Statistical Science
1987, Vol. 2, No. 3, 292–316

The Role of a Second Control Group in an Observational Study

Paul R. Rosenbaum

Abstract. An observational study is an attempt to estimate the effects of a treatment when subjects are not randomly assigned to treatment or control. The possibility of using more than one control group has often been briefly mentioned in general discussions of observational studies, and many observational studies have used two control groups. Here, the limited and dispersed literature on this subject is reviewed, and the topic is developed in several directions by using a formal notation for observational studies. The value of a second control group depends on the supplementary information that is available about unobserved biases that are suspected to exist. A second control group provides a test of the assumption that conventional adjustments for observed covariates suffice in estimating treatment effects. Under the best of circumstances, this test is consistent and unbiased, and its power exceeds the probability of falsely detecting a treatment effect. Indeed, under the best of circumstances, two control groups can yield consistent and unbiased estimates of bounds on the treatment effect when conventional adjustments fail. In contrast, however, in the worst of circumstances, a second control group can be of little value.

Key words and phrases: Observational studies, control groups, adjustable treatment assignment, ignorable treatment assignment, case-control studies, unobserved covariates.

1. INTRODUCTION—SOME EXAMPLES

1.1 Observational Studies

Cochran (1965, page 234) defined an observational study as an empirical investigation 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."

In such studies, because of the absence of random assignment, treated and control subjects may be quite different prior to treatment. Differences in outcomes after treatment may, therefore, reflect either effects caused by the treatment, or the initial lack of comparability, or both to some extent. When a relevant pretreatment variable or covariate, X, is accurately measured, adjustments by matching, subclassification or model-based procedures can often reduce the bias due to pretreatment differences (e.g., Cochran and Rubin, 1973; Rubin, 1977). Still, there is usually reason for concern that treated and control subjects may differ in ways that have not been measured, in which case adjustments for X need not eliminate the bias. When adjustments for X suffice to remove all of the bias, treatment assignment is said to be X-adjustable, a term defined formally in Section 2.

The current paper considers the role that multiple control groups can play in investigating unobserved pretreatment differences. The paper is organized as follows. To place the role of multiple control groups in proper context, Sections 1.2 and 1.3 briefly review the two basic approaches to assessing the impact of unobserved pretreatment differences, namely sensitivity analyses and tests of X-adjustable treatment assignment. These two approaches are complementary, and the use of multiple control groups is an instance of the second approach. Section 1.4 contains brief discussions of several examples of the use of multiple

Paul R. Rosenbaum is Associate Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104. *An early version of this article was the basis for a talk at the Workshop on Nonrandomized Experimentation sponsored by the Committee on Toxicology of the National Research Council and held at the National Academy of Sciences on November 19, 1985.*

control groups, whereas Sections 4.2 and 4.4 contain detailed discussion of one example. In Section 2, some general concepts and notation for observational studies are reviewed. The role of a second control group in cohort (or prospective) studies is discussed in Section 3. Case-control studies are somewhat different because the "control" groups in such studies are not untreated groups, but rather noncase groups. Section 4 discusses the role of multiple "control" groups in a case-control study.

1.2 Addressing Biases Due to Unobserved Pretreatment Differences: Sensitivity Analysis

There are two basic ways of addressing biases due to unobserved pretreatment differences. One involves a sensitivity analysis: the sensitivity of conclusions to a range of plausible assumptions about an unobserved covariate, U, is investigated. For example, Cornfield, Haenszel, Hammond, Lilienfeld, Shimkin and Wynder (1959) found that, if failure to adjust for an unobserved covariate is to account for the entire apparent effect of heavy cigarette smoking on the risk of lung cancer, then that covariate must have an extremely strong—indeed, implausibly strong—association with both cigarette smoking and lung cancer risk. This is true because of the very strong relationship between heavy cigarette smoking and lung cancer. Other discussions of methods for sensitivity analysis in observational studies are given by Bross (1966, 1967), Rubin (1978, Section 4.2), Rosenbaum and Rubin (1983b) and Rosenbaum (1984c, Section 4; 1986, 1987).

It is always useful to know and to report whether an inference is sensitive to plausible biases, and such sensitivity is always reason for caution in interpretation. Still, sensitivity to modest bias, by itself, does not imply that bias is present, but only that such a bias, if present, would substantially alter the study's conclusion. Some treatments have important but small effects; conclusions from studies of such effects will inevitably be judged sensitive. For example, heavy smoking is believed to increase the risk of death from cardiovascular diseases by the relatively small factor of about 1.5; an effect of this size would be judged quite sensitive to bias by the method of Cornfield, Haenszel, Hammond, Lilienfeld, Shimkin and Wynder (1959). Still, if this is a real effect, as it is now generally thought to be, then more people die of smoking-induced cardiovascular disease than of smoking-induced lung cancer, simply because the base risk of death from cardiovascular disease is so much greater (Bayne-Jones, Burdette, Cochran, Farber, Fieser, Furth, Hickam, LeMaistre, Schuman and Seevers, 1964, page 317). The point here is that a treatment may have an effect that is small in the sense relevant to a sensitivity analysis but large in other important respects. We cannot, therefore, adopt a methodological rule of rejecting as unsubstantiated every conclusion that is sensitive to modest biases; however, much such sensitivity may properly concern us. There is, then, a need for another complementary approach to addressing biases due to unobserved pretreatment differences.

1.3 Addressing Biases Due to Unobserved Pretreatment Differences: Tests of X-Adjustable Treatment Assignment

The second basic approach to addressing possible biases due to unobserved covariates involves, first, assuming tentatively that adjustments for X suffice to remove bias and, second, testing this tentative assumption at each point of contact with observable data. This process invariably requires the use of some supplementary information; often, this is qualitative information about the process by which the treatment is thought to produce its effects. If data contradict testable consequences of the conjunction of this supplementary information and the tentative assumption that adjustments for X suffice to remove bias, then there may be some uncertainty as to which part of the conjunction is contradicted. Still, since we began with the concern that adjustments for X might not suffice, such a finding would considerably heighten that concern.

In contrast, to find no contradiction with observable data is to *corroborate* the hypothesis that adjustments for X suffice to remove bias. Here, the term corroborate is used in the precise and slightly technical sense defined by Popper (1959): to corroborate a theory or hypothesis is to expose it to possible refutation but fail to refute it. Corroboration is a matter of degree: a theory is more thoroughly corroborated if it has survived severe, extensive attempts of refutation. In observational studies, little formal work has been done to measure and clarify the severity of tests of the hypothesis that adjustments for X suffice, and, therefore, the degree of corroboration provided by passing such tests has often been unclear. At several points, the current paper attempts to provide such clarification.

The Doll and Hill (1966) study of the effect of smoking on coronary thrombosis contains a good example. Within subclasses defined by covariates X, Doll and Hill classified smokers by the amount smoked, and cross-classified former smokers by the amount last smoked and the time since quitting. If adjustments for X suffice to eliminate bias, then within subclasses defined by X, we expect an actual effect of smoking to cause the highest risks for continuing heavy smokers, lower risks for both continuing light smokers and former heavy smokers, still lower risks for former light

smokers and the lowest risks for nonsmokers. If the data are inconsistent with these predictions, then there is evidence that adjustments for X alone do not suffice to remove bias.

Good, informal discussions of this approach to addressing possible biases due to unobserved covariates are given by Yerushalmy and Palmer (1959), Campbell and Stanley (1963), Cochran (1965, Section 5; 1972, pages 88 and 89), Hill (1965) and Lilienfeld and Lilienfeld (1980, Section 12.B). A somewhat more formal discussion, treating the approach as a part of statistical theory, is given by Rosenbaum (1984a).

The purpose of the current paper is to consider the role that a second control group plays in checking the hypothesis that adjustments for X suffice to remove bias. As will be seen, under the best circumstances a second control group can bring us fairly close to a consistent and unbiased test of the hypothesis that adjustments for X suffice, and, moreover, can provide bounds on the size of the biases of conventional estimates of treatment effects. In contrast, in the worst of circumstances a second control may be of little value, and may even foster misinterpretation (see Section 3.7). The issue turns on the supplementary information that can be brought to bear, and on whether control groups can be selected to address specific biases.

1.4 Multiple Control Groups: Some Examples

The possibility of using more than one control group in an observational study has been given a brief mention in several general discussions of observational studies; e.g., Mantel and Haenszel (1959, pages 726 and 727), Yerushalmy and Palmer (1959, page 34), Cochran (1963, pages 485 and 486; 1965, page 248; 1967, page 324), Campbell (1969) and Cole (1979, pages 22 and 24). Several examples follow. In each instance, the second control group addresses specific limitations of the first, but has limitations of its own.

