Observational Study

Paul R. Rosenbaum

Volume 3, pp. 1451–1462

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Observational Study

Observational Studies Defined

In the ideal, the effects caused by treatments are investigated in experiments that randomly assign subjects to treatment or control, thereby ensuring that comparable groups are compared under competing treatments [1, 5, 15, 23]. In such an experiment, comparable groups prior to treatment ensure that differences in outcomes after treatment reflect effects of the treatment (*see* **Clinical Trials and Intervention Studies**). Random assignment uses chance to form comparable groups; it does not use measured characteristics describing the subjects before treatment. As a consequence, random assignment tends to make the groups comparable both in terms of measured characteristics and characteristics that were not or could not be measured. It is the unmeasured characteristics that present the largest difficulties when **randomization** is not used. More precisely, random assignment ensures that the only differences between treated and control groups prior to treatment are due to chance – the flip of a coin in assigning one subject to treatment, another to control – so if a common statistical test rejects the hypothesis that the difference is due to chance, then a treatment effect is demonstrated [18, 22].

Experiments with human subjects are often ethical and feasible when (a) all of the competing treatments under study are either harmless or intended and expected to benefit the recipients, (b) the best treatment is not known, and in light of this, subjects consent to be randomized, and (c) the investigator can control the assignment and delivery of treatments. Experiments cannot ethically be used to study treatments that are harmful or unwanted, and experiments are not practical when subjects refuse to cede control of treatment assignment to the experimenter. When experiments are not ethical or not feasible, the effects of treatments are examined in an observational study. Cochran [12] defined an *observational study* as an empiric comparison of treated and control groups in which:

> the objective is to elucidate cause-and-effect relationships [... in which it] is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover, or to assign subjects at random to different procedures.

When subjects are not assigned to treatment or control at random, when subjects select their own treatments or their environments inflict treatments upon them, differing outcomes may reflect these initial differences rather than effects of the treatments [12, 59]. Pretreatment differences or *selection biases* are of two kinds: those that have been accurately measured, called *overt biases*, and those that have not be measured but are suspected to exist, called *hidden biases*. Removing overt biases and addressing uncertainty about hidden biases are central issues in observational studies. Overt biases are removed by adjustments, such as **matching**, **stratification** or covariance adjustment (*see* **Analysis of Covariance**), which are discussed in the Section titled 'Adjusting for Biases Visible in Observed Covariates'. Hidden biases are addressed partly in the design of an observational study, discussed in the Section titled 'Design of Observational Studies' and 'Elaborate Theories', and partly in the analysis of the study, discussed in the Section titled 'Appraising Sensitivity to Hidden Bias' and 'Elaborate Theories' (*see* **Quasi-experimental Designs**; **Internal Validity**; and **External Validity**).

Examples of Observational Studies

Several examples of observational studies are described. Later sections refer to these examples.

Long-term Psychological Effects of the Death of a Close Relative

In an attempt to estimate the long-term psychological effects of bereavement, Lehman, Wortman, and Williams [30] collected data following the sudden death of a spouse or a child in a car crash. They matched 80 bereaved spouses and parents to 80 controls drawn from 7581 individuals who came to renew a drivers license. Specifically, they matched for gender, age, family income before the crash, education level, number, and ages of children. Contrasting their findings with the views of Bowlby and Freud, they concluded:

> Contrary to what some early writers have suggested about the duration of the major symptoms

> of bereavement ... both spouses and parents in our study showed clear evidence of depression and lack of resolution at the time of the interview, which was 5 to 7 years after the loss occurred.

Effects on Criminal Violence of Laws Limiting Access to Handguns

Do laws that ban purchases of handguns by convicted felons reduce criminal violence? It would be difficult, perhaps impossible, to study this in a randomized experiment, and yet an observational study faces substantial difficulties as well. One could not reasonably estimate the effects of such a law by comparing the rate of criminal violence among convicted felons barred from handgun purchases to the rate among all other individuals permitted to purchase handguns. After all, convicted felons may be more prone to criminal violence and may have greater access to illegally purchased guns than typical purchasers of handguns without felony convictions. For instance, in his ethnographic account of violent criminals, Athens (1997, p. 68) depicts their sporadic violent behavior as consistent with stable patterns of thought and interaction, and writes: "... the self-images of violent criminals are always congruent with their violent criminal actions."

Wright, Wintemute, and Rivara [68] compared two groups of individuals in California: (a) individuals who attempted to purchase a handgun but whose purchase was denied because of a prior felony conviction, and (b) individuals whose purchase was approved because their prior felony arrest had not resulted in a conviction. The comparison looked forward in time from the attempt to purchase a handgun, recording arrest charges for new offenses in the subsequent three years. Presumably, group (b) is a mixture of some individuals who did not commit the felony for which they were arrested and others who did. If this presumption were correct, group (a) would be more similar to group (b) than to typical purchasers of handguns, but substantial biases may remain.

Effects on Children of Occupational Exposures to Lead

Morton, Saah, Silberg, Owens, Roberts, and Saah [39] asked whether children were harmed by lead brought home in the clothes and hair of parents who were exposed to lead at work. They matched 33 children whose parents worked in a battery factory to 33 unexposed control children of the same age and neighborhood, and used Wilcoxon's signed rank test to compare the level of lead found in the children's blood, finding elevated levels of lead in exposed children.

