pieces of information in a stronger sense than independence under random assignment.

3.3.4. Technical details of the sensitivity analysis performed in R. This section, which may be skipped, gives the technical details of the sensitivity analyses described in Section 3.3.3. The analysis of MO-minus-GO pairs applied the `senmv` and `amplify` functions of the `sensitivitymv` package in `R` to the 344 matched pair differences in toxicity using `method="i"` for inner trimming. The `senmv` function implements a sensitivity analysis for Maritz's (1979) M-tests derived from Huber's M-estimates, including the permutational t-test, as discussed in Rosenbaum (2007). Theoretical arguments and simulations in Rosenbaum (2013a, section 3) indicate that inner trimming of matched pairs increases design sensitivity. If an M-statistic with inner trimming (`method="i"` in `senmv`) is applied to the 344 pair differences in toxicity, a maximum p-value of 0.041 could be produced by a bias of magnitude $\Gamma = 2$, which corresponds to treatment assignment probabilities between $1/(1+\Gamma) = 1/3$ and $\Gamma/(1+\Gamma) = 2/3$. A bias of $\Gamma = 2$ is the same as (or amplifies to) an unobserved covariate that either triples the odds of treatment and increases the odds of a positive response difference by fivefold, or else increases the odds of treatment by fivefold and triples the odds of a positive response difference [see Rosenbaum & Silber (2009) and calculate `amplify(2, c(3,4,5))`]. In parallel, for nonrandom assignment to rows of **Table 1** to spuriously produce a p-value less than or equal to 0.05 in the absence of an effect of dose difference on toxicity difference would require a bias greater than $\Gamma = 15$ by the method proposed by Rosenbaum (2002, section 4.4.6) for a 2×2 table.

3.4. Other Examples of Evidence Factors

3.4.1. Comparisons within and between institutions that provide treatment: regional versus general anesthesia for knee surgery. Many treatments are provided to individuals within institutions. Patients receive certain types of surgery in hospitals. Addiction treatment may be provided within prisons. Educational interventions may be provided within schools. An individual may find his way to treatment or control by first ending up at a particular institution, then being selected for treatment (rather than control) within that institution. Each of these two steps may be biased by nonrandom selection, but the biases operating at the two steps are often different. For instance, one ends up in a federal prison rather than a state prison by committing a federal crime, whereas procedures within an institution and individual motivation may affect entry into addiction treatment inside a particular prison. This situation can produce two evidence factors.

Zubizarreta et al. (2012) were interested in whether the use of regional anesthesia had health benefits or harms compared with that of general anesthesia for knee surgery. There is no consensus about this topic, and both forms of anesthesia are widely used for knee surgery. Some hospitals often use regional anesthesia for knee surgery, others rarely use it, and still others use it for many patients but not for many others. Zubizarreta et al. (2012) built two nonoverlapping sets of pairs matched for covariates, in which each pair contained one patient who received regional anesthesia, and the other patient received general anesthesia. In the first group of pairs, the within-hospital pairs, each hospital contributed the same number of patients to the regional and general groups. In the second group of pairs, the usual-practice pairs, each hospital contributed patients only to the type of anesthesia it typically performs, resulting in a comparison between hospitals that usually use regional anesthesia and those that usually use general anesthesia. Neither factor or group of pairs is perfect, but they likely have different problems. The first factor is affected by the decisions of the anesthesiologists within a hospital, such as the decision to use regional anesthesia for one patient and general anesthesia for another. Those individual decisions may introduce bias.

The second factor is affected by unmeasured differences among hospitals that prefer to use either regional or general anesthesia. The two factors or sets of pairs are independent, not redundant, simply because they do not overlap. One hopes to see the same benefits or harms from regional anesthesia whether the comparison is within hospitals or between them, but of course that may not happen.

In this example, the two factors pointed in different directions, suggesting that unmeasured biases may be present. Regional anesthesia was associated with lower mortality, less frequent readmission, and a lower rate of deep vein thrombosis in the within-hospital pairs, all of which are biologically plausible effects. However, there was no convincing evidence of any of these effects in the usual-practice pairs. Within hospitals, regional anesthesia looks superior. Between hospitals that usually use regional anesthesia and hospitals that usually use general anesthesia, there is little sign of benefit from the use of regional anesthesia. A study that focused exclusively on only within-hospital or between-hospital comparisons might have reached a mistakenly firm conclusion, but the two factors together provide reason to refrain from such a conclusion. Similarly, a study that did not distinguish between within- and between-hospital information might mistakenly reach a firm conclusion (who knows which one) without realizing that different sources of variation in anesthesia type produce different associations with clinical outcomes.