EXAMPLE 1 (Spouses and Siblings). Gutensohn, Li, Johnson and Cole (1975) studied the possible effect of tonsillectomy on the risk of Hodgkin's disease in a case-control study. One control group consisted of the spouses of the Hodgkin's patients and a second consisted of the patient's siblings. Clearly, spouses tend to share the patient's adult home environment, while siblings tend to share the patient's childhood environment; neither group typically shares the patient's work environment. The estimated risk ratio with spouses as controls was substantially higher than with siblings as controls (3.1 versus 1.4). Moreover, the risk ratio differed significantly from 1 for the 78 spouse pairs, but not for the 119 sibling matched sets, containing 315 sibling controls. This example is discussed briefly by Cole (1979).

EXAMPLE 2 (Unaccepted and Unoffered Treatments). A subject may become a control either because the treatment was not offered or because although offered, it was declined. For example, the College Board's Advanced Placement (AP) Program provides high school students with the opportunity to earn college credit for work done in high school. Not all high schools offer the AP program, and in those that do, only a small, typically elite minority of students elect to participate. Recently, the Educational Testing Service (ETS) sought to evaluate the impact of the AP program on college achievement, and the possibility of using two control groups was considered. One control group would have been formed by matching AP students with other students from the same school who declined participation. The second control group would have matched AP students in one school with students in another school where the AP program was not available. In both cases, the matching would have been based on measures of academic performance prior to the senior year, the year in which AP programs are offered. The concerns here were twofold. First, qualified students who declined participation in the AP program might be less motivated for academic work, in which case the impact of the greater motivation of AP students might be mistaken for an effect of the program itself. Second, schools that declined to offer the AP program may be smaller, or may have more limited resources or may be in economically poorer neighborhoods, and the impact of these differences might also be mistaken for a program effect. (Ultimately, only the first control group was used in this study, primarily because of cost limitations, but partly because of the considerations discussed in Section 3.7). This example is connected with the work of Warren Willingham at ETS.

EXAMPLE 3 (Rejected for Treatment versus Ineffectively Treated Controls). The Committee on Toxicology of the National Research Council is completing a study of historical and follow-up data on the long term effects of a nonrandomized experimental exposure of soldiers to a wide range of chemical agents. Relatively few soldiers were exposed to each agent, and exposures to multiple agents were occasional and haphazard, not at all resembling a factorial experiment. The study compares soldiers receiving one chemical agent with two other groups, an untreated group and a second group consisting of all individuals exposed to all other chemical agents. Usually, there would be a clear preference for the untreated group as the control group, but in this instance the untreated group consisted of soldiers specifically rejected for treatment, in part on the basis of a medical examination, the results of which were not retained. The study's authors suspected that most of the chemical agents had little or no long term effects, so that most individuals in the

"other agent" group were suspected to have received ineffective treatments. The concerns were twofold: the untreated controls might be in poorer health because they were rejected on the basis of a medical examination, whereas the "other agent" control group might include some individuals exposed to active agents.

There are many other examples of the use of multiple control groups, including Solomon (1949), Collaborative Group for the Study of Stroke in Young Women (1973), Hiller, Giacometti and Yuen (1977), Halsey, Modlin, Jabbour, Dubey, Eddins and Ludwig (1980) and MacMahon (1984). These examples include: hospital and neighborhood controls; hospital and playmate controls in a study of young children; controls based on several disease groups in a case-control study; and controls subject to different pretreatment measurement procedures.

2. OBSERVATIONAL STUDIES: A SHORT REVIEW OF CERTAIN CONCEPTS AND NOTATION

2.1 Treatments and the Effects Caused By Treatments

For later use, the current section briefly reviews some concepts and notation for observational studies that have been developed in detail by Rubin (1974, 1977, 1978), Hamilton (1979), Holland and Rubin (1980, 1983), Rosenbaum and Rubin (1983a, 1983b, 1985), Rosenbaum (1984a, 1984b, 1984c, 1987) and Holland (1986b). The notation adapts and extends that used in the traditional literature on experimental design, for example, in the books by Fisher (1935), Kempthorne (1952) and Cox (1958). The review that follows is a quick summary; for a more precise discussion of the associated sampling distributions as used here, see Rosenbaum and Rubin (1985, Section 1 and Figure 1).

A *treatment* is an intervention that can, in principle, be given to or withheld from any experimental subject. Exposure to a hazardous substance is a treatment in this sense, whereas age and gender are not treatments.

Consider a response that a subject may exhibit, such as a cognitive test score or the development of a particular disease. A treatment's effect on a specific subject is a comparison of the two responses that the subject would exhibit if the treatment were applied or withheld. More formally, each subject has two possible versions of the response: one response, R_T, that would be exhibited if the *t*reatment were applied, and a second response, R_C, that would be exhibited if the treatment were withheld, i.e., if the *c*ontrol were applied. The treatment has no effect if the response of each subject is unchanged by the application of the treatment, that is, if $R_T = R_C$ for each subject. This is the formal definition of "no effect" used in Fisher's (1935) randomization test in randomized experiments (cf. Rubin, 1980). More generally, the effect of the treatment on a subject is a comparison of the two responses that subject could exhibit, e.g., $R_T - R_C$ (e.g., Welch, 1937; Kempthorne, 1952, Section 8; Rubin, 1974, 1977, 1978; Hamilton, 1979; Holland, 1986b).

2.2 Randomized Experiments

In the simplest randomized experiment, subjects are sampled from a population and are assigned by the flip of a fair coin to one of two groups, a treated group and a single control group. For each subject, let Z indicate the group to which the subject is assigned, with $Z = 0$ for treated subjects and $Z = 1$ for control subjects. In this case, R_T is observed for subjects with $Z = 0$, and R_C is observed for subjects with $Z = 1$.

Randomized assignment implies that the treatment Z assigned to a subject is unrelated to any attribute of that subject, and, in particular, is unrelated to (or statistically independent of) the subject's pair of responses, (R_T, R_C), or in Dawid's (1979) notation for independence:

$$(R_T, R_C) \perp\!\!\!\perp Z. \tag{2.1}$$

(Recall Dawid's notation: $A \perp\!\!\!\perp B$ if A and B are independent, and $A \perp\!\!\!\perp B \mid C$ if A and B are conditionally independent given C).

It follows from (2.1) that the expected response of control subjects, namely $E(R_C \mid Z = 1)$, equals the expected response of the treated subjects had they instead been exposed to the control, namely $E(R_C \mid Z = 0)$. In randomized experiments, therefore, the average effect of the treatment on the treated subjects, namely $E(R_T - R_C \mid Z = 0)$, satisfies

$$\begin{aligned} &E(R_T - R_C \mid Z = 0) \\ &= E(R_T \mid Z = 0) - E(R_C \mid Z = 0) \\ &= E(R_T \mid Z = 0) - E(R_C \mid Z = 1), \end{aligned} \tag{2.2}$$

so that $E(R_T - R_C \mid Z = 0)$ may be estimated by the difference in mean responses in the treated and control groups.

In general, (2.2) need not hold in observational studies because (2.1) does not generally hold; in other words, treated ($Z = 0$) and control ($Z = 1$) groups may differ in ways that are relevant to the response (R_T, R_C). To control for these pretreatment differences, adjustments are often made for a vector X of observed covariates or pretreatment variables. Several methods of adjustment are common, including matching on X, subclassification or stratification on X, and model based adjustments for X (cf. Cochran and Rubin, 1973). Because such adjustments are confined to observed covariates although the groups may also

differ in ways that have not been observed, there is no guarantee that adjustments for X will yield appropriate estimates of treatment effects. Such adjustments will suffice, however, when treatment assignment is X-adjustable.

2.3 X-Adjustable Treatment Assignment: An Assumption Implicit in the Use of Conventional Methods of Adjustment

Treatment assignment is X-adjustable when two conditions hold: first, the treatment group indicator, Z, is unrelated to the response pair, (R_T, R_C), within each subpopulation defined by a value of X, and, second, at each value of X, a fraction of the population falls in each treatment group. Formally, treatment assignment is X-adjustable if, in Dawid's (1979) notation for conditional independence,

$$(R_T, R_C) \perp\!\!\!\perp Z \mid X \tag{2.3a}$$

and

$$0 < \text{pr}(Z = z \mid X = x) < 1 \quad \text{for each } z \text{ and } x. \tag{2.3b}$$

At times, it is useful to be explicit about the response variables involved in X-adjustable assignment, in which case, (2.3) will be called $(R_T, R_C \mid X)$-adjustable assignment. For example, if (2.3a) were replaced by $R_C \perp\!\!\!\perp Z \mid X$, then the condition would be $(R_C \mid X)$-adjustable assignment, but the converse is untrue without additional conditions. (In Rosenbaum and Rubin (1983a), condition (2.3) is called "strongly ignorable treatment assignment for (R_T, R_C) given X"; however, the term "$(R_T, R_C \mid X)$-adjustable assignment" is more compact and suggestive, because adjustments for X suffice when (2.3) holds.) In particular, treatment assignment is $(R_T, R_C \mid X)$-adjustable, and hence also $(R_C \mid X)$-adjustable, (i) in conventional randomized experiments in which treatment groups are formed by flipping a fair coin, (ii) in randomized experiments in which treatment groups are formed by flipping biased coins, where the bias is a (possibly unknown) function of X alone (Rubin, 1977) and (iii) in observational studies in which treatment groups are formed on the basis of X and some other irrelevant covariates (Rosenbaum, 1984a, Section 2.3).