In addition, they compared exposed children whose parents had varied levels of exposure to lead at the factory, finding that parents who had higher exposures on the job in turn had children with more lead in their blood. Finally, they compared exposed children whose parents had varied hygiene upon leaving the factory at the end of the day, finding that poor hygiene of the parent predicted higher levels of lead in the blood of the child.

Design of Observational Studies

Observational studies are sometimes referred to as *natural experiments* [36, 56] or as *quasi-experiments* [61] (*see* **Quasi-experimental Designs**). These differences in terminology reflect certain differences in emphasis, but a shared theme is that the early stages of planning or designing an observational study attempt to reproduce, as nearly as possible, some of the strengths of an experiment [47].

A *treatment* is a program, policy, or intervention which, in principle, may be applied to or withheld from any subject under study. A variable measured prior to treatment is not affected by the treatment and is called a *covariate*. A variable measured after treatment may have been affected by the treatment and is called an *outcome*. An analysis that does not carefully distinguish covariates and outcomes can introduce biases into the analysis where none existed previously [43]. The *effect caused by a treatment* is a comparison of the outcome a subject exhibited under the treatment the subject actually received with the potential but unobserved outcome the subject would have exhibited under the alternative treatment [40, 59]. Causal effects so defined are sometimes said to be *counterfactual* (*see* **Counterfactual Reasoning**), in the specific sense that they contrast what did happen to a subject under one treatment with what would have happened under the other treatment. Causal effects cannot be calculated for individuals, because each individual is observed under treatment or under control, but not both. However, in a randomized experiment, the treated-minus-control difference in mean outcomes is an unbiased and consistent estimate of the average effect of the treatment on the subjects in the experiment.

In planning an observational study, one attempts to identify circumstances in which some or all of the following elements are available [47].

- *Key covariates and outcomes are available for treated and control groups*. The most basic elements of an observational study are treated and control groups, with important covariates measured before treatment, and outcomes measured after treatment. If data are carefully collected over time as events occur, as in a *longitudinal study* (*see* **Longitudinal Data Analysis**), then the temporal order of events is typically clear, and the distinction between covariates and outcomes is clear as well. In contrast, if data are collected from subjects at a single time, as in a *cross-sectional study* (*see* **Cross-sectional Design**) based on a single survey interview, then the distinction between covariates and outcomes depends critically on the subjects' recall, and may not be sharp for some variables; this is a weakness of cross-sectional studies. Age and sex are covariates whenever they are measured, but current recall of past diseases, experiences, moods, habits, and so forth can easily be affected by subsequent events.
- *Haphazard treatment assignment rather than self-selection*. When randomization is not used, treated and control groups are often formed by deliberate choices reflecting either the personal preferences of the subjects themselves or else the view of some provider of treatment that certain subjects would benefit from treatment. Deliberate selection of this sort can lead to substantial biases in observational studies. For instance, Campbell and Boruch [10] discuss the substantial systematic biases in many observational studies of compensatory programs intended to offset some disadvantage, such as the US Head Start Program for preschool children. Campbell and Boruch note that the typical study compares disadvantaged subjects eligible for the program to controls who were not eligible because they were not sufficiently disadvantaged. When randomization is not possible, one should try to identify circumstances in which an ostensibly irrelevant event, rather than deliberate choice, assigned subjects to treatment or control. For instance, in the United States, class sizes in government run schools are largely determined by the degree of wealth in the local region, but in Israel, a rule proposed by Maimonides in the 12th century still requires that a class of 41 must be divided into two separate classes. In Israel, what separates a class of size 40 from classes half as large is the enrollment of one more student. Angrist and Lavy [2] exploited Maimonides rule in their study of the effects of class size on academic achievement in Israel. Similarly, Oreopoulos [41] studies the economic effects of living in a poor neighborhood by exploiting the policy of Toronto's public housing program of assigning people to housing in quite different neighborhoods simply based on their position in a waiting list. Lehman et al. [30], in their study of bereavement in the Section titled 'Long-term Psychological Effects of the Death of a Close Relative', limited the study to car crashes for which the driver was not responsible, on the grounds that car crashes for which the driver was responsible were relatively less haphazard events, perhaps reflecting forms of addiction or psychopathology. Random assignment is a fact, but haphazard assignment is a judgment, perhaps a mistaken one; however, haphazard assignments are preferred to assignments known to be severely biased.
- *Special populations offering reduced self-selection*. Restriction to certain subpopulations may diminish, albeit not eliminate, biases due to self-selection. In their study of the effects of adolescent abortion, Zabin, Hirsch, and Emerson [69] used as controls young women who visited a clinic for a pregnancy test, but whose test result came back negative, thereby ensuring that the controls were also sexually active. The use in the Section titled 'Effects on Criminal Violence of Laws Limiting Access to Handguns' of controls who had felony arrests without convictions may also reduce hidden bias.
- *Biases of known direction*. In some settings, the direction of unobserved biases is quite clear even if their magnitude is not, and in certain special circumstances, a treatment effect that overcomes a bias working against it may yield a relatively unambiguous conclusion. For