3.4.2. Subdivisions of the treated group into more or less intensely exposed groups: widespread hospital closures in Philadelphia. In Philadelphia between 1997 and 2007, a massive reduction in the number of hospitals with obstetrics units occurred. Of the 19 such hospitals in 1997, 12 had closed their obstetrics units by 2007. Nothing similar happened in other US cities. Zhang et al. (2011) asked whether these closures had harmful effects on newborns or their mothers. They built a "control-Philadelphia" longitudinally over the years 1997–2007, pairing all 132,786 newborn children in neighborhoods in Philadelphia to children born in the same years in similar neighborhoods in other US cities, where similarity of neighborhood was judged by census tract data. These matched pairs also controlled for numerous individual characteristics of baby and mother, such as age, education, parity, prenatal care, health insurance, birth weight, and gestational age.

The study created two evidence factors and one test for unmeasured biases. The first evidence factor compared newborn pairs before closures to newborn pairs after closures in Philadelphia and "control-Philadelphia," a difference-in-differences. The second factor focused on the time period over which closures occurred. In that time period, the second factor compared neighborhoods in Philadelphia that had hospital closures to neighborhoods in Philadelphia that lacked them, in comparison with "control-Philadelphia," where closures were sporadic and rare. This factor is a different difference-in-differences. The test for bias repeated the comparison in the second factor but did so in the time period before closures, when the various neighborhoods existed but there were no closures to cause effects, yet another difference-in-differences. The three comparisons were all consistent with elevated rates of serious birth injuries [International Statistical Classification of Disease (ICD-9) code 767.3] during the period of closures in Philadelphia. Here, unlike in Section 3.4.1, the three comparisons could have produced inconsistent evidence or evidence of bias, but they did not. Zhang et al. (2011) further discuss sensitivity analyses concerned with the magnitudes of biases that would need to be present to alter these findings.

The second factor subdivided Philadelphia into more or less affected neighborhoods. In principle, the closures probably affected everyone in Philadelphia, even people remote from closures, because the hospitals that remained open initially experienced overcrowding. However, one expects the effect to be largest in the (relatively disadvantaged) neighborhoods that lost their local obstetrics units.

Another example in which the treated group subdivides into more- or less-affected parts is discussed in Rosenbaum (2011). This second small example is available as the `mtm` data in the `sensitivitymv` package in `R`. The examples in that package reproduce the evidence-factor analysis presented in Rosenbaum (2011), including the combination of p-values from separate factors.

4. DIFFERENTIAL EFFECTS AND THE CONTROL OF GENERIC UNMEASURED BIASES

People in observational studies often receive treatments for reasons that are incompletely recorded in measured covariates, with the consequence that comparisons of treated and untreated individuals may be severely biased. Sometimes, these unmeasured biases promote several treatments in a similar way, a so-called generic unmeasured bias, and there is no particular reason people receive any one treatment rather than another. The differential effect of two treatments is the effect of receiving one treatment in lieu of the other. Obviously, this effect is a comparison of two possibly active treatments, so it does not typically equal the main effect of either of those treatments versus no treatment. Nonetheless, in some contexts, the differential effect is unbiased by unmeasured covariates, whereas the main effects are severely biased. In such contexts, a careful study of both main effects and differential effects may rule out certain types of unmeasured bias—specifically, generic unmeasured bias—as competing explanations of an ostensible treatment effect. The current section describes the topic briefly, informally, and without detail. The reader is referred to Rosenbaum (2006, 2013b) for a formal development and detailed examples of this topic, and Gibbons et al. (2010) provide a related discussion.

Given two treatments, say A and B, four (2×2) combinations of treatments and control exist, as in a 2×2 factorial experiment: (not A, not B), (A, not B), (not A, B), or (A, B). The probability that a person falls into any one of these four cells may be severely biased by unmeasured covariates. The differential comparison is that of two of the four cells, specifically (A, not B) and (not A, B). In some contexts, people receive treatment A rather than nothing for a good reason, or they receive treatment B rather than nothing for a good reason, but given that they get A or B but not both, there may be no reason why they get A rather than B. That is to say, under a variety of familiar choice models described in a different article by Rosenbaum (2006, section 3.3), the probability of (A, not B) rather than (not A, not B) is biased by unmeasured covariates, and the probability of (not A, B) rather than (not A, not B) is biased by unmeasured covariates. Thus, both main effects are biased. However, under these models, the conditional probability of (A, not B) given either (A, not B) or (not A, B) is not affected by unmeasured covariates. In this situation, the main effects are biased, but the differential effect is not.

As an illustration, consider the study by Anthony et al. (2000). There is a theory, perhaps correct, that frequent use of nonsteroidal anti-inflammatory drugs (NSAIDs) reduces the risk of Alzheimer disease. Ibuprofen (e.g., brand name Advil) is one such drug. Presumably, a regular user of ibuprofen is suffering from some sort of chronic pain, perhaps from arthritis, headaches, or back pain, and such a person may differ in various ways from a person who is not a regular user of ibuprofen—that is, the main effect of ibuprofen may be biased by the reasons people take it. Indeed, it has been suggested that a person who is in the early undiagnosed stages of Alzheimer disease may be either less aware of pain or less able to act effectively in response to an awareness of pain, potentially producing a spurious negative association between a subsequent diagnosis of Alzheimer disease and regular use of ibuprofen. Pain relievers that are not NSAIDs exist, however, such as acetaminophen (e.g., brand name Tylenol). A lack of awareness of pain or a limited ability to respond effectively to it might reduce the use of ibuprofen for pain, but it is hard to see why it would shift people away from ibuprofen and toward acetaminophen. Why would undiagnosed