It is easy to show that when (2.3) holds, a wide variety of adjustments for X provide appropriate estimates of treatment effects (e.g., Rubin, 1977; Rosenbaum and Rubin, 1983a; Rosenbaum, 1984b). In the simplest case, we might randomly sample a treated $(Z = 0)$ subject from the population, note that subject's value x of X, and then randomly sample a control $(Z = 1)$ subject from among control subjects having the same value x of X. When (2.3) holds, the (conditional) expected difference in responses in this pair given the common value x of X is:

$$\begin{aligned} &E\{R_T \mid Z = 0, X = x\} - E\{R_C \mid Z = 1, X = x\} \\ &= E\{R_T \mid Z = 0, X = x\} \\ &\quad - E\{R_C \mid Z = 0, X = x\} \quad \text{by (2.3a).} \end{aligned} \tag{2.4}$$

Continuing when (2.3) holds, the (marginal) expected difference in responses in pairs obtained in this way is (2.4) averaged with respect to the distribution of X in the treated group, $\text{pr}(X \mid Z = 0)$; so the expected difference in matched pairs reduces to $E(R_T - R_C \mid Z = 0)$, namely the average effect of the treatment on the treated population. (The inability to find exact matches for some subjects introduces various biases; see Rosenbaum and Rubin (1985) for specifics.)

Because conventional methods of adjustment rely on the assumption that treatment assignment is X-adjustable, we must check this assumption in every way possible. As it turns out, a second control group provides a check on the hypothesis of $(R_C \mid X)$-adjustable assignment, and thereby only a partial check on $(R_T, R_C \mid X)$-adjustable assignment. This distinction is relevant to discussions of the properties of tests of X-adjustable assignment, for example, the properties of consistency and unbiasedness (Section 3.4). It is, therefore, useful for later reference to be clear about what follows from just $(R_C \mid X)$-adjustable assignment. First, (2.4) holds under $(R_C \mid X)$-adjustable assignment, so exact matching yields an unbiased estimate of $E(R_T - R_C \mid Z = 0)$. By an analogous argument, $E(R_T - R_C \mid Z = 0)$ may be estimated by subclassification or model-based adjustments under the same condition. Second, $(R_C \mid X)$-adjustable assignment and the null hypothesis of no effect, H_0: $R_T = R_C$, together imply $(R_T, R_C \mid X)$-adjustable assignment, so that $(R_C \mid X)$-adjustable assignment suffices for tests of this null hypothesis (e.g., Rosenbaum, 1984b). Moreover, if the treatment has an additive effect, $R_T = R_C + \tau$ for fixed τ (cf. Kempthorne, 1952, Section 8), or indeed if it has any deterministic effect—i.e., if $R_T = f(R_C, X)$ for some function $f(\cdot, \cdot)$—then $(R_C \mid X)$-adjustable assignment implies $(R_T, R_C \mid X)$-adjustable assignment. Nonetheless, in general, $(R_C \mid X)$-adjustable assignment is insufficient by itself if we wish to estimate the average treatment effect in the population as a whole, namely $E(R_T - R_C)$, for this requires an adjustment of the responses of both treated and control subjects, rather than adjustments of only the control responses.

2.4 (X, U)-Adjustable Treatment Assignment: A Simple Alternative to X-Adjustable Assignment

A central concern in observational studies is that even after adjustments have been made for the ob-

served covariates, X, treated and control groups may still differ with respect to an unobserved covariate, say U, that is relevant to both treatment assignment and response. To say that U is the unobserved covariate for which adjustments are required is, in effect, to say that treatment assignment is $(R_C \mid X, U)$-adjustable, i.e., that

$$R_C \perp\!\!\!\perp Z \mid (X, U) \tag{2.5a}$$

and

$$0 < \text{pr}(Z = z \mid X = x, U = u) < 1 \quad \text{for each } (z, x, u). \tag{2.5b}$$

Specifically, this says that adjustments for (X, U) would have been sufficient to estimate $E(R_T - R_C \mid Z = 0)$ or to test H_0: $R_T = R_C$, but these adjustments were not feasible because U was not observed. In this paper, the alternatives to the hypothesis of $(R_C \mid X)$-adjustable assignment will be stated in terms of $(R_C \mid X, U)$-adjustable assignment. The closely related condition, $(R_T, R_C \mid X, U)$-adjustable assignment, has been used in connection with sensitivity analyses (Rosenbaum and Rubin, 1983b; Rosenbaum, 1984c, 1986, 1987) and tests of X-adjustable assignment (Rosenbaum, 1984a, Section 3.4).

3. THE ROLE OF A SECOND CONTROL GROUP IN COHORT STUDIES

3.1 Notation: A Second Control Group Is a Second Group of Subjects Who Display Their Responses to the Control

In the case of two control groups, the group to which a subject belongs is indicated by Z, where $Z = 0$ for treated subjects, $Z = 1$ for subjects in the first control group and $Z = 2$ for subjects in the second control group. Only the case of two control groups will be considered, although additional control groups beyond two would not involve new principles. As before, every subject has two potential responses: the response, R_T, that would be observed were the treatment applied to this subject, and the response, R_C, that would be observed from this subject if the treatment were withheld. All subjects in the treated ($Z = 0$) group exhibit their responses (R_T) to treatment, and all subjects in the control ($Z = 1$ or $Z = 2$) groups exhibit their responses (R_C) to control. Adjustable treatment assignment is defined as before; for example, conditions (2.3) and (2.5) continue to define, respectively, $(R_T, R_C \mid X)$-adjustable and $(R_C \mid X, U)$-adjustable treatment assignment. This notation describes cohort studies with multiple control groups; a slightly different approach is required to describe case-control studies (see Section 4).

It is important to observe that, *when treatment assignment is X-adjustable, appropriate estimates of treatment effects may be obtained from either control group, providing adjustments are made for X.* This follows from arguments parallel to those in Section 2.3.

3.2 Using a Second Control Group to Test the Hypothesis That Treatment Assignment Is X-Adjustable

A second control group provides the basis for a test of X-adjustable treatment assignment of the type described in Rosenbaum (1984a). Specifically, $(R_C \mid X)$-adjustable assignment implies that R_C has the same conditional distribution given X in the two control groups ($Z = 1$ and $Z = 2$), or formally that

$$\text{pr}(R_C \le r \mid Z = 1, X) = \text{pr}(R_C \le r \mid Z = 2, X) \tag{3.1}$$

for each r, so any test of (3.1) is a test of X-adjustable treatment assignment. For example, if the response R_C were dichotomous and X were categorical with K possible values, one test of (3.1) would be to apply the Mantel-Haenszel (1959)-Birch (1964) statistic to the $2 \times 2 \times K$ contingency table recording R_C by ($Z = 1$ versus $Z = 2$) by X.

Suppose, now, that treatment assignment is $(R_C \mid X, U)$-adjustable for some unobserved covariate U. It is easy to show (e.g., Rosenbaum, 1984a, Section 3.4) that rejection of the hypothesis (3.1) is also rejection of the hypothesis that U has the same distribution in the two control groups, i.e., rejection of

$$\text{pr}(U \le u \mid Z = 1, X) = \text{pr}(U \le u \mid Z = 2, X) \quad \text{for all } u. \tag{3.2}$$

In other words, rejection of (3.1) implies that, even after adjustment for X, the two control groups are not comparable with respect to U, an unobserved covariate that is related to the response. Furthermore, if (3.2) does not hold, then *at least* one of the two control groups differs from the treated ($Z = 0$) group with respect to the distribution of U—i.e., either

$$\begin{aligned} &\text{pr}(U \le u \mid Z = 0, X) \\ &\qquad \neq \text{pr}(U \le u \mid Z = 1, X) \quad \text{or} \\ &\text{pr}(U \le u \mid Z = 0, X) \\ &\qquad \neq \text{pr}(U \le u \mid Z = 2, X) \end{aligned} \tag{3.3}$$

—so at least one of the control groups is not comparable to the treated group even after adjustment for X.

In short, if, after adjustment for X, the two control groups differ with respect to the response R_C, then treatment assignment is not X-adjustable, and at least one of the control groups is not comparable to the treated group.

3.3 "Control by Systematic Variation"

So far, a second control group plays a negative role: we look for evidence of departures from X-adjustable treatment assignment, and are comforted if we do not find any. To a certain extent, this is the way all scientific theories are checked: a scientific theory is never proved true, for it is not a mathematical proposition; rather, it is exposed to empirical refutation, and confidence in the theory grows as it resists refutation (Popper, 1959). We wish to select control groups so as to provide the severest possible test of X-adjustable treatment assignment. The question, then, is how this is to be done.