instance, in the Section titled ‘Effects on Criminal Violence of Laws Limiting Access to Handguns’, one expects that the group of convicted felons denied handguns contains fewer innocent individuals than does the arrested-but-not-convicted group who were permitted to purchase handguns. Nonetheless, Wright et al. [68] found fewer subsequent arrests for gun and violent offenses among the convicted felons, suggesting that the denial of handguns may have had an effect large enough to overcome a bias working in the opposite direction. Similarly, it is often claimed that payments from disability insurance provided by US Social Security deter recipients from returning to work by providing a financial disincentive. Bound [6] examined this claim by comparing disability recipients to rejected applicants, where the rejection was based on an administrative judgment that the injury or disability was not sufficiently severe. Here, too, the direction of bias seems clear: rejected applicants should be healthier. However, Bound found that even among the rejected applicants, relatively few returned to work, suggesting that even fewer of the recipients would return to work without insurance. Some general theory about studies that exploit biases of known direction is given in Section 6.5 of [49].

- *An abrupt start to intense treatments*. In an experiment, the treated and control conditions are markedly distinct, and these conditions become active at a specific known time. Lehman et al.’s [30] study of the psychological effects of bereavement in the Section titled ‘Long-term Psychological Effects of the Death of a Close Relative’ resembles an experiment in this sense. The study concerned the effects of the sudden loss of a spouse or a child in a car crash. In contrast, the loss of a distant relative or the gradual loss of a parent to chronic disease might possibly have effects that are smaller, more ambiguous, more difficult to discern. In a general discussion of studies of stress and depression, Kessler [28] makes this point clearly:

  > “... a major problem in interpret[ation] ... is that both chronic role-related stresses and the chronic depression by definition have occurred for so long that deciding unambiguously which came first is difficult ... The researcher, however, may focus on stresses that can be assumed to have occurred randomly with respect to other risk factors of depression and to be inescapable, in which case matched comparison can be used to make causal inferences about long-term stress effects. A good example is the matched comparison of the parents of children having cancer, diabetes, or some other serious childhood physical disorder with the parents of healthy children. Disorders of this sort are quite common and occur, in most cases, for reasons that are unrelated to other risk factors for parental psychiatric disorder. The small amount of research shows that these childhood physical disorders have significant psychiatric effects on the family.” (p. 197)

- *Additional structural features in quasi-experiments intended to provide information about hidden biases*. The term quasi-experiment is often used to suggest a design in which certain structural features are added in an effort to provide information about hidden biases; see Section titled ‘Elaborate Theories’ for detailed discussion. In the Section titled ‘Effects on Children of Occupational Exposures to Lead’, for instance, data were collected for control children whose parents were not exposed to lead, together with data about the level of lead exposure and the hygiene of parents exposed to lead. An actual effect of lead should produce a quite specific pattern of associations: more lead in the blood of exposed children, more lead when the level of exposure is higher, more lead when the hygiene is poor. In general terms, Cook et al. [14] write that:

  > “... the warrant for causal inferences from quasi-experiments rests [on] structural elements of design other than random assignment–pretests, comparison groups, the way treatments are scheduled across groups ... – [which] provide the best way of ruling out threats to internal validity ... [C]onclusions are more plausible if they are based on evidence that corroborates numerous, complex, or numerically precise predictions drawn from a descriptive causal hypothesis.” (pp. 570-1)

Randomization will produce treated and control groups that were comparable prior to treatment, and it will do this mechanically, with no understanding of the context in which the study is being conducted. When randomization is not used, an understanding of the context becomes much more important.

Context is important whether one is trying to identify what covariates to measure, or to locate settings that afford haphazard treatment assignments or subpopulations with reduced selection biases, or to determine the direction of hidden biases. Ethnographic and other qualitative studies (e.g., [3, 21]) may provide familiarity with context needed in planning an observational study, and moreover qualitative methods may be integrated with quantitative studies [55].

Because even the most carefully designed observational study will have weaknesses and ambiguities, a single observational study is often not decisive, and replication is often necessary. In replicating an observational study, one should seek to replicate the actual treatment effects, if any, without replicating any biases that may have affected the original study. Some strategies for doing this are discussed in [48].

Adjusting for Biases Visible in Observed Covariates

Matched Sampling

Selecting from a Reservoir of Potential Controls. Among methods of adjustment for overt biases, the most direct and intuitive is matching, which compares each treated individual to one or more controls who appear comparable in terms of observed covariates. Matched sampling is most common when a small treated group is available together with a large reservoir of potential controls [57].

The structure of the study of bereavement by Lehman et al. [30] in the Section titled 'Long-term Psychological Effects of the Death of a Close Relative' is typical. There were 80 bereaved spouses and parents and 7581 potential controls, from whom 80 matched controls were selected. Routine administrative records were used to identify and match bereaved and control subjects, but additional information was needed from matched subjects for research purposes, namely, psychiatric outcomes. It is neither practical nor important to obtain psychiatric outcomes for all 7581 potential controls, and instead, matching selected 80 controls who appear comparable to treated subjects.

Most commonly, as in both Lehman et al.'s [30] study of bereavement in the Section titled 'Long-term Psychological Effects of the Death of a Close Relative' and Morton et al.'s [39] study of lead exposure in the Section titled 'Effects on Children of Occupational Exposures to Lead', each treated subject is matched to exactly one control, but other matching structures may yield either greater bias reduction or estimates with smaller standard errors or both. In particular, if the reservoir of potential controls is large, and if obtaining data from controls is not prohibitively expensive, then the standard errors of estimated treatment effects can be substantially reduced by matching each treated subject to several controls [62]. When several controls are used, substantially greater bias reduction is possible if the number of controls is not constant, instead varying from one treated subject to another [37].

