Drug Information Journal, Vol. 22, pp. 343-352, 1988
Printed in the USA.

0092-8615/88/$3.00 + .00

THE EFFECT OF STUDY DESIGN ON GAIN IN EVALUATIONS OF NEW TREATMENTS IN MEDICINE AND SURGERY*

GRAHAM A. COLDITZ, MB, BS, DRPH

Channing Laboratory, Department of Medicine, Harvard Medical School and Brigham and Women's Hospital, Boston, Massachusetts

JAMES N. MILLER, MPP

Center for Science and International Affairs, John F. Kennedy School of Government, Harvard University, Cambridge, Massachusetts

FREDERICK MOSTELLER, PHD

Department of Health Policy and Management, Harvard School of Public Health, Boston, Massachusetts

We analyzed the results of 128 comparisons of an innovation with a standard treatment in medicine, and 221 comparisons in surgery, to relate features of study design to the magnitude of gain. The mean gain (measured by the Mann-Whitney statistic) for the innovation over standard therapy was relatively constant across study designs, except for nonrandom trials with sequential assignment to therapy. These trials showed a significantly higher likelihood that a patient would do better on the innovation than on standard therapy.

In medical trials that used sequential assignment of refractory patients, the mean gain (measured as the Mann-Whitney statistic) was 0.94, compared to the mean gain for randomized controlled trials of 0.62 ($p < .01$). In surgery, the mean Mann-Whitney statistic for nonrandom sequential studies evaluating primary treatments was 0.78, compared to 0.56 for randomized controlled trials ($p < .01$). In the evaluation of both medical and surgical therapies, randomized control trials that used a placebo control were significantly more likely to produce a gain for the innovation. In medicine, the Mann-Whitney statistic was 0.72 for placebo controlled trials and 0.61 for nonplacebo controlled trials ($p = 0.04$). In surgery, the average Mann-Whitney statistic was 0.60 for 12 double-blind trials that used placebos, and 0.52 for 10 double-blind trials that did not. When interpreting an evaluation of a new therapy, we may consider adjusting the results of studies that have used sequential assignment so that the average bias, as reported in this article, may be taken into account. Likewise, in studies that have used a placebo control, the Mann-Whitney statistic could also be adjusted downward by .10 if a standard treatment is available. These adjustments provide a more realistic appraisal of the new treatment until stronger studies supersede them.

Key Words: Comparative studies; Clinical studies; Observational studies; Randomized controlled trials; Meta-analysis; Research synthesis; Bias; Blinding; Placebo control

---

Presented at the Drug Information Association Workshop "Cost Benefit Analysis of Pharmaceutical Products," March 14 to 16, 1988, Hilton Head Island, South Carolina.

*Supported by the National Center for Health Services Research and Health Care Technology Assessment Grant No. 1 R01 HS 05156-01, the Macy Foundation and the Rockefeller Foundation.

Reprint address: Graham A. Colditz, MB, BS, DrPH, Channing Laboratory, Department of Medicine, Harvard Medical School, Boston, MA 02115.

ALTHOUGH randomized controlled trials are usually preferred for the evaluation of efficacy of innovations in therapy, often this kind of study is not available. For many reasons, we cannot expect all innovations to be evaluated in this manner, and so evaluators have increasingly been asked to consider and supply alternative methods. Many of these have been described in the Institute of Medicine's book *Assessing Medical Technologies* (1). This article makes a contribution to the comparison of methods by examining the gains for innovations evaluated by different study designs, especially comparative clinical studies, and considering their implications for the assessment of the innovation.

Previous investigations have suggested that when a new treatment in medicine or surgery is compared to standard practice, the average gain attributed to the innovation is greater in nonrandomized studies than in randomized studies (2–6). (Although not all investigations have shown such a relationship (7), this is the typical result.) In light of such evidence, how should the results from nonrandomized studies be interpreted and combined with results from randomized trials?

First, all studies, random and nonrandom, could simply be combined with equal weight. Given the previously cited evidence, however, this does not seem appropriate. Second, results from nonrandomized studies might simply be ignored. This seems to be a vast waste of data, particularly when little or no data based on randomized designs are available.

Third, the randomized controlled trials might be used for quantitative estimates, and the nonrandomized trials employed only for more qualitative judgments, or to test for consistency with randomized trials. Although this would be less of a waste than neglecting nonrandomized trials altogether, it would by no means be an efficient use of data. If few or no randomized studies have been carried out on a new treatment, this approach could require abandoning attempts to obtain a quantitative estimate of gain.

Fourth, less weight could be given to results from nonrandomized studies. This approach may seem reasonable at first; however, to the extent that nonrandomized studies were biased in favor of the innovation, adding in their estimates without correction—even at reduced weights—would still bias the overall estimate.

In this study, we pursued a fifth alternative: adjusting the results from nonrandomized studies to account for their average bias, so the data they provide can be used without adversely affecting the accuracy of estimates. To provide a basis for making such adjustments, we investigated the average gains found by a large number of randomized and nonrandomized evaluations in medicine and surgery.