Alzheimer disease lead people to reach for Tylenol rather than Advil? The differential effect compares regular users of ibuprofen (without acetaminophen) to regular users of acetaminophen (without ibuprofen), and individuals in both groups are likely to be similar in certain unmeasured ways: They are likely to be in pain, aware of pain, and able to act on their awareness. If there were an ostensible main effect of ibuprofen versus nothing but no differential effect of ibuprofen versus acetaminophen, then that effect could not easily be attributed to a benefit from NSAIDs because acetaminophen is not an NSAID.

Generic unmeasured biases are only one kind of unmeasured bias. A differential unmeasured bias might exist, one that leads some people to prefer *A* to *B* and others to prefer *B* to *A*. Perhaps some unknown polymorphism leads ibuprofen to be more effective as a pain reliever for some people and acetaminophen to be more effective for others. It is not difficult to conduct sensitivity analyses for differential biases in a study that has removed generic biases using differential comparisons (see Rosenbaum 2006, 2013b).

If differential effects are to be enlightening, care is needed in the choice of treatments, *A* and *B*. Ideally, the second comparison treatment, *B*, is plausibly promoted by the same biases that promote *A*, but the causal theory under study says that *A* should have effects that *B* does not have. In the example discussed in Anthony et al. (2000), this is true of ibuprofen, which is an NSAID, versus acetaminophen, which is not. Aspirin would be a poor choice for *B* in this example; aspirin has anti-inflammatory properties, so the differential effect of ibuprofen and aspirin might be zero because both are effective, not because of unmeasured bias.

Of course, the treatments need not be medical drugs. The two treatments, *A* and *B*, may be two health-promoting behaviors that are expected to have very different effects, such as use of sunscreen and/or dental floss. The investigator in such a situation might be seeking to distinguish a generic concern with good health from the specific effect of a treatment on an outcome likely to effected only by one of the treatments. The two treatments, *A* and *B*, might be two foods, say, spinach and tomatoes, with different nutritional components, and the context might be one in which the investigator is seeking to distinguish between generic benefits of consuming vegetables and a specific effect of, say, lycopene in tomatoes. The analysis in this context might then seek to identify effects specific to *A* through differential comparisons with several treatments, *B*, *C*, . . . , promoted by the same generic biases. For example, tomatoes might be compared with several vegetables not containing lycopene. In each of these cases, the investigator is trying to remove a generic unobserved bias by overcompensating for certain observable quantities, comparing people who are visibly different in terms of *B* in an effort to make them more similar in terms of unobservables that promote both *A* and *B*.

5. COMPUTERIZED CONSTRUCTION OF OPTIMAL QUASI-EXPERIMENTS

5.1. Introduction: Extracting Quasi-Experiments from Administrative Data Streams

When Campbell began writing about quasi-experiments in the 1950s, quasi-experimental devices were built from materials that were ready to hand, as described in Section 2.2. The computer has changed this in two ways. First, observational studies increasingly use computerized data, a by-product of the administrative record-keeping needed for some large public or private program. Examples include data from Medicare in the United States (e.g., Silber et al. 2007, 2013), the Kaiser Permanente health system in California (e.g., Silber et al. 2009), or the national System of Measurement of Quality in Education (SIMCE) in Chile (e.g., Zubizarreta et al. 2014). These

administrative data may contain detailed individual records for hundreds of thousands or even millions of people. Second, the computer may be used to extract a quasi-experiment from an administrative data system. Section 5 reviews recent developments related to the second task, the computerized construction of quasi-experiments.

5.2. Constructing Multiple Control Groups

Having two control groups is of value only if the control groups differ in some consequential way (see Campbell 1969, Rosenbaum 1987, Pomp et al. 2010). If having two control groups were of value merely because there are two rather than one, then any control group could be divided in half at random to produce two control groups, but, of course, this would provide no insight into unmeasured biases. Campbell (1969) observed that although a covariate may be unmeasured, it might be known that two specific control groups differ substantially with respect to it. Finding that two such control groups have similar outcomes thereby provides some evidence that the inability to measure this covariate does not greatly bias comparisons. In **Figure 1**, for instance, the treated and control groups were more similar before treatment started than after, so the mere presence of tuberculosis in the treated group does not fully account for the much higher level of chromosome aberrations after the start of antituberculosis drugs. It is not always easy to find such control groups, but related computerized strategies can help.