Multivariate Matching Using Propensity Scores. In matching, the first impulse is to try to match each treated subject to a control who appears nearly the same in terms of observed covariates; however, this is quickly seen to be impractical when there are many covariates. For instance, with 20 binary covariates, there are 2^{20} or about a million types of individuals, so even with thousands of potential controls, it will often be difficult to find a control who matches a treated subject on all 20 covariates.

Randomization produces covariate balance, not perfect matches. Perfect matches are not needed to balance observed covariates. Multivariate matching methods attempt to produce matched pairs or sets that balance observed covariates, so that, in aggregate, the distributions of observed covariates are similar in treated and control groups. Of course, unlike randomization, matching cannot be expected to balance unobserved covariates.

The **propensity score** is the conditional probability (*see* **Probability: An Introduction**) of receiving the treatment rather than the control given the *observed* covariates [52]. Typically, the propensity score is unknown and must be estimated, for instance, using **logistic regression** [19] of the binary category, treatment/control on the observed covariates. The propensity score is defined in terms of the *observed* covariates, even when there are concerns about hidden biases due to *unobserved* covariates, so estimating the propensity score is straightforward because the needed data are available. For nontechnical surveys of methods using propensity scores, see [7, 27], and see [33] for discussion of propensity scores for doses of treatment.

Matching on one variable, the propensity score, tends to balance all of the observed covariates, even though matched individuals will typically differ on many observed covariates. As an alternative, matching on the propensity score and one or two other key covariates will also tend to balance all of the observed covariates. If it suffices to adjust for the observed covariates – that is, if there is no hidden bias due to unobserved covariates – then it also suffices to adjust for the propensity score alone. These results are Theorems 1 through 4 of [52]. A study of the psychological effects of prenatal exposures to barbiturates balanced 20 observed covariates by matching on an estimated propensity score and sex [54].

One can and should check to confirm that the propensity score has done its job. That is, one should check that, after matching, the distributions of observed covariates are similar in treated and control groups; see [53, 54] for examples of this simple process. Because theory says that a correctly estimated propensity score should balance observed covariates, this check on covariate balance is also a check on the model used to estimate the propensity score. If some covariates are not balanced, consider adding to the logit model interactions or quadratics involving these covariates; then check covariate balance with the new propensity score.

Bergstralh, Kosanke, and Jacobsen [4] provide SAS software for an optimal matching algorithm.

Stratification

Stratification is an alternative to matching in which subjects are grouped rather than paired. Cochran [13] showed that five strata formed from a single continuous covariate can remove about 90% of the bias in that covariate. Strata that balance many covariates at once can often be formed by forming five strata at the quintiles of an estimated propensity score. A study of coronary bypass surgery balanced 74 covariates using five strata formed from an estimated propensity score [53].

The optimal stratification – that is, the stratification that makes treated and control subjects as similar as possible within strata – is a type of matching called *full matching* in which a treated subject can be matched to several controls or a control can be matched to several treated subjects [45]. An optimal full matching, hence also an optimal stratification, can be determined using network optimization.

Model-based Adjustments

Unlike matched sampling and stratification, which compare treated subjects directly to actual controls who appear comparable in terms of observed covariates, model-based adjustments, such as covariance adjustment, use data on treated and control subjects without regard to their comparability, relying on a model, such as a linear regression model, to predict how subjects would have responded under treatments they did not receive. In a case study from labor economics, Dehejia and Wahba [20] compared the performance of model-based adjustments and matching, and Rubin [58, 60] compared performance using simulation. Rubin found that model-based adjustments yielded smaller standard errors than matching when the model is precisely correct, but model-based adjustments were less robust than matching when the model is wrong. Indeed, he found that if the model is substantially incorrect, model-based adjustments may not only fail to remove overt biases, they may even increase them, whereas matching and stratification are fairly consistent at reducing overt biases. Rubin found that the combined use of matching and model-based adjustments was both robust and efficient, and he recommended this strategy in practice.

Appraising Sensitivity to Hidden Bias

With care, matching, stratification, model-based adjustments and combinations of these techniques may often be used to remove overt biases accurately recorded in the data at hand, that is, biases visible in imbalances in observed covariates. However, when observational studies are subjected to critical evaluation, a common concern is that the adjustments failed to control for some covariate that was not measured. In other words, the concern is that treated and control subjects were not comparable prior to treatment with respect to this unobserved covariate, and had this covariate been measured and controlled by adjustments, then the conclusions about treatment effects would have been different. This is not a concern in randomized experiments, because randomization balances both observed and unobserved covariates. In an observational study, a **sensitivity analysis** (*see* **Sensitivity Analysis in Observational Studies**) asks how such hidden biases of various magnitudes might alter the conclusions of the study. Observational studies vary greatly in their sensitivity to hidden bias.

