Again, the SUTVA is implicit in Rosenbaum's analysis and its violation might make the application of his results ineffective. For his medical examples, the SUTVA is not satisfied if the model only specifies a single control response, R_C rather than an R_{C_i} for the ith control group and, in fact, subjects in the first control group are exposed to a different kind of treatment than are those in the second control group. On the other hand, Rosenbaum shows that when such an assumption holds there is a clear benefit for the design and analysis of observational studies.

Because Rosenbaum focuses on the problem of assessing biases due to pretreatment differences, it is a little unfair to ask him to solve these other problems as well, but from the excellence of the present paper I am sure he is up to the task.

ADDITIONAL REFERENCES

HOLLAND, P. W., JAMISON, D. and RAGOSTA, M. (1979). *Computer-assisted Instruction and Compensatory Education: The ETS/LAUSD Study. Data Analysis: Fiscal year 1978.* Educational Testing Service, Princeton, N. J.

RAGOSTA, M., HOLLAND, P. W. and JAMISON, D. (1982). *Computer-assisted Instruction and Compensatory Education: The ETS/LAUSD Study. Final Report.* Educational Testing Service, Princeton, N. J.

RUBIN, D. B. (1986). What ifs have causal answers, discussion of "Statistics and causal inference," by P. Holland. *J. Amer. Statist. Assoc.* **81** 961–962.

Comment: The Use of Multiple Control Groups in Designed Experiments

Barry H. Margolin

I want to commend Dr. Rosenbaum on a most lucid presentation regarding the role of a second control group in an observational study. My contribution to this discussion is motivated by a remark of Cochran (1965), that in discussions of topics in observational studies, "it is relevant to indicate how the problem is tackled in controlled experimentation . . . ." With this in mind, I will discuss briefly and illustrate the roles that multiple control groups have played in designed experiments; these illustrations draw heavily upon my own research simply because they are readily accessible to me. Three distinct roles are discernible: (i) to detect the presence of unsuspected systematic effects; (ii) to determine whether there are hidden sources of extraneous random variability; and (iii) to assemble sufficient control data to permit a meaningful assessment of sampling model assumptions.

The earliest experiments whose design included features resembling multiple control groups appear to be uniformity trials in agricultural research (Cochran, 1937). These are agricultural experiments in which the land is divided into a number of plots of the same size. A single variety of the crop of interest is planted, although other factors such as fertilizer are kept constant from plot to plot, and the yield of each plot is observed. As Cochran (1937) noted, the primary purpose of a uniformity trial is to study the effects of amalgamation of the original plots into "larger plots of various sizes and shapes" and "to provide information on the optimum size and shape of plot" with regard to experimental error. As such, a uniformity trial is viewable as an experiment with a control group but no treatment group.

The analogy with multiple control groups becomes clearer, however, when one observes that uniformity trials are also conducted to validate the applicability of tests of significance that are based on analysis of variance (ANOVA). As Cochran (1937) writes,

> "A preliminary requirement for the application of the analysis of variance to be possible is that the experimental design used should be chosen at random from a set of designs such that, in the absence of any treatment effect, the average treatment mean square over the set should equal the average error mean square. . . . The further question arises: how good an approximation to the tabulated z distribution is generated by the process of randomization used? There again the question may be tested from uniformity trial data."

When any particular design is imposed on the uniformity trial data, which involve no true treatments, the results hopefully appear as if they derived from a

Barry H. Margolin is Head, Statistical Methodology Section, National Institute of Environmental Health Sciences, P. O. Box 12233, Research Triangle Park, North Carolina 27709.

designed experiment with k homogeneous groups, i.e., k control groups.

I have commented elsewhere (Margolin, 1985) that the concepts behind uniformity trials suggest the desirability of running laboratory experiments in which the factor of interest is omitted, but all other features of the experimental protocol are retained. For example, in experiments to study the genotoxic effects of a suspect chemical, it is common to design an experiment to include a control group and $k - 1$ treatment groups suitably spaced on the dosage scale. If the test chemical is replaced with distilled water, the "treatment" is one that is known not to induce genotoxicity. If, in addition, the individuals conducting the experiment are blinded to the identity of this "treatment," the result is an experiment with k distinct control groups. The experimental results obtained can then be used to fulfill the three roles for multiple control groups listed earlier.

As an illustration, Table 1 contains the results of four sister chromatid exchange (SCE) *in vitro* assays with distilled water tested under code. Each entry is associated with a unique flask and represents the average SCE count from 50 Chinese hamster ovary cells after their exposure to the distilled water treatment. The four experiments can be viewed as consisting of five control groups differentiated only by their temporal order of creation. These data were used by Margolin, Resnick, Rimpo, Archer, Galloway, Bloom and Zeiger (1986) to check whether the homogeneous Poisson model, which they documented for SCE readings from cells within a common control flask, extended to multiple control flasks created within a given day. Using the data in Table 1, the authors concluded that (i) there was neither evidence of systematic differences nor an indication of hidden sources of random variability among the flasks within an experiment, and (ii) that the homogeneous Poisson model is a good approximation to the sampling behavior of these data.

TABLE 1
Average SCE/cell data for four coded experiments with distilled water

Dose (μg/ml)	No S-9 experiments[a]		With S-9 experiments[a]	
	1	2	1	2
0	8.64	8.12	9.18	8.34
0.16×10^3	7.72	—	9.30	—
0.50×10^3	8.48	—	8.82	—
0.16×10^4	7.94	8.14	10.08	9.04
0.30×10^4	—	8.32	—	7.62
0.40×10^4	—	8.74	—	8.04
0.50×10^4	8.86	8.48	9.30	8.06
Average SCE count in experiment	8.33	8.36	9.34	8.22

[a] 50 cells scored for each dose per experiment.