In building a matched observational study using a large reservoir of potential controls, one often finds that many potential controls are of limited use because they exhibit limited overlap on measured covariates with the treated group. Traditionally, these potential controls were not used. An alternative is to identify a subset of covariates thought to strongly influence treatment assignment but perhaps not outcomes—so-called seemingly innocuous covariates—and to build a second control group without matching for this subset of covariates. Heller et al. (2010, section 4) provide a formal definition of innocuous covariates. This can be done in one of two ways: permitting overlap in the two control groups (Rosenbaum & Silber 2013) or preventing it by tapered matching (Daniel et al. 2008).

The example in Heller et al. (2010) built upon a paper by Rouse (1995) concerned with the potential effects of two-year versus four-year colleges on total years of college education. Students may enter a two-year or four-year program and continue on beyond the initial commitment of two or four years. Similarly, students who enter a two-year or four-year program may fail to complete the program, resulting in spending less than two or four years in college. Some argue that two-year colleges provide both an initial college education in a more affordable format, perhaps living at home without dorm expenses and/or while working part-time, and a near-term goal of a two-year associate's degree. Able students may then transfer to four-year colleges. Others argue that two-year colleges are without advantage to students, that a more limited goal is simply limiting. Heller et al. (2010) constructed two control groups by tapered matching. The first control group matched students in two-year colleges to students in four-year colleges, controlling for 20 covariates including cognitive test scores. Three, possibly innocuous, indicator variables were difficult to control in this 20-variable match; indeed, they were so difficult to control that only matched pairs could be constructed if there was to be control for all 20 covariates, and matching 2-to-1 was not possible. These three indictors represented four regions of the US; some regions, such as the Midwest, have comparatively few two-year colleges. One might suspect that region acts mostly to make two-year colleges convenient or inconvenient and conceivably otherwise matters little for years of college attained, but one cannot be certain. In this sense, region seems innocuous: Why not compare students in Chicago to students in Houston or Seattle? Perhaps region only seems innocuous, however; perhaps Seattle really is different from Chicago. The second control group

was matched for 17 covariates, the original 20 less the 3 region indicators. Tapered matching (Daniel et al. 2008) prevented overlap and optimally allocated students to control groups. In the first control group, region is the same within matched pairs. In the second control group, region may, and often does, differ. Perhaps the second control group is biased by the failure to control for region. Alternatively, perhaps the second control group is less biased than the first: After all, in the second control group, people more often went to two- or four-year colleges simply because of what was available locally, not as the result of an individual decision. Thus, matched triples were similar on 17 covariates, the treated and first control of each triple came from the same region, and the second control often came from a region with fewer two-year colleges. As it turned out, the two control groups had very similar outcomes, so deciding against a two-year college was not much different from living in a region with few such colleges. In terms of the conclusions, the median difference in years of college attained was one year less for students in two-year colleges, not the initial commitment of two years less. This difference occurred in part because a substantial fraction, perhaps a quarter, of controls in four-year colleges did not complete their full four-year commitments. The important point for the present discussion is that tapered matching used more data than pair matching for 20 covariates and used it specifically to shed a little light on unmeasured biases.

The example in Heller et al. (2010) used the tapered matching method proposed by Daniel et al. (2008), thereby preventing any one control from being used twice, once in each of the two matched control groups. Tapered matching is easy to implement: Two distance matrices, one for each definition of the control group, are stacked to yield twice as many rows, after which an optimal match is constructed, say using the `pairmatch` function of Hansen's (2007) `optmatch` package in `R`. Unlike tapered matching, Rosenbaum & Silber (2013) and Silber et al. (2013) use several matched control groups that share some controls. A device called the exterior match permits the comparison of two control groups that share some controls. Tapered matching optimally avoids overlap of control groups, but the control groups have been unnaturally separated: They form interesting comparisons, but they do not represent naturally occurring populations. For other constructions of multiple control groups, the reader is referred to Lu & Rosenbaum (2004), Stuart & Rubin (2008b), and Lu et al. (2011).

5.3. Constructing Evidence Factors Within and Between Institutions

The study by Zubizarreta et al. (2012) of regional anesthesia discussed in Section 3.4.1 created two evidence factors, two sets of matched pairs: One balanced treated and control patients within 47 hospitals, and the other compared patients at different hospitals that typically use different types of anesthesia. Section 3.4.1 discusses the motivation for these two comparisons.

This study was built using a combinatorial optimization algorithm to first allocate individual patients to the two types of pairs and then pair them to be as similar as possible. Specifically, it used fine balance with optimal subsetting in the first type of pair to ensure that each hospital was represented with exactly the same frequency in the treated and control groups while pairing patients as closely as possible for covariates such as their Acute Physiology and Chronic Health Evaluation (APACHE) scores and American Society of Anesthesiologists (ASA) physical status classifications. Because some hospitals typically use regional anesthesia for knee surgery and others typically use general anesthesia, this first, exactly balanced match left behind a large remnant of unused patients representing the type of anesthesia typically used at each hospital. The remnant was optimally matched for covariates, creating the second factor, a comparison purely between different hospitals. We may then ask the following questions: Does the treatment seem to have the

same effect when applied within hospitals as between hospitals? Or are there signs that something besides the treatment is affecting outcomes?