In a thoughtful discussion of artifacts and controls in behavioral research, Campbell (1969) quotes Bitterman's (1965) discussion of the principle of "control by systematic variation." Bitterman studied the nature of differences in behavior between species. For example, he wished to conclude that fish are less likely than rats to display some learned behavior not because the fish are less motivated (in this case less hungry), but rather because fish learn more slowly. He writes:

> "I do not, of course, know how to arrange a set of conditions for the fish which will make sensory and motor demands exactly equal to those which are made upon the rat in some experimental situation. Nor do I know how to equate drive level or reward value in the two animals. Fortunately, however, meaningful comparisons still are possible, because for *control by equation* we may substitute what I call *control by systematic variation.* Consider, for example, the hypothesis that the difference between the [learning] curves . . . is due to a difference, not in learning, but in degree of hunger. The hypothesis implies that there is a level of hunger at which the fish will show progressive improvement, and put this way, the hypothesis becomes easy to test. We have only to vary level of hunger widely in different groups of fish, which we know well how to do. If, despite the widest possible variation in hunger, progressive improvement fails to appear in the fish, we may reject the hunger hypothesis. Hypotheses about other variables also may be tested by systematic variation."

There are, then, two principles from the design of experiments for use in observational studies. First, control by equation: compare treated and control subjects who were comparable prior to treatment, perhaps by matching or subclassification on observed covariates (X). But, second, if this is impossible because a possibly relevant covariate (U) was not observed, then apply control by systematic variation: find two control groups in which the distribution of U is quite different, even if unobserved, and check that despite the difference in U, the responses in the two control groups are similar. Even though we cannot measure U and, therefore, cannot ensure that the treated and control groups are comparable with respect to U, we may nonetheless be able to determine whether imbalances in U could explain observed differences in the responses of treated and control groups. (In a factorial experiment, these two principles are combined: treatments are systematically varied and balanced at the same time. There are fewer options in the design of observational studies, and the principles may have to be applied sequentially.)

3.4 A Consistent and Unbiased Test

Recall that a statistical test is *consistent* against a family of alternative hypotheses if the power of the test against each alternative in the family tends to one as the sample size increases. Recall also that a statistical test is *unbiased* against a family of alternatives if the power of the test exceeds its level for each alternative in the family. See Lehmann (1959) for detailed discussion of the formal notions of consistent and unbiased statistical tests.

The test in Section 3.2 of X-adjustable treatment assignment is not generally consistent or unbiased; that is, the test may fail to detect certain violations of X-adjustable treatment assignment no matter how large the sample size or how often the study is replicated. A consistent test would be preferable, for with such a test, failure to detect violations of the hypothesis of X-adjustable treatment assignment in a series of large observational studies on a single topic would begin to suggest that, in those studies, this hypothesis is not far from the truth. This section shows that the principle of "control by systematic variation," as discussed by Bitterman (1965) and Campbell (1969), brings us closer to a consistent test, at least against one broad and interesting class of alternatives.

In this section, treatment assignment is assumed to be $(R_C \mid X, U)$-adjustable for some specific unobserved covariate, U. We wish to ask whether treatment assignment is also $(R_C \mid X)$-adjustable, that is, whether the unobserved U may be ignored in estimating treatment effects. We may certainly ignore U in estimating $E(R_T - R_C \mid Z = 0)$ or in testing H_0: $R_T = R_C$ if

$$R_C \perp\!\!\!\perp U \mid X \tag{3.4}$$

or equivalently if

$$\text{pr}(R_C > r \mid X, U) = \text{pr}(R_C > r \mid X) \quad \text{for each } r.$$

If (3.4) holds, then $(R_C \mid X, U)$-adjustable treatment assignment implies $(R_C \mid X)$-adjustable treatment assignment, by elementary properties of conditional independence or by Dawid's (1979) Lemma 4.

It is often possible to select control groups in such a way as to systematically vary the U distribution. In Example 3, for instance, the untreated control group

consisted of soldiers rejected for treatment on the basis of a medical examination. It is not clear that such an examination, given to young men healthy enough to be accepted for military service, could predict chronic conditions such as cancers decades later. It is clear that there are other untreated groups of soldiers who had to pass special medical examinations before being selected for some special service. Military pilots might be an example. Although it is not known that either of these two untreated control groups is comparable to the treated group, it is known that the two control groups differ from each other in the sense that one was selected and the other rejected on the basis of a medical examination.

When the control groups have been selected to systematically vary the U distribution, a consistent test of (3.4) is possible against the alternative hypothesis that R_C is (strictly) stochastically increasing in U given X; i.e., against the alternative that:

$$\text{pr}(R_C > r \mid X, U = u) \leq \text{pr}(R_C > r \mid X, U = u') \quad \text{if } u < u' \quad \text{for all } r, \tag{3.5a}$$

with strict inequality for some r, i.e., for some r,

$$\text{pr}(R_C > r \mid X, U = u) < \text{pr}(R_C > r \mid X, U = u') \quad \text{if } u < u'. \tag{3.5b}$$

Here, (3.5) states that higher values of R_C are associated with higher values of U even after adjustment for X. Condition (3.5) would hold, for example, if R_C had a linear regression on (X, U) with independent and identically distributed errors, and a strictly positive coefficient for U. Similarly, (3.5) would hold if R_C were dichotomous and followed a linear logit model in (X, U), again with a strictly positive coefficient for U. In other words, (3.5) is a very general (nonparametric) description of positive dependence, known as positive regression dependence (Lehmann, 1966).

Suppose that the control groups have been selected according to the principle of "control by systematic variation," or specifically, suppose that the control groups have been selected so that values of U are typically higher in control groups $Z = 2$ than in group $Z = 1$ at each X; i.e., so that

$$\text{pr}(U > u \mid Z = 1, X = x) < \text{pr}(U > u \mid Z = 2, X = x) \quad \text{for each } u \text{ and } x. \tag{3.6}$$

Condition (3.6) would hold if, for example, the distribution of U given X in the two control groups had the same shape with a higher mean in the $Z = 2$ group. It is often possible to find control groups for which (3.6) is very plausible for some specific U.

To Summarize. (a) Treatment assignment is assumed $(R_C \mid X, U)$-adjustable, and (b) the control groups have been selected to systematically vary U so that (3.6) holds. We seek a consistent test of the null hypothesis of $(R_C \mid X)$-adjustable assignment against the alternative hypothesis that higher R_C values are associated with higher U values even after adjustment for X—i.e., against the alternative (3.5). But, $(R_C \mid X)$-adjustable assignment implies

$$\text{pr}(R_C > r \mid Z = 1, X) = \text{pr}(R_C > r \mid Z = 2, X) \quad \text{for all } r, \tag{3.7}$$

whereas by an elementary argument (a), (b) and (3.5) imply

$$\text{pr}(R_C > r \mid Z = 1, X) \leq \text{pr}(R_C > r \mid Z = 2, X) \quad \text{for all } r; \tag{3.8a}$$

and

$$\text{pr}(R_C > r \mid Z = 1, X) < \text{pr}(R_C < r \mid Z = 2, X) \quad \text{for some } r. \tag{3.8b}$$

In words, if U can be ignored, then the conditional distribution of the response R_C given X is the same in the two control groups, whereas if R_C still increases with U after adjustment for X, then higher values of R_C are expected in control group $Z = 2$ that was selected for its higher values of U. It follows that any test of (3.7) that is consistent (respectively, unbiased) against (3.8) is a consistent (unbiased) test of $(R_C \mid X)$-adjustable assignment against (3.5). For example, if R_C is dichotomous and X is categorical, then the Mantel-Haenszel (1959)-Birch (1964) test is consistent and unbiased, whereas, if R_C is continuous then the stratified Mann-Whitney-Wilcoxon test is consistent and unbiased.

Although it is often possible to select two control groups in such a way as to systematically vary the distribution of a particular U, it will rarely be certain that the control groups differ solely on (X, U) and not on other unobserved covariates as well. Formally, it will rarely be certain that treatment assignment is $(R_C \mid X, U)$-adjustable. In some instances it may be possible to use larger numbers of control groups to study several unobserved differences simultaneously. More commonly, a second control group will address the possibility of bias due to a single unobserved U assuming, in effect, that other unobserved differences are absent. See Section 3.7 for related discussion.

Consistency and unbiasedness are fairly weak properties. Still, we have found that "control by systematic variation" is a sound approach, in that it yields the stronger properties of statistical tests that are not generally available without systematic variation of the U distribution. We routinely evaluate experimental and survey designs in part by considering their formal statistical properties; this principle should apply in the design of observational studies as well. A stronger

version of this principle is discussed in the next section.

