Rejoinder

Paul R. Rosenbaum

REPLY TO PAUL HOLLAND

It is always a great pleasure to receive comments from Paul Holland. The reader should understand that it takes a certain amount of effort if Paul and I are to find something to disagree about, but Paul has helpfully raised the issue of SUTVA, and so there is a small difference of language and emphasis worthy of discussion. I will not discuss Paul's intriguing CAI experiment in detail, primarily because I am not sure I absorbed enough about it from his brief description to offer useful comments. Of course, I cannot but agree with the conclusion he reaches at the end of this description in his paragraph 4: we should randomize whenever we can; we can far more often than we do; we should teach every beginning student about the importance of randomization in experiments; nonrandomized controls should be used only when the ethical or practical obstacles to experimentation are overwhelming. I suppose there has been little dispute among statisticians on these points for more than half a century.

Toward the end, Paul discusses what Don Rubin calls SUTVA, the stable unit-treatment value assumption. I would like to make a few general remarks about SUTVA, and then return to Paul's specific comments. This assumption concerns the notation that expresses treatment effects as comparisons of two potential responses for each subject; it says that this notation is adequate for the problem at hand. One might say it is the assumption, or perhaps the indefinite collection of assumptions, implicit in the notation. Don and Paul are certainly correct in emphasizing that virtually any notation carries assumptions buried within it, and they are especially correct in this case, for there are many ways in which the notation can be inadequate, and quite a few are of practical importance. Nonetheless, I do not love SUTVA as a generic label for all of these, for it seems to bear a distinct resemblance to an attic trunk; what does not fit is neatly folded and packed away. The more capacious the trunk, the more likely we are to have difficulty remembering precisely what is packed away. Periodically, we might open the lid and scan the top layer to illustrate what the trunk contains, but because it is so large, we are not inclined to take everything out, to sort the contents into piles: useful in season, useful if altered to fit, damaged beyond repair; still less are we inclined to begin the alterations, for the repair of each garment entails considerable effort. To press the metaphor, I would like to see the trunk opened, the contents sorted, alterations and repairs effected and to see what is beyond repair identified and clearly labeled. In other words, I would like to see SUTVA divided up into a series of more tangible assumptions with practical interpretations, so that violations could be quickly discerned and perhaps addressed. I know that Don and Paul do not want inadequacies of the notation to be forgotten, but I am concerned that the expansive concept SUTVA may tend to have this effect. Let me mention two parts to SUTVA, the second being relevant to Paul's comments.

One violation of SUTVA, "interference between units," is discussed by Cox (1958, Section 2.4) in the context of randomized experimentation. It is possible that the treatment assigned to one unit affects not just that unit, but other units as well. An example from observational studies is passive smoking: whether or not you get lung cancer from cigarette smoke depends not only on the treatment assigned to you—whether or not you smoke—but also on the treatments assigned to others—whether coworkers and family members smoke. In this case, the notation is not adequate: you do not have two potential responses depending solely on whether or not you smoke, but rather a multitude of potential responses depending on whether you smoke and whether those around you smoke. The usual solution to such interference in experiments is to amalgamate small units into larger units that do not interfere with one another; e.g., to replace students by classrooms in educational experiments. This solution works in some but not all observational studies. In studies of macroeconomics or of oligopolistic markets, it is usually impossible to identify two units that do not interfere with one another; in this case, the problem of interference between units must be tackled head on. I am not convinced that adequate methods for this task yet exist, although the interest economists have in simultaneous equation systems and cooperative games is clearly an attempt to grapple with this sort of problem.

A second violation of SUTVA is what Campbell and Stanley (1963) call "history" and I will call "intervening treatments." An intervening treatment is a second treatment, not the treatment of primary interest. The intervening treatment is applied after the primary treatment, and so the intervening treatment is not a covariate, but it is applied before the responses are observed, and so it may affect the responses. This is different from a factorial experiment in that the

primary treatment may affect not only the response but also whether or not you receive the intervening treatment. Again, the notation is not adequate in this case. An adequate notation would involve measurements ($Z, Z_1, Z_0, R_{00}, R_{01}, R_{10}, R_{11}$) for each subject, where (i) R_{ij} is the response a subject would exhibit if the subject received the primary treatment at level i and the intervening treatment at level j, for $i, j = 0, 1$, (ii) Z indicates whether the primary treatment is applied and (iii) Z_i indicates whether the subject would have received the secondary treatment if the primary treatment had been applied at level $Z = i$. If $Z = i$ and $Z_j = j$, then only response R_{ij} is observed from the subject. If the three Z's were determined by the flip of fair coins, it would be straightforward to estimate, say, the average effect of the primary treatment in the absence of the intervening treatment, but without randomization new complications arise. Paul's "flu outbreak" might be thought of as an intervening treatment. To mention this example is to realize that one often has fairly strong beliefs about the relationships among ($Z, Z_1, Z_0, R_{00}, R_{01}, R_{10}, R_{11}$); for instance, one might be willing to believe that computer-aided instruction will have rather slight effects or no effect on whether or not one gets the flu, so one might accept that $Z_1 = Z_0$.