5.4. Constructing Stronger Instrumental Variables

An instrument or instrumental variable is an ostensibly random push or encouragement to accept a particular treatment in a context in which encouragement alone can affect outcomes only if it does, indeed, alter the treatment received (see Holland 1988, Angrist et al. 1996). The Vietnam War draft lottery was essentially random and pushed some people into military service who would not have entered voluntarily. Many people served without being drafted, however, and others found ways around the draft, so the lottery was merely a randomized push to accept a treatment. The goal in using an instrument is to extract a bit of random treatment assignment from a context in which treatment assignment is typically very biased.

An instrument is said to be strong if it usually decides treatment, and it is said to be weak if it affects treatment decisions only slightly. Weak instruments create inferential problems (see Bound et al. 1995, Imbens & Rosenbaum 2004). More importantly, a study with a weak instrument is invariably sensitive to small biases in the random assignment of instrument levels (Small & Rosenbaum 2008). For these reasons, strong instruments are preferable.

Several recent studies have used combinatorial optimization to construct stronger instruments; for example, the reader is referred to Baiocchi et al. (2010) and Zubizarreta et al. (2013). In particular, Baiocchi et al. (2010) asked whether hospitals in Pennsylvania with advanced neonatal intensive care units (NICUs) reduced the risk of death of premature infants. The focus on Pennsylvania was a matter of convenience, namely, the availability of data; their question is about the effect of advanced NICUs and has no particular interest in Pennsylvania per se. Following a long tradition in health services research, the instrument used was excess travel time from the mother's zip code to a hospital with an advanced NICU, the reasoning being that geography is strongly related to accessibility of certain hospitals but not a major risk factor for infant mortality. A mother in labor does not have the option of traveling for several hours to reach a more capable hospital. In Pennsylvania, most people live in or near major cities such as Philadelphia and Pittsburgh, so they are near numerous hospitals of varied capabilities, and distance is a weak instrument for hospital choice. The advanced NICU at the highly capable medical school of Pennsylvania State University in Hershey, Pennsylvania, however, is in a mid-sized town surrounded by farm country. Elsewhere in Pennsylvania, farm country is often remote from hospitals with advanced NICUs. Using a form of optimal nonbipartite matching (e.g., Lu et al. 2011 and `R` package `nbpMatching`), Baiocchi et al. (2010) rebuilt Pennsylvania into a smaller, more rural state in which geography is a strong instrument for hospital choice. Their analysis revealed that advanced NICUs do appear to reduce mortality among premature infants.

5.5. Constructing Stronger Discontinuity Designs

Some treatment groups have well-defined, abruptly changing standards for entry. There may be a cut point on some score that grants entry into some government-run entitlement program. Adjacent school districts with different policies may have well-defined geographic boundaries. When this is true, treatment may change abruptly at a boundary, but people near the boundary who receive different treatments may be very similar in most other ways. Thistlewaite & Campbell (1960) proposed exploiting this idea as a quasi-experimental design; they compared similar people with different treatments near the abrupt boundary for treatment, and their technique has become widely popular.

Keele et al. (2015) employed optimal matching to strengthen such a design. They studied the effect of ballot initiatives on voter turnout by locating the geographic boundary of the electoral districts in which a ballot initiative is present. In their study design, the boundary is long and heterogeneous, bordering on many other electoral districts, and the potential voters are themselves heterogeneous. Keele et al. (2015) formed optimally matched pairs of potential voters, just inside or just outside a district near the same location on the long boundary, matching similar potential voters. Conceptually, the treatment is identified only near the boundary of eligibility and only among similar individuals facing each other from opposite sides of it (Hahn et al. 2001). The design used by Keele et al. (2015) is an actual structure that closely resembles this underlying concept.

6. SUMMARY

Quasi-experimental devices attempt to detect anticipated patterns of unmeasured biases, often by conducting several comparisons unequally or differently affected by such biases. Are the same ostensible treatment effects seen in several comparisons differently affected by anticipated biases? There is often the realistic hope that quasi-experimental devices will detect large unmeasured biases of anticipated form. A sensitivity analysis may show that, no matter what form they take, small or moderate biases in treatment assignment cannot explain the observed association between treatment and outcome. Taken together in observational studies, quasi-experimental devices and sensitivity analyses can reduce ambiguity about the effects caused by treatments. The discussion here has emphasized three recent developments: evidence factors, differential effects to remove generic unmeasured biases, and computerized constructions of quasi-experiments.

SUMMARY POINTS

1. Quasi-experimental devices enlarge the set of considered associations in an effort to disambiguate the association between treatment and outcome.
2. Quasi-experimental devices may detect or rule out large unmeasured biases of anticipated form.
3. Quasi-experimental devices complement sensitivity analyses that speak to small or moderate biases of any form.
4. Evidence factors produce, within a single study, two statistically independent tests of no treatment that are likely to be affected by different unmeasured biases.
5. The study of differential effects can remove generic unmeasured biases that promote several treatments in a similar way.
6. Combined with large computerized administrative record systems, combinatorial optimization algorithms can be used to construct quasi-experiments with desired properties.