3.5 Bracketing

As noted informally by Campbell (1969, page 365), the ideal control groups would bracket the treated group. To make this concept precise, assume that treatment assignment is $(R_C \mid X, U)$-adjustable for some specific unobserved covariate U. Then the control groups bracket the treated group if

$$\begin{aligned} &\text{pr}(U > u \mid Z = 1, X) \\ &\leq \text{pr}(U > u \mid Z = 0, X) \\ &\leq \text{pr}(U > u \mid Z = 2, X) \quad \text{for each } u; \end{aligned} \tag{3.9}$$

i.e., if the control groups have been selected so that, after adjustment for X, the $Z = 1$ control group has lower values of U than the treated group, and the $Z = 2$ control group has higher values of U. In Example 2 of Section 1, with U as the unobserved measure of motivation for academic work, condition (3.6) is very plausible, whereas (3.9) it is not: the students who actually took part in the AP program seem likely to have had higher motivation than either of the other groups, because both of the control groups contain individuals who would have declined participation given the choice.

In this section, it is assumed that, if R_C is associated with U after adjustment for X, then the association is positive; i.e.,

$$\text{pr}(R_C > r \mid X, U = u) \leq \text{pr}(R_C > r \mid X, U = u') \quad \text{if } u < u'. \tag{3.10}$$

Note that (3.10) differs from (3.5) in not requiring *strict* monotonicity; (3.10) permits either (3.4) or (3.5), as well as various intermediate cases.

The assumptions of this section imply:

$$\begin{aligned} \text{pr}(R_C > r \mid Z = 1, X) &\leq \text{pr}(R_C > r \mid Z = 0, X) \\ &\leq \text{pr}(R_C > r \mid Z = 2, X); \end{aligned} \tag{3.11}$$

i.e., after adjustment for X, the population distribution of the (unobserved) control responses for the treated ($Z = 0$) group is bounded by the two distributions of (observed) control responses for the two control groups ($Z = 1$ and $Z = 2$).

Consider the simplest case: a treated subject is randomly sampled, and then exactly matched on the basis of X with two controls, one from each control group. The inequality of (3.11) implies the following.

(i) The difference in responses of the matched treated and $Z = 1$ control subjects overestimates the average effect of the treatment on treated subjects—i.e., it overestimates $E(R_T - R_C \mid Z = 0)$—whereas the difference between the matched treated and $Z = 2$ control subjects underestimates the effect. In other words, when the control groups bracket the treated group, the two matched pair differences estimate bounds on the treatment effect. This is true even though adjustments have been made only for X, and treatment assignment is not X-adjustable.

(ii) Consider a test of the null hypothesis of no treatment effect—i.e., of the hypothesis that $R_T = R_C$ for all subjects—against the one-sided alternative of typically higher responses under the treatment. Here, the matched treated subjects are compared with the matched controls from one or the other of the two control groups. For concreteness, suppose the one-sided McNemar test is used if the response is binary and the one-sided Wilcoxon signed rank test is used if the response is continuous. These tests will not generally have their nominal level because treatment assignment is not assumed to be X-adjustable. Put another way, the matched treated and control subjects may differ systematically with respect to U and, as a result, rejection of the null hypothesis may occur too frequently when the null hypothesis is, in fact, true. Assume the null hypothesis of no effect is true, and let α_1 and α_2 be the actual levels of the one-sided test when applied to the $Z = 1$ and $Z = 2$ controls, respectively. Here, α_1 and α_2 are the true probabilities of rejection for the two control groups using the McNemar or Wilcoxon tests with some fixed nominal level, say .05. In addition, let β be the power of the one-sided McNemar and Wilcoxon test when applied with the same level, .05, to compare the responses in the two control groups as a test of X-adjustable treatment assignment, as in Section 3.2. From (3.11) and the one-to-one matching of treated and control subjects, we have:

$$\beta \geq \alpha_1 \geq .05 \geq \alpha_2. \tag{3.12}$$

In other words, when the control groups bracket the treated group, the one-sided test based on the second control group is conservative, although the test based on the first control group is liberal. Moreover, the chance, β, of detecting the bias due to U by comparing the two control groups exceeds the chance (α_1 or α_2) of falsely rejecting the hypothesis of no treatment effect. This last point is considerably stronger than the consistency/unbiasedness result of Section 3.4.

3.6 The Power of Tests of *X*-Adjustable Assignment in the Absence of Bracketing

In Section 3.5, the power β of the test of X-adjustable treatment assignment was contrasted with the true levels, α_1 and α_2, of the tests comparing the treated group with each of the singly matched control groups. There, due to the bracketing in (3.9) and the positive relationship between R_C and U in (3.10), it

was concluded that $\beta \geq \alpha_1$ and $\beta \geq \alpha_2$. If, instead of (3.9), both control groups had lower distributions of U than the treated group—i.e., if

$$\begin{aligned} &\text{pr}(U > u \mid Z = 0, X) \geq \text{pr}(U > u \mid Z = 1, X) \\ &\qquad\qquad \geq \text{pr}(U > u \mid Z = 2, X) \end{aligned} \tag{3.13}$$

—then we would have $\beta \leq \alpha_2$ and no definite relationship between β and α_1. This is unsatisfactory, for it implies that in the absence of an actual treatment effect, we are more likely to falsely detect a treatment effect than we are to correctly detect bias due to an unobserved covariate U.

In many observational studies, it is feasible to design the study to include more controls than treated subjects, and this could raise β relative to α_1 and α_2. It is, however, important to have some idea of the sample sizes required for reasonable power. For this purpose, consider the following idealized model: (i) zero treatment effect, $R_T = R_C$; (ii) a linear model for R_C,

$$R_C = \alpha + \beta^T X + \gamma U + E;$$

(iii) a homogeneous normal conditional distribution for $(U, E)^T$ in the three groups, $z = 0, 1, 2$,

$$\text{pr}(U, E \mid Z = z, X) = N\left\{\begin{pmatrix} \mu_z \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_U^2 & 0 \\ 0 & \sigma_E^2 \end{bmatrix}\right\};$$

(iv) exact matching on X of K controls from *each* of the $Z = 1$ and $Z = 2$ control groups with each treated subject; and (v) group comparisons based on a paired t test, where the within pair differences are based on mean responses of control subjects in that pair. For a single matched pair or set, the within pair difference between the treated ($Z = 0$) response and average response of the K controls from group $Z = z$ has expectation $\gamma(\mu_0 - \mu_z)$ and variance $\gamma^2(\sigma_U^2 + \sigma_E^2)(1 + 1/K)$. In contrast, a single within pair difference between the two control groups has expectation $\gamma(\mu_1 - \mu_2)$ and variance $\gamma^2(\sigma_U^2 + \sigma_E^2)(2/K)$. It follows easily that we will have $\beta \geq \alpha_1$, α_2 if

$$\left|\frac{\mu_1 - \mu_2}{\mu_0 - \mu_z}\right| \geq \sqrt{\frac{2}{K+1}} \quad \text{for } z = 1, 2. \tag{3.14}$$

In the case of bracketing (Section 3.5), we have $|\mu_1 - \mu_2| > |\mu_0 - \mu_z|$ for $z = 1, 2$, so that the left-hand side of (3.14) is at least 1; then, since $K \geq 1$, the inequality in (3.14) holds so $\beta \geq \alpha_1$, α_2, in general agreement with Section 3.6. In contrast, if (3.13) holds, then $|\mu_1 - \mu_2| \leq |\mu_0 - \mu_2|$, and a K larger than 1 is required for $\beta \geq \alpha_1$, α_2. Table 1 gives values of the bound in (3.14) for various values of K. For example, if $\mu_1 = (\mu_0 + \mu_2)/2$, so the first control group falls half way between the treated group and the second control group, then we need at least $K = 7$ controls from each control group matched to each treated subject to ensure $\beta \geq \alpha_1$, α_2. Clearly in Table 1, the bound falls very slowly as K increases.

TABLE 1

In the absence of bracketing, how many controls are required to test X-adjustable treatment assignment?

No. K of controls from each control group matched to each treated subject	Smallest value $\left(\sqrt{\frac{2}{K+1}}\right)$ of $\frac{\lvert\mu_1 - \mu_2\rvert}{\lvert\mu_0 - \mu_2\rvert}$ for which $\beta \geq \alpha_1, \alpha_2$
1	1.00
2	.82
3	.71
4	.63
5	.58
7	.50 = ½ exactly
10	.43
17	.33 = ⅓ exactly
25	.28
50	.20
100	.14

The power calculations in this section are certainly idealized, and the requirement that $\beta \geq \alpha_1$, α_2 is a limited one. Still, Table 1 serves to emphasize the importance of selecting control groups to systematically vary the U distribution as widely as possible, for otherwise we are unlikely to detect bias due to U even with surprisingly large numbers of controls. Moreover, in the absence of bracketing, it is certainly not unreasonable to seek two control groups that are each an order of magnitude larger than the treated group in an effort to provide a serious test of X-adjustable treatment assignment.