Cornfield et al. [17] conducted the first formal sensitivity analysis in a discussion of the effects of cigarette smoking on health. The objection had been raised that smoking might not cause lung cancer, but rather that there might be a genetic predisposition both to smoke and to develop lung cancer, and that this, not an effect of smoking, was responsible for the association between smoking and lung cancer. Cornfield et al. [17] wrote:

> . . . if cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone X-producers among cigarette smokers must be at least 9 times greater than among nonsmokers. (p. 40)

Though straightforward to compute, their sensitivity analysis is an important step beyond the familiar fact that association does not imply causation. A sensitivity analysis is a specific statement about the *magnitude* of hidden bias that would need to be present to explain the associations actually observed. Weak associations in small studies can be explained away by very small biases, but only a very large bias can explain a strong association in a large study.

A simple, general method of sensitivity analysis introduces a single sensitivity parameter Γ that measures the degree of departure from random assignment of treatments. Two subjects with the same observed covariates may differ in their odds of receiving the treatment by at most a factor of Γ. In an experiment, random assignment of treatments ensures that $\Gamma = 1$, so no sensitivity analysis is needed. In an observational study with $\Gamma = 2$, if two subjects were matched exactly for observed covariates, then one might be twice as likely as the other to receive the treatment because they differ in terms of a covariate not observed. Of course, in an observational study, Γ is unknown. A sensitivity analysis tries out several values of Γ to see how the conclusions might change. Would small departures from random assignment alter the conclusions? Or, as in the studies of smoking and lung cancer, would only very large departures from random assignment alter the conclusions? For each value of Γ, it is possible to place bounds on a statistical inference – perhaps for $\Gamma = 3$, the true P value is unknown, but must be between 0.0001 and 0.041. Analogous bounds may be computed for point estimates and confidence intervals. How large must Γ be before the conclusions of the study are qualitatively altered? If for $\Gamma = 9$, the P value for testing no effect is between 0.00001 and 0.02, then the results are highly insensitive to bias – only an enormous departure from random assignment of treatments could explain away the observed association between treatment and outcome. However, if for $\Gamma = 1.1$, the P value for testing no effect is between 0.01 and 0.3, then the study is extremely sensitive to hidden bias – a tiny bias could explain away the observed association. This method of sensitivity analysis is discussed in detail with many examples in Section 4 of [49] and the references given there.

For instance, Morton et al.'s [39] study in the Section titled 'Effects on Children of Occupational Exposures to Lead' used Wilcoxon's **signed-ranks test** to compare blood lead levels of 33 exposed and 33 matched control children. The pattern of matched pair differences they observed would yield a P value less than 10^{-5} in a randomized experiment. For $\Gamma = 3$, the range of possible P values is from about 10^{-15} to 0.014, so a bias of this magnitude could not explain the higher lead levels among exposed children. In words, if Morton et al. [39] had failed to control by matching a variable strongly related to blood lead levels and three times more common among exposed children, this would not have been likely to produce a difference in lead levels as large as the one they observed. The upper bound on the P value is just about 0.05 when $\Gamma = 4.35$, so the study is quite insensitive to hidden bias, but not as insensitive as the studies of heavy smoking and lung cancer. For $\Gamma = 5$ and $\Gamma = 6$, the upper bounds on the P value are 0.07 and 0.12, respectively, so biases of this magnitude could explain the observed association. Sensitivity analyses for point estimates and confidence intervals for this example are in Section 4.3.4 and Section 4.3.5 of [49].

Several other methods of sensitivity analysis are discussed in [16, 24, 26], and [31].

Elaborate Theories

Elaborate Theories and Pattern Specificity

What *can* be observed to provide evidence about hidden biases, that is, biases due to covariates that were *not* observed? Cochran [12] summarizes the view of **Sir Ronald Fisher**, the inventor of the randomized experiment:

> About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: "Make your theories elaborate." The reply puzzled me at first, since by Occam's razor, the advice usually given is to make theories as simple as is consistent with known data. What Sir Ronald meant, as subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold.

Similarly, Cook & Shadish [15] (1994, p. 565): "Successful prediction of a complex pattern of multivariate results often leaves few plausible alternative explanations. (p. 95)" Some patterns of response are scientifically plausible as treatment effects, but others are not [25], [65]. "[W]ith more pattern specificity," writes Trochim [63], "it is generally less likely that plausible alternative explanations for the observed effect pattern will be forthcoming. (p. 580)"

Example of Reduced Sensitivity to Hidden Bias Due to Pattern Specificity

Morton et al.'s [39] study of lead exposures in the Section titled 'Effects on Children of Occupational Exposures to Lead' provides an illustration. Their elaborate theory predicted: (a) higher lead levels in the blood of exposed children than in matched control children, (b) higher lead levels in exposed children whose parents had higher lead exposure on the job, and (c) higher lead levels in exposed children whose parents practiced poorer hygiene upon leaving the factory. Since each of these predictions was consistent with observed data, to attribute the observed associations to hidden bias rather than an actual effect of lead exposure, one would need to postulate biases that could produce all three associations.