FUTURE ISSUES

1. Improvements in computerized administrative record systems provide expanding opportunities for constructing quasi-experiments.
2. To date, the construction of quasi-experiments has emphasized network optimization techniques, but the growing literature on approximation algorithms offers the potential for using these techniques to construct quasi-experiments.

3. If matched pairs successfully control many covariates, there may be no risk to confidentiality in publicly disclosing (*a*) summary measures of matched-pair covariate balance for many covariates and (*b*) individual-level or microdata for just a few outcomes in matched pairs.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I would like to thank Larry Brown for helpful comments. This work was supported in part by grant SES-1260782 from the US National Science Foundation.

LITERATURE CITED

Alam K. 1974. Some nonparametric tests of randomness. *J. Am. Stat. Assoc.* 69:738–39

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91:444–55; discussion pp. 455–68

Anthony JC, Breitner JC, Zandi PP, Meyer MR, Jurasova I, et al. 2000. Reduced prevalence of AD in users of NSAIDs and H2 receptor antagonists. *Neurology* 54:2066–71

Baiocchi M, Small DS, Lorch S, Rosenbaum PR. 2010. Building a stronger instrument in an observational study of perinatal care for premature infants. *J. Am. Stat. Assoc.* 105:1285–96

Basu AP. 1983. Identifiability. In *Encyclopedia of Statistical Sciences*, Vol. 4, p. 2. New York: Wiley

Blank SV, Curtin JP. 2007. More than a name. *J. Clin. Oncol.* 25:3551

Bound J, Jaeger DA, Baker RM. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Am. Stat. Assoc.* 90:443–50

Brien CJ, Bailey RA. 2006. Multiple randomizations. *J. R. Stat. Soc. A* 68:571–99; discussion pp. 599–609

Campbell DT. 1957. Factors relevant to the validity of experiments in social settings. *Psychol. Bull.* 54:297–312. Reprinted in Campbell 1988

Campbell DT. 1969. Prospective: artifact and control. In *Artifact in Behavioral Research*, ed. R Rosenthal, R Rosnow, pp. 351–82. New York: Academic Press. Reprinted in Campbell 1988

Campbell DT. 1988. *Methodology and Epistemology for Social Science: Selected Papers*. Chicago: Univ. Chicago Press

Campbell DT, Boruch RF. 1975. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs*, ed. CA Bennett, AA Lumsdaine, pp. 195–296. New York: Academic

Campbell DT, Stanley JC. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally

Cannistra SA. 2007. Gynecologic oncology or medical oncology: What's in a name? *J. Clin. Oncol.* 25:1157–59

Cochran WG. 1965. The planning of observational studies of human populations. *J. R. Stat. Soc. A* 128:234–65

Cook TD, Campbell DT. 1979. *Quasi-Experimentation*. Boston: Houghton Mifflin

Cornfield J, Haenszel W, Hammond E, Lilienfeld A, Shimkin M, Wynder E. 1959. Smoking and lung cancer. *J. Nat. Cancer Inst.* 22:173–203

Daniel SR, Armstrong K, Silber JH, Rosenbaum PR. 2008. An algorithm for optimal tapered matching, with application to disparities in survival. *J. Comp. Graph. Stat.* 174:914–24

Diprete TA, Gangl M. 2004. Assessing bias in the estimation of causal effects. *Sociol. Methodol.* 34:271–310

Dwass M. 1960. Some *k*-sample rank order tests. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, ed. I Olkin, pp. 198–202. Stanford, CA: Stanford Univ. Press
Fisher RA. 1935. *The Design of Experiments*. Edinburgh, UK: Oliver & Boyd
Gastwirth JL. 1992. Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics* 33:19–34
Gibbons RD, Amatya AK, Brown CH, Hur K, Marcus SM, et al. 2010. Post-approval drug safety surveillance. *Annu. Rev. Public Health* 31:419–37
Grodstein F, Stampfer MJ, Manson JE, Colditz GA, Willet WC, et al. 1996. Postmenopausal estrogen and progestin use and the risk of cardiovascular disease. *New Engl. J. Med.* 335:453–61
Hammond EC. 1964. Smoking in relation to mortality and morbidity. *J. Nat. Cancer Inst.* 32:1161–88
Hahn J, Todd P, Van der Klaauw W. 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69:201–9

R package `optmatch`.