3.7 Partial Comparability

A common argument for using multiple control groups begins with the observation that each of several possible control groups resembles the treated group in some way but not in several other ways. In the simplest case, there are two possible control groups, $Z = 1$ and $Z = 2$, the unobserved covariate $U = (U_1, U_2)$ is two dimensional and (2.5) holds. Then there is partial comparability if

$$\text{pr}(U_1 \leq u \mid Z = 0, X) = \text{pr}(U_1 \leq u \mid Z = 1, X) \quad \text{for all } u, \tag{3.15a}$$

and

$$\text{pr}(U_2 \leq u \mid Z = 0, X) = \text{pr}(U_2 \leq u \mid Z = 2, X) \quad \text{for all } u. \tag{3.15b}$$

In words, control group $Z = 1$ resembles the treated

group with respect to U_1, but not necessarily U_2, and control group $Z = 1$ resembles the treated group with respect to U_2 but not necessarily U_1.

Control groups selected in an effort to achieve partial comparability provide a test of X-adjustable treatment assignment—any second control group does this—but partial comparability by itself does not ensure that the test is a particularly good one. To see this, consider Example 2, the AP program example, where perhaps U_1 is a measure of economic resources available to the student's school, the U_2 is a measure of individual motivation for academic work. Plausibly, the AP students—the treated group—will frequently come from schools with greater resources and will also have high motivation for academic work. Possibly, both U_1 and U_2 are positively related to college performance even after adjustment for recorded sophomore year test scores and course grades. If so, even if partial comparability were achieved with the two control groups discussed in Section 3.1, and even if the AP program had no effect, the AP students are likely to outperform both control groups, because each control group has lower typical values of one coordinate of U. Indeed, similar results for the two control groups might give us a false impression that treatment assignment is X-adjustable. By itself, partial comparability need not yield a more severe test of X-adjustable treatment assignment.

Partial comparability does have one advantage, however. When X-adjustable assignment is rejected, the knowledge that (3.15) holds may aid interpretation. For instance, in discussing the results in Example 1 in Section 1.4, Cole (1979, page 22) bases various conjectures on the knowledge that siblings had similar childhood environments and also similar risks of Hodgkin's disease. It also seem probable that partial comparability could be used to restrict the range of hypotheses about (U_1, U_2) that would be investigated in a sensitivity analysis (Section 1.2), although the mechanics still need to be developed.

4. THE ROLE OF A SECOND "CONTROL" GROUP IN CASE-CONTROL STUDIES

4.1 Case-Noncase Comparisons

A case-control study compares the treatment histories of groups of subjects defined by their responses, that is, a case group and one or more noncase or "control" groups. For example, Doll and Hill (1952) compared the smoking histories of lung cancer patients (the cases) and patients with other diseases in the same hospital (the "control" or noncases). Note that this is quite different from a cohort study in which smokers and nonsmokers are compared with respect to subsequent outcomes. There is, however, a key result due to Cornfield (1951) with some helpful amplification due to Mantel (1973). It states that certain population odds ratios estimated from case-control studies are equal to the corresponding population odds ratios obtained from cohort studies of the same population; this point is reviewed in Section 4.3. Enlightening theoretical discussions of case-control studies are given by Cornfield (1951), Mantel (1973), Hamilton (1979) and Holland and Rubin (1980); the last two references use the two-response (R_T, R_C) notation. Reviews of practical issues in the design of such studies are given by MacMahon and Pugh (1970) and Lilienfeld and Lilienfeld (1980). Statistical methods for case-control studies are surveyed by Mantel and Haenszel (1959) and Breslow and Day (1980).

Case-control studies require slight modifications of the notation of Sections 2 and 3. In most case-control studies, the two responses, (R_T, R_C), are each binary; e.g., 1 for subjects who develop a particular disease and 0 otherwise. Write R^* for the observed binary response from a subject: R^* equals R_T for subjects who actually received the treatment and R^* equals R_C for subjects who actually received the control. Also, write Z for the binary variable indicating whether or not the treatment has been applied, so that, as in Section 2.2, $Z = 0$ for treated subjects and $Z = 1$ for untreated subjects. A case-control study compares the distribution of Z for cases (i.e., for subjects with $R^* = 1$) to the distribution of Z for "controls" or noncases (i.e., for subjects with $R^* = 0$).

In case-control studies, the term "control" is used in a loose, specialized and nonstandard way. In contrast with cohort studies, a "control" group in a case-control study typically contains many subjects who received the treatment. To avoid confusion, alternatives to the phrase "control group" have been proposed by many authors, including "comparison group" and "referent group." Perhaps the simplest and most suggestive term for such a "control" group is a "noncase group," because the group is selected to consist of subjects who are not cases. A case-"control" study is really a case-noncase comparison.

In case-control studies, cases of a particular disease are often obtained from a registry of the cases reported in a particular region, or from all recent patients having the disease in specific hospitals. Typically, the intention is to view cases obtained in this way as representative of all cases in some population, although exactly what population this is can be unclear. Finding noncases or controls who are representative of noncases in the same population is, therefore, often problematic. Strictly speaking, it is only in the "synthetic" case-control studies described by Mantel (1973) that the cases and noncases under study can confidently be viewed as representatives of cases and noncases in a well-defined common population.

4.2 Noncase or "Control" Groups in Case-Control Studies

Often, the noncases actually used are patients with other diseases in the same hospital or registry, or else relatives or neighbors of the cases. In some instances, more than one such group of noncases is used. It is clear that such noncases are not truly representative of noncases in any interesting population; e.g., most noncases of lung cancer are neither hospitalized for another disease nor related to a lung cancer patient. The circumstances under which noncases obtained in this way can form a reasonable comparison group requires clarification.

For this purpose, assume first that, at each value of X, the cases are representative of cases in a specific population, so that the cases provide a direct estimate of $\text{pr}(Z \mid R^* = 1, X)$ in the population. In contrast, assume that: (i) the noncases in the same population are divided into mutually exclusive and exhaustive strata, indicated by a variable S; (ii) the available noncases are drawn from some but not all of the strata; i.e., from some but not all values of S; and (iii) the noncases drawn from a stratum are, at each X, representative of noncases in that stratum, and so they can be used to directly estimate $\text{pr}(Z \mid R^* = 0, S = s, X)$ for some values of s.

For illustration, Table 2 contains selected data from an interesting and thorough study by Hiller, Giacometti and Yuen (1977) of the effects of sunlight on the risk of cataract. The treatment consisted of exposure to more than 3000 hours of sunshine each year ($Z = 0$) as opposed to exposure to less than 2400 hours ($Z = 1$). Here, "exposure" refers simply to living in a region of the United States with these levels of total annual exposure. Cataract cases ($R^* = 1$) were obtained from a registry, the "Model Reporting Area for Blindness Statistics." Noncases were drawn from three strata of noncases in the population: noncases in the same registry having diabetic retinopathy ($R^* = 0, S = 1$), (severe) myopia ($R^* = 0, S = 2$) or optic nerve disease ($R^* = 0, S = 3$). For precise definitions of these strata, and in particular for discussion of the classification of a few noncases with multiple diseases, see Hiller, Giacometti and Yuen (1977). The complete population also contains noncases from another stratum from which no noncases are available, namely the stratum (say $R^* = 0, S = 0$) of all noncases *not* in the registry, possibly because of no eye disease.

The *selection of control groups will be said to be X-adjustable if*

$$Z \perp\!\!\!\perp S \mid (R^* = 0, X), \tag{4.1}$$

that is if, among noncases at each X, the treatment or exposure (Z) is unrelated to the source (S) of the controls. In other words, condition (4.1) says that the treatment/exposure distribution for noncases in the population, namely $\text{pr}(Z \mid R^* = 0, X)$, may be estimated from each of the several sources of noncases, because (4.1) is equivalent to saying that $\text{pr}(Z \mid R^* = 0, X) = \text{pr}(Z \mid R^* = 0, S = s, X)$ for each s.

In the example cited above, the selection of control groups would *not* be X-adjustable if sunlight exposure caused an increase in the risk of the other diseases, namely diabetic retinopathy, myopia and optic nerve disease. Hiller, Giacometti and Yuen (1977, page 57) mention, in effect, that the three noncase groups were selected in the hope of avoiding this possibility. (This issue can be developed formally at the expense of additional notation. In brief outline, S is the observed version of a two version "other disease" indicator (S_T, S_C); cf. Section 2. Then S_T is the other disease status that would be observed from a subject under the treatment (i.e., if $Z = 0$) and S_C is the other disease status that would be observed under the control (i.e., if $Z = 1$). Condition (4.1) would hold if among noncases (i.e., among subjects with $R^* = 0$), (a) treatment assignment were $(S_T, S_C \mid X)$-adjustable and (b) the treatment had no effect on (S_T, S_C), so that $S_T = S_C$).