In a formal sense, elaborate theories play two roles: (a) they can aid in detecting hidden biases [49], and (b) they can make a study less sensitive to hidden bias [50, 51]. In Section titled 'Effects on Children of Occupational Exposures to Lead', if exposed children had lower lead levels than controls, or if higher exposure predicted lower lead levels, or if poor hygiene predicted lower lead levels, then this would be difficult to explain as an effect caused by lead exposure, and would likely be understood as a consequence of some unmeasured bias, some way children who appeared comparable were in fact not comparable. Indeed, the pattern specific comparison is less sensitive to hidden bias. In detail, suppose that the exposure levels are assigned numerical scores, 1 for a child whose father had either low exposure or good hygiene, 2 for a father with high exposure and poor hygiene, and 1.5 for the several intermediate situations. The sensitivity analysis discussed in the Section titled 'Appraising Sensitivity to Hidden Bias' used the signed rank test to compare lead levels of the 33 exposed children and their 33 matched controls, and it became sensitive to hidden bias at $\Gamma = 4.35$, because the upper bound on the P value for testing no effect had just reached 0.05. Instead, using the dose-signed-rank statistic [46, 50] to incorporate the predicted pattern, the comparison becomes sensitive at $\Gamma = 4.75$; that is, again, the upper bound on the P value for testing no effect is 0.05. In other words, some biases that would explain away the higher lead levels of exposure children are not large enough to explain away the pattern of associations predicted by the elaborate theory. To explain the entire pattern, noticeably larger biases would need to be present.

A reduction in sensitivity to hidden bias can occur when a correct elaborate theory is strongly confirmed by the data, but an increase in sensitivity can occur if the pattern is contradicted [50]. It is possible to contrast competing design strategies in terms of their 'design sensitivity;' that is, their ability to reduce sensitivity to hidden bias [51].

Common Forms of Pattern Specificity

There are several common forms of pattern specificity or elaborate theories [44, 49, 61].

- *Two control groups.*Campbell [8] advocated selecting two control groups to systematically vary an unobserved covariate, that is, to select two different groups not exposed to the treatment, but known to differ in terms of certain unobserved covariates. For instance, Card and Krueger [11] examined the common claim among economists that increases in the minimum wage cause many minimum wage earners to loose their jobs. They did this by looking at changes in employment at fast food restaurants – Burger Kings, Wendy's, KFCs, and Roy Rogers' – when New Jersey increased its minimum wage by about 20%, from $4.25 to $5.05 per hour, on 1 April 1992, comparing employment before and after the increase. In certain analyses, they compared New Jersey restaurants initially paying $4.25 to two control groups: (a) restaurants in the same chains across the Delaware River in Pennsylvania where the minimum wage had not increased, and (b) restaurants in the same chains in affluent sections of New Jersey where the starting wage was at least $5.00 before 1 April 1992. An actual effect of raising the minimum wage should have negligible effects on both control groups. In contrast, one anticipates differences between the two control groups if, say, Pennsylvania Burger Kings were poor controls for New Jersey Burger Kings, or if employment changes in relatively affluent sections of New Jersey are very different from those in less affluent sections. Card and Krueger found similar changes in employment in the two control groups, and similar results in their comparisons of the treated group with either control group. An algorithm for optimal pair matching with two control groups is illustrated with Card and Krueger's study in [32].

- *Coherence among several outcomes and/or several doses*. Hill [25] emphasized the importance of a coherent pattern of associations and of dose-response relationships, and Weiss [66] further developed these ideas. Campbell [9] wrote: "... great inferential strength is added when each theoretical parameter is exemplified in two or more ways, each mode being as independent as possible of the other, as far as the theoretically irrelevant components are concerned (p. 33)." Webb [64] speaks of triangulation. The lead example in the Section titled 'Example of Reduced Sensitivity to Hidden Bias Due to Pattern Specificity' provides one illustration and Reynolds and West [42] provide another. Related statistical theory is in [46, 51] and the references given there.
- *Unaffected outcomes; ineffective treatments.*An elaborate theory may predict that certain outcomes should not be affected by the treatment or certain treatments should not affect the outcome; see Section 6 of [49] and [67]. For instance, in a *case-crossover study* [34], Mittleman et al. [38] asked whether bouts of anger might cause myocardial infarctions or heart attacks, finding a moderately strong and highly significant association. Although there are reasons to think that bouts of anger might cause heart attacks, there are also reasons to doubt that bouts of curiosity cause heart attacks. Mittleman et al. found curiosity was not associated with myocardial infarction, writing: "the specificity observed for anger ... as opposed to curiosity ... argue against recall bias." McKillip [35] suggests that an unaffected or 'control' outcome might sometimes serve in place of a control group, and Legorreta et al. [29] illustrate this possibility in a study of changes in the demand for a type of surgery following a technological change that reduced cost and increased safety.

Summary

In the design of an observational study, an attempt is made to reconstruct some of the structure and strengths of an experiment. Analytical adjustments, such as matching, are used to control for overt biases, that is, pretreatment differences between treated and control groups that are visible in observed covariates. Analytical adjustments may fail because of hidden biases, that is, important covariates that were not measured and therefore not controlled by adjustments. Sensitivity analysis indicates the magnitude of hidden bias that would need to be present to alter the qualitative conclusions of the study. Observational studies vary markedly in their sensitivity to hidden bias; therefore, it is important to know whether a particular study is sensitive to small biases or insensitive to quite large biases. Hidden biases may leave visible traces in observed data, and a variety of tactics involving pattern specificity are aimed at distinguishing actual treatment effects from hidden biases. Pattern specificity may aid in detecting hidden bias or in reducing sensitivity to hidden bias.