Hansen BB. 2007. Optmatch: flexible, optimal matching for observational studies. *R News* 7:18–24

Hedayat AS, Sloane NJA, Stufken J. 1999. *Orthogonal Arrays: Theory and Applications*. New York: Springer
Heller R, Rosenbaum PR, Small DS. 2010. Using the cross-match test to appraise covariate balance in matched pairs. *Am. Stat.* 64:299–309
Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, et al. 2008. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 19:766–79
Holland PWH. 1988. Causal inference, path analysis, and recursive structural equations models. *Sociol. Methodol.* 18:449–84
Hosman CA, Hansen BB, Holland PWH. 2010. The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann. Appl. Stat.* 4:849–70
Hsu JY, Small DS, Rosenbaum PR. 2013. Effect modification and design sensitivity in observational studies. *J. Am. Stat. Assoc.* 108:135–48
Imbens GW. 2003. Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.* 93:126–32
Imbens GW, Rosenbaum PR. 2004. Robust, accurate confidence intervals with a weak instrument. *J. R. Stat. Soc. A* 168:109–26
Imbens GW, Wooldridge JM. 2009. Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* 47:5–86
Keele L, Titiunik R, Zubizarreta JR. 2015. Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *J. R. Stat. Soc. A.* 178:223–39
Liu W, Kuramoto SJ, Stuart EA. 2013. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevent. Sci.* 14:570–80

R package `nbpmatching`.

Lu B, Greevy R, Xu X, Beck C. 2011. Optimal nonbipartite matching and its statistical applications. *Am. Stat.* 65:21–30

Lu B, Rosenbaum PR. 2004. Optimal pair matching with two control groups. *J. Comp. Graph. Stat.* 13:422–34
Lund E, Bønaa KH. 1993. Reduced breast cancer mortality among fisherman's wives in Norway. *Cancer Causes Cont.* 4:283–87
Manski CF. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard Univ. Press
Manski CF, Nagin DS. 1998. Bounding disagreements about treatment effects: a case study of sentencing and recidivism. *Sociol. Methodol.* 28:99–137
Marden JI. 1992. Use of nested orthogonal contrasts in analyzing rank data. *J. Am. Stat. Assoc.* 87:307–18
Maritz JS. 1979. A note on exact robust confidence intervals for location. *Biometrika* 66:163–66
Masjedi MR, Heidary A, Mohammadi F, Velayati AA, Dokouhaki P. 2000. Chromosome aberrations and micronuclei in lymphocytes of patients before and after exposure to anti-tuberculosis drugs. *Mutagenesis* 15:489–94
Meyer BD. 1995. Natural and quasi-experiments in economics. *J. Bus. Econ. Stat.* 13:151–61
Mudholkar GS, McDermott MP. 1989. A class of tests for the equality of ordered means. *Biometrika* 76:161–68
Nagin DS, Weisburd D. 2013. Evidence and public policy: the example of evaluation of research in policing. *Criminol. Public Policy* 12:651–79

Neel J. 2002. The marketing of menopause: historically, hormone therapy heavy on promotion, light on science. Washington, DC: Nat. Public Radio (8 August 2002)
Peto R. 1981. The horse-racing effect. *Lancet* 318:467–68
Pomp ER, Van Stralen KJ, Le Cessie S, Vandenbroucke JP, Rosendaal FR, Doggen CJM. 2010. Experience with multiple control groups in a large population-based case–control study on genetic and environmental risk factors. *Eur. J. Epidemiol.* 25:459–66
Randles RH, Hogg RV. 1971. Certain uncorrelated and independent rank statistics. *J. Am. Stat. Assoc.* 66:569–74
Resnick SI. 1999. *A Probability Path.* Berlin: Birkhauser
Robins JM, Rotnitzky A, Scharfstein D. 1999. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology*, ed. E Halloran, D Berry, pp. 1–94. New York: Springer
Rosenbaum PR. 1987. The role of a second control group in an observational study. *Stat. Sci.* 2:292–316
Rosenbaum PR. 1988. Sensitivity analysis for matching with multiple controls. *Biometrika* 75:577–81
Rosenbaum PR. 1991. Discussing hidden bias in observational studies. *Ann. Intern. Med.* 115:901–5
Rosenbaum PR. 2001a. Replicating effects and biases. *Am. Stat.* 55:223–27
Rosenbaum PR. 2001b. Stability in the absence of treatment. *J. Am. Stat. Assoc.* 96:210–19
Rosenbaum PR. 2002. *Observational Studies*. New York: Springer. 2nd ed.
Rosenbaum PR. 2006. Differential effects and generic biases in observational studies. *Biometrika* 93:573–86
Rosenbaum PR. 2007. Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics* 63:456–64
Rosenbaum PR. 2008. Testing hypotheses in order. *Biometrika* 95:248–52
Rosenbaum PR. 2010a. *Design of Observational Studies*. New York: Springer
Rosenbaum PR. 2010b. Evidence factors in observational studies. *Biometrika* 97:333–45
Rosenbaum PR. 2011. Some approximate evidence factors in observational studies. *J. Am. Stat. Assoc.* 106:285–95
Rosenbaum PR. 2012a. An exact adaptive test with superior design sensitivity in an observational study of treatments for ovarian cancer. *Ann. Appl. Stat.* 6:83–105
Rosenbaum PR. 2012b. Testing one hypothesis twice in observational studies. *Biometrika* 99:763–74
Rosenbaum PR. 2013a. Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics* 69:118–27
Rosenbaum PR. 2013b. Using differential comparisons in observational studies. *Chance* 26:18–25
Rosenbaum PR, Rubin DB. 1983. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Stat. Soc. B* 45:212–18
Rosenbaum PR, Silber JH. 2009. Amplification of sensitivity analysis in observational studies. *J. Am. Stat. Assoc.* 104:1398–405
Rosenbaum PR, Silber JH. 2013. Using the exterior match to compare two entwined matched control groups. *Am. Stat.* 67:67–75
Rouse CE. 1995. Democratization or diversion? The effect of community colleges on educational attainment. *J. Bus. Econ. Stat.* 13:217–24
Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66:688–701
Rutter M, Acad. Med. Sci. Work. Group. 2007. *Identifying the Environmental Causes of Disease: How Should We Decide What to Believe and When to Take Action?* London, UK: Acad. Med. Sci. **http://www.acmedsci.ac.uk/policy/policy/identifying-the-environmental-causes-of-disease/**
Savage IR. 1957. On the independence of tests of randomness and other hypotheses. *J. Am. Stat. Assoc.* 52:53–57
Shadish WR, Cook TD, Campbell DT. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin
Silber JH, Lorch SA, Rosenbaum PR, Medoff-Cooper B, Bakewell-Sachs S, et al. 2009. Time to send the preemie home? Additional maturity at discharge and subsequent health care costs and outcomes. *Health Serv. Res.* 44:444–63
Silber JH, Rosenbaum PR, Clark AS, Giantonio BJ, Ross RN, et al. 2013. Characteristics associated with differences in survival among black and white women with breast cancer. *J. Am. Med. Assoc.* 310:389–97