TABLE 2
Sunlight and cataract: a case group and three noncase groups by age and sex

Age	Males (Z)		Females (Z)	
	0	1	0	1
20–44 ($R^* = 1$)	8	33	6	23
($R^* = 0, S = 1$)	9	96	3	54
($R^* = 0, S = 2$)	11	56	6	29
($R^* = 0, S = 3$)	45	204	26	139
46–64 ($R^* = 1$)	19	139	30	114
($R^* = 0, S = 1$)	18	172	13	222
($R^* = 0, S = 2$)	16	79	7	95
($R^* = 0, S = 3$)	48	226	25	134
65–74 ($R^* = 1$)	33	76	26	99
($R^* = 0, S = 1$)	3	90	13	185
($R^* = 0, S = 2$)	9	36	3	48
($R^* = 0, S = 3$)	11	84	11	49
75+ ($R^* = 1$)	121	172	165	364
($R^* = 0, S = 1$)	2	41	14	123
($R^* = 0, S = 2$)	8	22	5	50
($R^* = 0, S = 3$)	15	70	13	64

Notes: (i) $Z = 0$ if annual sunlight > 3000+ hours; $Z = 1$ if annual sunlight ≤ 2400 hours; (ii) ($R^* = 1$) for cataract cases; ($R^* = 0$, $S = 1$) for noncases with diabetic retinopathy; ($R^* = 0$, $S = 2$) for noncases with myopia; ($R^* = 0$, $S = 3$) for noncases with optic nerve disease. This material is from Hiller, Giacometti and Yuen (1977).

4.3 A Version of Cornfield's Result: Odds Ratios in Case-Control Studies

At a fixed $X = x$, define $\tau(x)$ to be the ratio of the odds of disease under the treatment and under the control, i.e.,

$$\tau(x) = \frac{\text{pr}(R_T = 1 \mid X = x)/\text{pr}(R_T = 0 \mid X = x)}{\text{pr}(R_C = 1 \mid X = x)/\text{pr}(R_C = 0 \mid X = x)}.$$

Thus, $\tau(x)$ is a measure of the effect of the *t*reatment on the subpopulation of subjects with $X = x$. A key result is due to Cornfield (1951) with elaboration and reexpression by Mantel (1973), Hamilton (1979) and Holland and Rubin (1980); it states that, in the absence of certain biases, $\tau(x)$ may be directly estimated in a case-control study. Specifically, let $\epsilon_s(x)$ be the ratio of the odds of *e*xposure to the treatment for cases and for noncases from groups $S = s$,

$$\epsilon_s(x) = \frac{\dfrac{\text{pr}(Z = 0 \mid R^* = 1, X = x)}{\text{pr}(Z = 1 \mid R^* = 1, X = x)}}{\dfrac{\text{pr}(Z = 0 \mid R^* = 0, S = s, X = x)}{\text{pr}(Z = 1 \mid R^* = 0, S = s, X = x)}}.$$

In a case control study, $\epsilon_s(x)$ can be estimated directly from the corresponding empirical odds ratio at $X = x$. Cornfield's (1951) key result states, in effect, that $\tau(x) = \epsilon_s(x)$ *for each s if treatment assignment is* $(R_T, R_C \mid X)$*-adjustable and the selection of control groups is X-adjustable.* In words, under the given conditions we may estimate $\tau(x)$ directly, since it equals $\epsilon_s(x)$. For brevity, the conjunction of the two assumptions, namely (2.3) and (4.1), will be called X-adjustable observation.

The proof of Cornfield's (1951) result is elementary:

$$\text{(4.2a)} \quad \epsilon_s(x) = \frac{\dfrac{\text{pr}(Z = 0 \mid R^* = 1, X = x)}{\text{pr}(Z = 1 \mid R^* = 1, X = x)}}{\dfrac{\text{pr}(Z = 0 \mid R^* = 0, X = x)}{\text{pr}(Z = 1 \mid R^* = 0, X = x)}} \quad \text{[by (4.1)]}$$

$$\text{(4.2b)} \quad = \frac{\dfrac{\text{pr}(R^* = 1 \mid Z = 0, X = x)}{\text{pr}(R^* = 0 \mid Z = 0, X = x)}}{\dfrac{\text{pr}(R^* = 1 \mid Z = 1, X = x)}{\text{pr}(R^* = 0 \mid Z = 1, X = x)}}$$

(by Bayes' Theorem)

$$\text{(4.2c)} \quad = \frac{\dfrac{\text{pr}(R_T = 1 \mid Z = 0, X = x)}{\text{pr}(R_T = 0 \mid Z = 0, X = x)}}{\dfrac{\text{pr}(R_C = 1 \mid Z = 1, X = x)}{\text{pr}(R_C = 0 \mid Z = 1, X = x)}}$$

(by the definition of R^*)

$$\text{(4.2d)} \quad = \tau(x) \quad \text{[by (2.3)]}.$$

The attractive, explicit form of steps (4.2b) to (4.2d) is due to Holland and Rubin (1980); see also Hamilton (1979) and Holland (1986a).

An entirely analogous argument tells us what to expect when one control (i.e., noncase) group is compared to another. Specifically, under X-adjustable observation, we have

$$\text{(4.3)} \quad \omega'_{ss}(x) = \frac{\dfrac{\text{pr}(Z = 1 \mid R^* = 0, S = s, X = x)}{\text{pr}(Z = 0 \mid R^* = 0, S = s, X = x)}}{\dfrac{\text{pr}(Z = 1 \mid R^* = 0, S = s', X = x)}{\text{pr}(Z = 0 \mid R^* = 0, S = s', X = x)}}$$

$$= 1 \text{ for all } s, s' \text{ and all } x.$$

Informally, if adjustments for X suffice to estimate the treatment effect, $\tau(x)$, then the control groups will not differ from one another after adjustment for X, in the sense that $\omega'_{ss}(x) = 1$.

In short, in a case-control study, a comparison of the treatment histories of two or more control groups provides a check of the (conjunction of) the two conditions that permit estimation of the treatment effect.

4.4 An Example

Table 3 contains an example based on the data in Table 2. Table 3 applies the Mantel-Haenszel (1959) test and estimator to compare the cases to each control group, and the control groups to one another. The Mantel-Haenszel estimator provides estimates of the odds ratios, $\epsilon_s(x)$ and $\omega'_{ss}(x)$, assuming they are constants, ϵ_s and ω'_{ss}, not depending on x. The significance levels refer to the null hypotheses that these odds ratios equal one.

Table 3 provides strong evidence that $\epsilon_s > 1$ for each s, so that cataract cases have greater exposures to sunlight than the other groups. However, the table also provides strong evidence that observation is not X-adjustable, so that the argument in (4.2) does not hold, and therefore the estimated ϵ_s cannot safely be taken as an estimate of $\tau(x)$. In particular, two of the control groups (myopia and optic nerve disease) differ less from the cases than they do from the third control group (diabetic retinopathy), sharply contradicting (4.3).

In examining the differences between the control groups, both the statistical significance of the differences and their magnitudes are important. In the current example, the unexplained but statistically significant differences among the control groups are comparable in size to several of the estimates of the effect of sunlight on cataract.

There is one final point that deserves emphasis. To reject the hypothesis of X-adjustable observation, as has been done here, is not to conclude that the treatment has no effect, nor to conclude that the study was

TABLE 3
Comparisons of sunlight exposure among case and noncase groups, adjusting for age and sex: (Mantel-Haenszel estimates of partial odds ratios, with significance levels)

	Myopia controls ($R^* = 0, S = 2$)	Optic nerve disease controls ($R^* = 0, S = 3$)	Cataract Cases ($R^* = 1$)
Diabetic retinopathy control ($R^* = 0, S = 1$)	1.98**	2.58**	4.00**
Myopia controls ($R^* = 0, S = 2$)		1.20	1.98**
Optic nerve disease controls ($R^* = 0, S = 3$)			1.62**
	Tests and measures of departures from X-adjustable observation		Tests and measures of treatment effects assuming X-adjustable observation

Two-sided significance levels: **, P-value ≤ .001; *, .001 < P-value ≤ .05; blank, .05 < P-value.

poorly conducted, nor to conclude that the study's results are uninteresting or undeserving of publication. Rather, to reject X-adjustable observation is to conclude that adjustment for X alone is insufficient to remove bias, and therefore that conventional estimates and significance levels cannot be taken at face value. [See Cole (1979, page 22) for related comments.] Indeed, a good observational study will be designed to permit several tests of the hypothesis of X-adjustable observation; it will therefore be more likely to lead to rejection of this hypothesis than will a poorly designed study that permits no checks at all. The results of tests of X-adjustable observation are simply part of the record of the study's results, intended to aid sober interpretation.

ACKNOWLEDGMENT

This work was supported by Grant SES-87-01890 from the Measurement Methods and Data Resources Program of the National Science Foundation.