The use of two control groups within an experiment involving a true treatment of interest is illustrated in studies by Probst reported in Ashby, de Serres, Draper, Ishidate, Margolin, Matter and Shelby (1985). In each of 14 Ames test experiments with chemicals of interest, Probst included two control groups of three plates each, one at the beginning of plating and one at the end. Table 2 contains the "before and after" control group averages, rounded to the nearest integer. Thirteen of the fourteen differences in Table 1 are nonnegative, indicating that a systematically decreasing effect, albeit small, is operative from the beginning to the end of the experiments.

As a final point in the discussion, I wish to mention the inclusion in laboratory studies of what is commonly referred to as a "positive control group," i.e., the creation of a treatment group whose treatment is known to be effective in the protocol being used. If this known effect is not detected concurrently in the course of the experiment with the factors under study, all findings of no effect with regard to these study factors are discounted as suspect and the experiment is repeated. This leads to an intriguing question: Within an observational study, what is the analogy to a positive control group and how should it be used?

ADDITIONAL REFERENCES

ASHBY, J., DE SERRES, F. J., DRAPER, M., ISHIDATE, M., JR., MARGOLIN, B. H., MATTER, B. E. and SHELBY, M. D., eds.

TABLE 2
Control group averages from a series of fourteen Ames test experiments[a] each with one control group at the start and one at the end of experimentation

	Experiments													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Start	27	25	18	32	21	19	22	19	27	25	28	17	23	19
End	19	21	18	29	18	21	16	17	20	22	25	16	19	18
Difference	8	4	0	3	3	−2	6	2	7	3	3	1	4	1

[a] Ames test with TA1535, no S9; data are the number of visible colonies averaged over three replicates and rounded to the nearest integer, (Ashby, de Serres, Draper, Ishidate, Margolin, Matter and Shelby, 1985, page 164).

(1985). *Evaluation of Short-Term Tests for Carcinogens: Report of the International Programme on Chemical Safety Collaborative Study on In Vitro Assays.* North Holland, Amsterdam.

COCHRAN, W. G. (1937). Catalogue of uniformity trial data. *J. Roy. Statist. Soc. Suppl.* **4** 233–253.

MARGOLIN, B. H. (1985). Statistical studies in genetic toxicology: A perspective from the U. S. National Toxicology Program. *Environ. Health Perspectives* **63** 187–194.

MARGOLIN, B. H., RESNICK, M. A., RIMPO, J. Y., ARCHER, P., GALLOWAY, S. M., BLOOM, A. D. and ZEIGER, E. (1986). Statistical analyses for in vitro cytogenetic assays using Chinese hamster ovary cells. *Environ. Mutagen.* **8** 183–204.

Comment

Richard G. Cornell

Rosenbaum's review provides a logical framework for thinking about the value of more than one control group in observational studies and contains useful discussion of the implications for the design and interpretation of investigations in which randomization to control and treatment groups is not feasible.

I found his discussion of case-control studies particularly interesting. Controls in these studies are "noncases" selected for comparison with a group of "cases" known to have a particular disease or other condition. Controls and cases are compared with respect to the extent of exposure to potential causative agents or with respect to other background variables. Rosenbaum emphasizes that a comparison of the histories of two or more control groups provides a check of the assumptions that underlie the estimation of the effect of exposure after covariance adjustment. The example he presents on the extent of exposure to sunlight of cataract cases and controls involves three control groups with other eye conditions. The use of multiple control groups enabled him to conclude that adjustment for age and sex is not sufficient for unbiased estimation of the effect of sunlight exposure on the prevalence of cataracts.

This example serves as a prototype for the interpretation of other case-control studies with more than one control group and a warning of possible undetected bias in case-control studies with only one control group. More importantly, it serves as a reminder that it is best to select more than one control group when the ideal control group cannot be formed through randomization. This allows a check on assumptions that cannot be attained through randomization and yet are crucial to conclusions on the effect of exposure after taking covariates into account.

Rosenbaum also refers to an example of a study in which the groups to be compared are exposure groups and the outcome is the prevalence of coronary thrombosis, with subclasses within each exposure group defined by covariates. Observational studies of this type are common in epidemiology and medicine. One example is a study, described by Cornell (1984), of cancer rates relative to exposure to the environment in steel plants, which produce stainless steel. One purpose of this study was to see if there is evidence of an increase in lung cancer rate attributable to the nickel and chromium used in stainless steel production. Exposure groups were formed by area worked within a plant. Comparisons were made after adjusting for age.

Another example is the comparison of survival rates for burn victims using registry data grouped by hospital, and subsequently by the speed of wound closure attained by the burn care practice in a hospital. Again, age is an important covariate for comparisons of burn survival. So is burn severity as measured by the extent of full thickness burn. A model that takes these variables as well as other demographic and severity variables into account for purposes of estimation, prediction and evaluation is discussed by Wolfe, Roi, Flora, Feller and Cornell (1983) and presented in detail by Cornell, Flora and Roi (1983).

These examples are typical of observational studies in epidemiology and medicine in that morbidity or mortality rates are compared between groups formed by exposure categories or type of treatment. Comparisons are made within categories defined by covariates or after adjustment for covariates. Common types of covariates are demographic variables, such as age and sex, and initial severity measures.

Rosenbaum gives guidance with respect to the design and analysis of such studies. He says that it is desirable to select two control groups in such a way that a possibly relevant, but unobserved, covariate has different distributions in the two control groups, and then to check to see if the responses in the two control groups are similar. If they are, then the unobserved

Richard G. Cornell is Professor of Biostastistics, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109.