These two examples are intended to show that the notation can be inadequate in important ways that have very little in common. My own view is that the language of statistics should be constructed to use different terms to refer to issues that differ markedly in substance.

In his comments, Paul quite correctly observes that in my use of a single R_C for both control groups I have implicitly defined what I mean by a control group. To say that one has two control groups is say that one has two distinct groups of subjects in which everyone received the same standard treatment, which might be no treatment. Transferring a particular subject from one group to the other would change the composition of the groups, but it would not alter the subject himself; in particular, it would not alter his response. If the two groups are defined by the fact that they received two different standard treatments—so transferring a subject might alter his response—then I would not call them control groups, and the notation I have used would not be adequate. I would not follow Paul in describing this as a violation of SUTVA; rather, I would say I do not have two control groups.

It is possible that one might discover differences in responses in two genuine control groups not because of pretreatment differences, but rather because of the presence of an intervening treatment that is applied at different rates in the two control groups. Presumably, the intervening treatment is also applied to some members of the treated group. Again, I would not describe this as a generic violation of SUTVA; rather, I would say I have two control groups and an intervening treatment.

Paul and I are, I think, in full agreement about the substantive issues here, and the question is simply the best language in which to organize discussion of these issues. My comments simply reflect my discomfort with the "attic trunk" aspect of the SUTVA concept.

REPLY TO BARRY MARGOLIN

I could not agree more with the point Barry Margolin emphasizes, namely that one of the most valuable sources of guidance in observational studies comes from the statistical theory of experimental design. I read his interesting descriptions of uniformity trials and multiple controls in experiments with this in mind. There are two points I would like to raise.

The first point concerns Margolin's observation that in experimental design, uniformity trials are used to check empirically the accuracy and applicability of statistical theory. These are randomized experiments in which it is known that there is no active treatment. This raises the possibility of uniformity trials for observational studies, that is, comparisons of existing populations known to have received identical treatments. Such a comparison would presumably incorporate adjustments for observed covariates, a judicious array of tests of X-adjustable assignment, and sensitivity analyses perhaps as in Rosenbaum (1987). If well-conducted, it seems quite possible that such uniformity trials could help to evaluate existing methods and suggest needed improvements.

The second point concerns the possibility of a positive control in an observational study. I have never seen a positive control *group*, but I have seen a positive control *variable*, that is, a coordinate of the (now multidimensional) responses for which the direction of the treatment effect is known. Again, positive control variables are used to test X-adjustable assignment.

REPLY TO RICHARD CORNELL

Richard Cornell suggests a number of ways in which control by systematic variation might be applied in his current research. The use of multiple groups for comparison seems especially pertinent in his attempts to examine the effects of nickel and chromium as contributors to lung cancer rates among stainless steel workers. His two control groups consist of (i) office workers and (ii) steel workers in a plant that does not use nickel or chromium. These groups were selected specifically to differ markedly in their exposures to other possible carcinogens associated with the production of steel. Because of this control by systematic

variation, a negative finding has a positive interpretation; that is, a finding of a clearly negligible difference between results obtained in the two control groups would go a long way toward eliminating other carcinogens involved in producing steel as an explanation of a difference in lung cancer rates in the treated and control groups. In contrast, had Cornell selected control groups in a thoughtless way—say using office workers and a random sample from the community surrounding the plant—a negative finding would have no positive interpretation; that is, similar results in the two control groups would not eliminate an important ambiguity in the study's conclusions.

Cornell's example does a good job of illustrating the point that one can often pick control groups on the basis of supplementary information about the distributions of unobserved covariates in the groups in such a way as to eliminate one or two key ambiguities. In fact, in this instance, a difference between the control groups also has a clear interpretation, one suggesting that group (ii) is not a control group in the strict sense, but rather a group that has received an active hazardous treatment.

REPLY TO NORMAN BRESLOW

A point requiring clarification concerns the main conclusion of my paper. In his second paragraph, Professor Breslow writes: "... [Rosenbaum's] main conclusion is that inference regarding the validity of a covariable-adjusted relative risk estimated for a particular exposure is strengthened if one can demonstrate equality in the covariable-adjusted odds ratios that contrast the exposures in two or more control groups."