Acknowledgment

This work was supported by grant SES-0345113 from the US National Science Foundation.

References

[1] Angrist, J.D. (2003). Randomized trials and quasi-experiments in education research. *NBER Reporter*, Summer, 11–14.

[2] Angrist, J.D. & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement, *Quarterly Journal of Economics* **114**, 533–575.

[3] Athens, L. (1997). *Violent Criminal Acts and Actors Revisited*, University of Illinois Press, Urbana.

[4] Bergstralh, E.J., Kosanke, J.L. & Jacobsen, S.L. (1996). Software for optimal matching in observational studies, *Epidemiology* **7**, 331–332. `http://www.mayo.edu/hsr/sasmac.html`

[5] Boruch, R. (1997). *Randomized Experiments for Planning and Evaluation*, Sage Publications, Thousand Oaks.

[6] Bound, J. (1989). The health and earnings of rejected disability insurance applicants, *American Economic Review* **79**, 482–503.

[7] Braitman, L.E. & Rosenbaum, P.R. (2002). Rare outcomes, common treatments: analytic strategies using propensity scores, *Annals of Internal Medicine* **137**, 693–695.

[8] Campbell, D.T. (1969). Prospective: artifact and Control, in *Artifact in Behavioral Research*, R. Rosenthal & R. Rosnow, eds, Academic Press, New York.

[9] Campbell, D.T. (1988). *Methodology and Epistemology for Social Science: Selected Papers*, University of Chicago Press, Chicago, pp. 315–333.

[10] Campbell, D.T. & Boruch, R.F. (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*, C.A. Bennett & A.A. Lumsdaine, eds, Academic Press, New York, pp. 195–296.

[11] Card, D. & Krueger, A. (1994). Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania, *American Economic Review* **84**, 772–793.

[12] Cochran, W.G. (1965). The planning of observational studies of human populations (with Discussion), *Journal of the Royal Statistical Society. Series A* **128**, 134–155.

[13] Cochran, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies, *Biometrics* **24**, 295–313.

[14] Cook, T.D., Campbell, D.T. & Peracchio, L. (1990). Quasi-experimentation, in *Handbook of Industrial and Organizational Psychology*, M. Dunnette & L. Hough, eds, Consulting Psychologists Press, Palo Alto, pp. 491–576.

[15] Cook, T.D. & Shadish, W.R. (1994). Social experiments: Some developments over the past fifteen years, *Annual Review of Psychology* **45**, 545–580.

[16] Copas, J.B. & Li, H.G. (1997). Inference for non-random samples (with discussion), *Journal of the Royal Statistical Society. Series B* **59**, 55–96.

[17] Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. & Wynder, E. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions, *Journal of the National Cancer Institute* **22**, 173–203.

[18] Cox, D.R. & Read, N. (2000). *The Theory of the Design of Experiments*, Chapman & Hall/CRC, New York.

[19] Cox, D.R. & Snell, E.J. (1989). *Analysis of Binary Data*, Chapmann & Hall, London.

[20] Dehejia, R.H. & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs, *Journal of the American Statistical Association* **94**, 1053–1062.

[21] Estroff, S.E. (1985). *Making it Crazy: An Ethnography of Psychiatric Clients in an American Community*, University of California Press, Berkeley.

[22] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.

[23] Friedman, L.M., DeMets, D.L. & Furberg, C.D. (1998). *Fundamentals of Clinical Trials*, Springer-Verlag, New York.

[24] Gastwirth, J.L. (1992). Methods for assessing the sensitivity of statistical comparisons used in title VII cases to omitted variables, *Jurimetrics* **33**, 19–34.

[25] Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* **58**, 295–300.

[26] Imbens, G.W. (2003). Sensitivity to exogeneity assumptions in program evaluation, *American Economic Review* **93**, 126–132.

[27] Joffe, M.M. & Rosenbaum, P.R. (1999). Propensity scores, *American Journal of Epidemiology* **150**, 327–333.

[28] Kessler, R.C. (1997). The effects of stressful life events on depression, *Annual Review of Psychology* **48**, 191–214.

[29] Legorreta, A.P., Silber, J.H., Costantino, G.N., Kobylinski, R.W. & Zatz, S.L. (1993). Increased cholecystectomy rate after the introduction of laparoscopic cholecystectomy, *Journal of the American Medical Association* **270**, 1429–1432.

[30] Lehman, D., Wortman, C. & Williams, A. (1987). Long-term effects of losing a spouse or a child in a motor vehicle crash, *Journal of Personality and Social Psychology* **52**, 218–231.

[31] Lin, D.Y., Psaty, B.M. & Kronmal, R.A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies, *Biometrics* **54**, 948–963.

[32] Lu, B. & Rosenbaum, P.R. (2004). Optimal pair matching with two control groups, *Journal of Computational and Graphical Statistics* **13**, 422–434.

[33] Lu, B., Zanutto, E., Hornik, R. & Rosenbaum, P.R. (2001). Matching with doses in an observational study of a media campaign against drug abuse, *Journal of the American Statistical Association* **96**, 1245–1253.

[34] Maclure, M. & Mittleman, M.A. (2000). Should we use a case-crossover design? *Annual Review of Public Health* **21**, 193–221.