REFERENCES

BAYNE-JONES, S., BURDETTE, W., COCHRAN, W., FARBER, E., FIESER, L., FURTH, J., HICKAM, J., LEMAISTRE, C., SCHUMAN, L. and SEEVERS, M. (1964). *Smoking and Health: Report of the Advisory Committee to the Surgeon General.* Van Nostrand, Princeton, N. J.

BIRCH, M. (1964). The detection of partial association, I: The 2 × 2 case. *J. Roy. Statist. Soc. Ser. B* **26** 313–324.

BITTERMAN, M. (1965). Phyletic differences in learning. *Amer. Psychol.* **20** 396–410.

BRESLOW, N. and DAY, N. (1980). *The Analysis of Case Control Studies.* World Health Organization, Lyon, France.

BROSS, I. D. J. (1966). Spurious effects from an extraneous variable. *J. Chronic Dis.* **19** 637–647.

BROSS, I. D. J. (1967). Pertinency of an extraneous variable. *J. Chronic Dis.* **20** 487–495.

CAMPBELL, D. (1969). Prospective: Artifact and control. In *Artifact in Behavioral Research* (R. Rosenthal and R. Rosnow, eds.). Academic, New York.

CAMPBELL, D. T. and STANLEY, J. C. (1963). *Experimental and Quasi-experimental Designs for Research.* Rand McNally, Chicago.

COCHRAN, W. G. (1963). Methodological problems in the study of human populations. *Ann. New York Acad. Sci.* **107** 476–489.

COCHRAN, W. G. (1965). The planning of observational studies of human populations (with discussion). *J. Roy. Statist. Soc. Ser. A* **128** 234–255.

COCHRAN, W. G. (1967). Planning and analysis of nonexperimental studies. In *Proceedings of the Twelfth Conference on the Design of Experiments in Army Research and Testing.* U. S. Army Research Office, Durham, N. C.

COCHRAN, W. G. (1972). Observational studies. In *Statistical Papers in Honor of George W. Snedecor* (T. A. Bancroft, ed.) 71–90. Iowa State Univ. Press, Ames.

COCHRAN, W. G. and RUBIN, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya Ser. A* **35** 417–446.

COLE, P. (1979). The evolving case-control study. *J. Chronic Dis.* **32** 15–27.

COLLABORATIVE GROUP FOR THE STUDY OF STROKE IN YOUNG WOMEN. (1973). Oral contraception and increased risk of cerebral ischemia or thrombosis. *New England J. Med.* **288** 871–878.

CORNFIELD, J. (1951). A method for estimating comparative rates from clinical data. *J. Nat. Cancer Inst.* **11** 1269–1275.

CORNFIELD, J., HAENSZEL, W., HAMMOND, E. C., LILIENFELD, A. M., SHIMKIN, M. B. and WYNDER, E. L. (1959). Smoking and lung cancer: Recent evidence and discussion of some questions. *J. Nat. Cancer Inst.* **22** 173–203.

COX, D. R. (1958). *The Planning of Experiments.* Wiley, New York.

DAWID, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 1–31.

DOLL, R. and HILL, A. B. (1952). A study of the aetiology of carcinoma of the lung. *British Med. J.* 1271–1286.

DOLL, R. and HILL, A. B. (1966). Mortality of British doctors in relation to smoking: Observations on coronary thrombosis. In *Epidemiological Approaches to the Study of Cancer and Other Chronic Diseases* (W. Haenszel, ed.). National Cancer Institute, Bethesda, Md.

FISHER, R. A. (1935). *The Design of Experiments.* Oliver and Boyd, Edinburgh.

GUTENSOHN, N., LI, F. P., JOHNSON, R. E. and COLE, P. (1975). Hodgkin's disease, tonsillectomy and family size. *New England J. Med.* **292** 22–25.

HALSEY, N., MODLIN, J., JABBOUR, J., DUBEY, L., EDDINS, D. and LUDWIG, D. (1980). Risk factors in subacute sclerosing pan-

encephalitis: A case-control study. *Amer. J. Epidemiol.* **111** 415–424.
HAMILTON, M. A. (1979). Choosing a parameter for 2 × 2 table or 2 × 2 × 2 table analysis. *Amer. J. Epidemiol.* **109** 362–375.
HILL, A. B. (1965). The environment and disease: Association or causation? *Proc. Roy. Soc. Med.* **58** 295–300.
HILLER, R., GIACOMETTI, L. and YUEN, K. (1977). Sunlight and cataract: An epidemiologic investigation. *Amer. J. Epidemiol.* **105** 450–459.
HOLLAND, P. W. (1986a). Which comes first, cause or effect? *New York Statist.* **38** 1–6.
HOLLAND, P. W. (1986b). Statistics and causal inference (with discussion). *J. Amer. Statist. Assoc.* **81** 945–970.
HOLLAND, P. W. and RUBIN, D. B. (1980). Causal inference in prospective and retrospective studies. J. Cornfield Memorial Lecture at the American Statistical Association Meetings. Unpublished manuscript.
HOLLAND, P. W. and RUBIN, D. B. (1983). On Lord's paradox. In *Principles of Modern Psychological Measurement: Festschrift for Frederick M. Lord* (H. Wainer and S. Messick, eds.). Lawrence Earlbaum, Hillsdale, N. J.
KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments.* Wiley, New York.
LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.
LEHMANN, E. L. (1966). Some concepts of dependence. *Ann. Math. Statist.* **37** 1137–1153.
LILIENFELD, A. M. and LILIENFELD, D. E. (1980). *Foundations of Epidemiology*, 2nd ed. Oxford Univ. Press, New York.
MACMAHON, B. (1984). Coffee and cancer of the pancreas: A review. In *Coffee and Health* (B. MacMahon and T. Sugimura, eds.) 109–115. Cold Spring Harbor Laboratory, Cold Spring Harbor, N. Y.
MACMAHON, B. and PUGH, T. (1970). *Epidemiology: Principles and Methods.* Little, Brown, Boston.
MANTEL, N. (1973). Synthetic retrospective studies and related topics. *Biometrics* **29** 479–486.
MANTEL, N. and HAENSZEL, W. (1959). Statistical aspects of retrospective studies of disease. *J. Nat. Cancer Inst.* **22** 719–748.
POPPER, K. (1959). *The Logic of Scientific Discovery.* Harper and Row, New York.
ROSENBAUM, P. R. (1984a). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *J. Amer. Statist. Assoc.* **79** 41–48.
ROSENBAUM, P. R. (1984b). Conditional permutation tests and the propensity score in observational studies. *J. Amer. Statist. Assoc.* **79** 565–574.
ROSENBAUM, P. R. (1984c). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. Roy. Statist. Soc. Ser. A* **147** 656–666.
ROSENBAUM, P. R. (1986). Dropping out of high school in the United States: An observational study. *J. Educat. Statist.* **11** 207–224.
ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26.
ROSENBAUM, P. R. and RUBIN, D. B. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.
ROSENBAUM, P. R. and RUBIN, D. B. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. Roy. Statist. Soc. Ser. B* **45** 212–218.
ROSENBAUM, P. R. and RUBIN, D. B. (1985). The bias due to incomplete matching. *Biometrics* **41** 103–116.
RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educat. Psychol.* **66** 688–701.
RUBIN, D. B. (1977). Assignment to treatment group on the basis of a covariate. *J. Educat. Statist.* **2** 1–26.
RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58.
RUBIN, D. B. (1980). Discussion of "Randomization analysis of experimental data: The Fisher randomization test," by D. Basu. *J. Amer. Statist. Assoc.* **75** 591–593.
SOLOMON, R. L. (1949). An extension of the control group design. *Psychol. Bull.* **46** 137–150.
WELCH, B. L. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika* **29** 21–52.
YERUSHALMY, J. and PALMER, C. E. (1959). On the methodology of investigations of etiologic factors in chronic diseases. *J. Chronic Dis.* **10** 27–40.

Comment

Paul W. Holland

I am pleased to see that Rosenbaum's work is included in *Statistical Science*—for two reasons. First, observational studies are very common in scientific work and yet from a theoretical perspective they are poorly understood, often maligned and rarely subjected to serious formal analysis. Rosenbaum's discussion here and elsewhere shows that a formal analysis can lead to useful, practical tools that can help in the design and analysis of nonrandomized studies. Such work ought to be widely publicized and *Statistical Science* is an attractive forum. Second, the particular formal analysis used here by Rosenbaum elaborates and extends the approach I call "Rubin's model" (Holland, 1986a, 1986b) and which I personally feel needs to become wider known and used by mathematical statisticians. My experience over the last 10 years has been that *any* problem involving causal inferences (e.g., inferences about the effects of treatments) is

Paul W. Holland is Distinguished Research Scholar and Director of the Research Statistics Group, Educational Testing Service, 21-T Rosedale Road, Princeton, New Jersey 08541.