In fact, this paraphrase is quite close to the opposite of what I said, not once, but repeatedly, in the abstract, in the introduction and in the majority of theoretical results. The abstract states: "The value of a second control group depends on the supplementary information that is available about unobserved biases that are suspected to exist ... [and in the absence of such information] a second control group can be of little value." The introduction observes that "little formal work has been done to measure and clarify the severity of tests of the hypothesis that adjustments for X suffice [in estimating treatment effects], and therefore, the degree of corroboration provided by passing such tests has often been unclear." That a second control group may or may not be of value is repeated in Section 1.3 in the paragraph that begins: "The purpose of the current paper ...", concluding "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." The theoretical argument at the heart of the paper is devoted almost entirely to showing that when certain kinds of supplementary information are available, statistical procedures have attractive formal properties, but when such information is absent, the properties need not hold. One such result in Section 3.4 concerns the consistency and unbiasedness of tests of X-adjustable treatment assignment when the control groups have been selected to systematically vary the unobserved covariate U suspected of introducing bias. The systematic variation of U is the essential supplementary information without which the test need not be consistent or unbiased. A second result in Section 3.5 states that the power of certain tests of X-adjustable assignment will exceed the probability of falsely rejecting the null hypothesis of no treatment effect if the control groups have been selected to bracket the treated group. Here the bracketing is the essential supplementary information without which the probability of detecting bias due to U may be smaller than the probability of falsely detecting a treatment effect. Section 3.6 considers the impact of the absence of bracketing in quantitative terms, that is, in terms of the sample size required to achieve $\beta \geq \alpha_1, \alpha_2$ with systematic variation alone. The conclusion there is that an order of magnitude increase in the size of the control groups may well be required to compensate for the absence of bracketing. Section 3.7 shows that the wrong kind of supplementary information—in this case, partial comparability—yields no improvement in the properties of statistical procedures. In several places, I went so far as to suggest that a second control group might even be harmful if it created the false impression that X-adjustable assignment had been exposed to a serious test when, in fact, it had not because the supplementary information on which desirable statistical properties depend was lacking.

This cannot be read as saying that I believe that finding similar results in just any two control groups supports meaningfully the conclusion that adjustments for X suffice to remove bias, or in Professor Breslow's terms that the "validity of a ... relative risk ... is strengthened." In fact, it says the opposite: the supplementary information one has is critical to interpretation when similar results are obtained in two control groups; without such information, a finding of a negligible difference between two control groups has no clear interpretation. To put this another way, it is clear from the results in the paper that tests of X-adjustable assignment based on two control groups need not, in general, have satisfactory properties—e.g., the tests need not be consistent. Finding a negligible difference between the results obtained in two control groups provides support for X-adjustable treatment assignment—that is, it corroborates X-adjustable treatment assignment in Popper's sense—*only if* the comparison of the two control groups had

a reasonable prospect of exhibiting a difference if X-adjustable assignment were false and some specific scientifically plausible alternative were true instead.

Professor Breslow says that the principles reviewed and developed in the paper are "well-known and widely used" in biostatistics and epidemiology. It appears that there is some question as to whether we are referring to the same principles, but if we are, then, in my experience, he is wrong about this. There are three issues. First, it is certainly not true in any discipline that observational study designs are routinely evaluated in terms of the formal properties—e.g., consistency, unbiasedness, power—of statistical tests for hidden biases, that is, tests of X-adjustable assignment. Second, the two existing principles for selecting control groups that were found here to yield attractive formal properties of tests for hidden biases—namely, the principles of systematic variation and bracketing—are due to social scientists, especially Donald Campbell, and to my knowledge do not appear in the epidemiological/biostatistical literature.

Third, there is the question of actual current practice. Five epidemiological studies using multiple control groups were cited in my paper. In all five, formal statistical procedures were used to compare the treated group and the control groups, but not one used formal statistics to compare the control groups to one another. In Table 3, we saw what happens in one of the studies when the control groups are compared. The excellent text by Lilienfeld and Lilienfeld (1980) presents detailed data from the Collaborative Group for the Study of Stroke in Young Women (1973) as an exercise for the student, but the data are tabulated in such a way as to permit comparisons only between the treated and control groups, and not comparisons among the control groups. The practice of using the same methods for all comparisons—among control groups and between treated and control groups—is not a standard part of the practice of the literature cited, a literature that includes the work of such leading epidemiologists as Brian MacMahon and the late Abraham Lilienfeld. If we use sensitive statisical procedures to look for treatment effects but not for hidden biases, we will find effects and not biases, even when the biases are real and the effects are not.