R packages `sensitivitymv` and `sensitivitymw`.

`senmv` and `truncatedP` functions and `mtm` data in the `R` package `sensitivitymv`.

`adaptive.noether`. `brown` function in the `R` package `SensitivityCaseControl`.

`R` packages `sensitivitymv` and `sensitivitymw`.

`amplify` function in the `R` package `sensitivitymv`.

Silber JH, Rosenbaum PR, Polsky D, Ross RN, Even-Shoshan O, et al. 2007. Does ovarian cancer treatment and survival differ by the specialty providing chemotherapy? *J. Clin. Oncol.* 25:1169–75

Small DS, Rosenbaum PR. 2008. War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *J. Am. Stat. Assoc.* 103:924–33

Stuart EA, Hanna DB. 2013. Should epidemiologists be more sensitive to design sensitivity? *Epidemiology* 24:88–89

Stuart EA, Rubin DB. 2008a. Best practices in quasi-experimental designs. In *Best Practices in Quantitative Methods*, ed. J Osborne, pp. 155–76. Thousand Oaks, CA: Sage

Stuart EA, Rubin DB. 2008b. Matching with multiple control groups with adjustment for group differences. *J. Educ. Behav. Stat.* 33:279–306

Susser M. 1987. Falsification, verification and causal inference in epidemiology: Reconsideration in the light of Sir Karl Popper's philosophy. In *Epidemiology, Health and Society: Selected Papers*, ed. M Susser, pp. 82–93. New York: Oxford Univ. Press

Terpstra TJ. 1952. Asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indag. Math.* 14:327–33

Thistlewaite DL, Campbell DT. 1960. Regression-discontinuity analysis: an alternative to the ex post facto experiment. *J. Educ. Psychol.* 51:309–17

West SG, Duan N, Pequegnat W, Gaist P, Des Jarlais DC, et al. 2008. Alternatives to the randomized controlled trial. *Am. J. Public Health* 98:1359–66

Wolfe DA. 1973. Some general results about uncorrelated statistics. *J. Am. Stat. Assoc.* 68:1013–18

Women's Health Initiative Writing Group. 2002. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *J. Am. Med. Assoc.* 288:321–33

Wu CFJ, Hamada MS. 2011. *Experiments: Planning, Analysis, and Optimization*. Hoboken, NJ: John Wiley & Sons

Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. 2002. Truncated product method for combining *p*-values. *Genet. Epidemiol.* 22:170–85

Zelen M. 1979. A new design for randomized clinical trials. *New Eng. J. Med.* 300:1242–45

Zhang K, Small DS, Lorch S, Srinivas S, Rosenbaum PR. 2011. Using split samples and evidence factors in an observational study of neonatal outcomes. *J. Am. Stat. Assoc.* 106:511–24

Zubizarreta JR, Neuman M, Silber JH, Rosenbaum PR. 2012. Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia. *J. Am. Stat. Assoc.* 107:901–15

Zubizarreta JR, Paredes RD, Rosenbaum PR. 2014. Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Ann. Appl. Stat.* 8:204–31

Zubizarreta JR, Small DS, Goyal NK, Lorch S, Rosenbaum PR. 2013. Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Ann. Appl. Stat.* 7:25–50

`truncatedP` **function in the** `R` **package** `sensitivitymv`.

Annual Review of Statistics and Its Application

Volume 2, 2015

Contents