[35] McKillip, J. (1992). Research without control groups: a control construct design, in *Methodological Issues in Applied Social Psychology*, F.B. Bryant, J. Edwards & R.S. Tindale, eds, Plenum Press, New York, pp. 159–175.

[36] Meyer, B.D. (1995). Natural and quasi-experiments in economics, *Journal of Business and Economic Statistics* **13**, 151–161.

[37] Ming, K. & Rosenbaum, P.R. (2000). Substantial gains in bias reduction from matching with a variable number of controls, *Biometrics* **56**, 118–124.

[38] Mittleman, M.A., Maclure, M., Sherwood, J.B., Mulry, R.P., Tofler, G.H., Jacobs, S.C., Friedman, R., Benson, H. & Muller, J.E. (1995). Triggering of acute myocardial infarction onset by episodes of anger, *Circulation* **92**, 1720–1725.

[39] Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M. & Saah, M. (1982). Lead absorption in children of employees in a lead-related industry, *American Journal of Epidemiology* **115**, 549–555.

[40] Neyman, J. (1923). On the application of probability theory to agricultural experiments: essay on principles, Section 9 (In Polish), *Roczniki Nauk Roiniczych*, **Tom X**, 1–51, Reprinted in English with Discussion in *Statistical Science* 1990,. **5**, 463–480.

[41] Oreopoulos, P. (2003). The long-run consequences of living in a poor neighborhood, *Quarterly Journal of Economics* **118**, 1533–1575.

[42] Reynolds, K.D. & West, S.G. (1987). A multiplist strategy for strengthening nonequivalent control group designs, *Evaluation Review* **11**, 691–714.

[43] Rosenbaum, P.R. (1984a). The consequences of adjustment for a concomitant variable that has been affected by the treatment, *Journal of the Royal Statistical Society. Series A* **147**, 656–666.

[44] Rosenbaum, P.R. (1984b). From association to causation in observational studies, *Journal of the American Statistical Association* **79**, 41–48.

[45] Rosenbaum, P.R. (1991). A characterization of optimal designs for observational studies, *Journal of the Royal Statistical Society. Series B* **53**, 597–610.

[46] Rosenbaum, P.R. (1997). Signed rank statistics for coherent predictions, *Biometrics* **53**, 556–566.

[47] Rosenbaum, P.R. (1999). Choice as an alternative to control in observational studies (with Discussion), *Statistical Science* **14**, 259–304.

[48] Rosenbaum, P.R. (2001). Replicating effects and biases, *American Statistician* **55**, 223–227.

[49] Rosenbaum, P.R. (2002). *Observational Studies*, 2nd Edition, Springer-Verlag, New York.

[50] Rosenbaum, P.R. (2003). Does a dose-response relationship reduce sensitivity to hidden bias? *Biostatistics* **4**, 1–10.

[51] Rosenbaum, P.R. (2004). Design sensitivity in observational studies, *Biometrika* **91**, 153–164.

[52] Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.

[53] Rosenbaum, P. & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score, *Journal of the American Statistical Association* **79**, 516–524.

[54] Rosenbaum, P. & Rubin, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score, *American Statistician* **39**, 33–38.

[55] Rosenbaum, P.R. & Silber, J.H. (2001). Matching and thick description in an observational study of mortality after surgery, *Biostatistics* **2**, 217–232.

[56] Rosenzweig, M.R. & Wolpin, K.I. (2000). Natural "natural experiments" in economics, *Journal of Economic Literature* **38**, 827–874.

[57] Rubin, D. (1973a). Matching to remove bias in observational studies, *Biometrics* **29**, 159–183.Correction: 1974,. **30**, 728.

[58] Rubin, D.B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies, *Biometrics* **29**, 185–203.

[59] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* **66**, 688–701.

[60] Rubin, D.B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies, *Journal of the American Statistical Association* **74**, 318–328.

[61] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton-Mifflin, Boston.

[62] Smith, H.L. (1997). Matching with multiple controls to estimate treatment effects in observational studies, *Sociological Methodology* **27**, 325–353.

[63] Trochim, W.M.K. (1985). Pattern matching, validity and conceptualization in program evaluation, *Evaluation Review* **9**, 575–604.

[64] Webb, E.J. (1966). Unconventionality, triangulation and inference, *Proceedings of the Invitational Conference on Testing Problems*, Educational Testing Service, Princeton, pp. 34–43.

[65] Weed, D.L. & Hursting, S.D. (1998). Biologic plausibility in causal inference: current method and practice, *American Journal of Epidemiology* **147**, 415–425.

[66] Weiss, N. (1981). Inferring causal relationships: elaboration of the criterion of 'dose-response', *American Journal of Epidemiology* **113**, 487–90.

[67] Weiss, N.S. (2002). Can the "specificity" of an association be rehabilitated as a basis for supporting a causal hypothesis? *Epidemiology* **13**, 6–8.

[68] Wright, M.A., Wintemute, G.J. & Rivara, F.P. (1999). Effectiveness of denial of handgun purchase to persons believed to be at high risk for firearm violence, *American Journal of Public Health* **89**, 88–90.

[69] Zabin, L.S., Hirsch, M.B. & Emerson, M.R. (1989). When urban adolescents choose abortion: effects on education, psychological status, and subsequent pregnancy, *Family Planning Perspectives* **21**, 248–255.

PAUL R. ROSENBAUM