MATERIALS AND METHODS

Data Extraction

We selected, by methods described in detail elsewhere (8,9), a number of medical journals with publication dates in 1980 and surgery journals with publication dates in 1983. This article synthesizes and compares these previous analyses. For medicine, five journals were selected from each of four disciplines to represent a range of therapies being evaluated and reported in the literature. The disciplines were cardiology, neurology, psychiatry, and respiratory medicine. For surgery, six general surgical journals were used.

To be eligible, an article had to report on a study that had not been previously published. The study was required to compare the efficacy of two or more treatments on 10 or more human patients (5 for cross-over studies), with response to therapy as the outcome measure. Two people, after several hours of training, read each article to decide whether it qualified for inclusion. Articles were included for further review if either reader accepted

it. A log of rejected articles was maintained.

Each article selected was then independently read by two people with training in statistical methods. These readers first underwent a training program to ensure uniform understanding and application of the definitions for data they extracted. These readers identified the innovation and the standard, defined the study type, and extracted the gains reported from each study (see below). Any differences were independently adjudicated and resolved by a third reader.

We measure the gain attributed to the innovation over the standard therapy differently for the medicine and surgery studies. In medicine, the Mann–Whitney statistic was used. The Mann–Whitney statistic estimates the probability that a randomly selected patient will perform better given the innovation than a randomly selected patient given the standard treatment. It can be computed from many different statistical measures that are often reported in comparisons of medical treatments, eg, proportion surviving, mean change in blood pressure (and standard deviation), and frequency of side effects. Readers were asked to select, when possible, a single endpoint for each study. If several endpoints were judged of equal importance, the average of the Mann–Whitney statistics based on each of these endpoints was taken as the average gain for the study.

For surgery, gain within a study was measured by the difference in the proportion of each treatment group considered to be successes (positive values favoring the innovation). To compare findings of the medicine and surgery analyses, results from surgery were also translated into an equivalent Mann–Whitney statistic, using the formula: Mann–Whitney statistic equals $1/2+D/2$, where D represents the difference in proportions.

For both medicine and surgery, we also undertook an analysis based on a scoring of authors' qualitative conclusions, employing a scale devised by Gilbert et al (5). Because those results closely paralleled those based on the quantitative measure of gain, they are not reported here.

The articles surveyed sometimes reported on more than one comparison of an innovation to a standard. In our analysis, each comparison of an innovation to a standard was assigned equal weight. For example, if an article reported on two separate studies, each with one comparison of an innovation to a standard therapy, this article would contribute two comparisons to the total. If a study compared three innovations with one standard therapy, then that study would contribute three comparisons to the analysis. We also carried out an analysis giving each article equal weight; because we found essentially no change in results, that work is not reported here.

We used the following classification into six study types, based on a scheme devised by Bailar et al (10): randomized controlled trial with parallel control groups; randomized controlled trial with sequential control (cross-over); nonrandomized controlled trial with parallel control groups; nonrandomized controlled trial with sequential control; externally controlled trial; and observational study.

We found two kinds of nonrandomized controlled trials with sequential control: those that used a formal, but nonrandom, cross-over procedure for patients; and pre/post comparisons, where patients were simply assessed before and after receiving treatment with the innovation. Many pre/post comparisons were based on patients who had been refractory to treatment with the standard therapy.

In externally controlled trials, investigators compare the results of their preplanned (and "single-armed") study of the innovation to other results, outside their trial, using the standard therapy. Data for the external control group may be taken from the literature or from unpublished results from the same or another institution.

Observational studies use retrospective record reviews, often including additional follow-up of patients. Such evaluations, since they are based exclusively on historical reviews of patient outcomes, are unlike all of the other study types considered, which are preplanned for at least one of the treatments considered.

Along with study type as defined above, we also considered blindness of patients to treatment received, blind assessment of outcome, and use of placebo.

Analysis

We compared average gains within each study type, using the Mann–Whitney statistic for medicine, and the difference in proportions of treatment successes for surgery.

To sharpen comparisons of average gains in surgery, we stratified all studies by whether primary or secondary treatments were evaluated. Primary treatments are "intended to cure or ameliorate the patient's primary disease," while secondary treatments are "improvements intended to prevent or treat such complications as infection or thrombo-embolic disease, or improvements in anesthesia or postoperative care." (5)

The p-values provided in our analysis correspond to two-sided tests of the null hypothesis that the average gain found by different study types is the same, using the normal approximation. In computing p-values, no adjustment is made for multiple tests.

RESULTS

The 113 medicine articles that qualified for inclusion provided 128 comparisons of treatments; 188 surgery articles provided 221 comparisons.

Gains for Medicine

Study Types. Table 1 presents the average gains by study type for medicine. For randomized controlled trials using parallel controls, the mean of the Mann–Whitney statistic was 0.61, while 0.50 would be a neutral result. Thus, averaging over all randomized controlled studies included, there was a 61% chance that a random patient receiving the innovation would do better than a random patient on standard therapy.

The average Mann–Whitney statistic was relatively constant for all but one study design, the nonrandomized controlled trials using sequential assignment to therapy. This study design showed a substantial increase in the likelihood that a patient would do better on the innovation than on the standard therapy according to the Mann–Whitney statistic (Table 1). Within these nonrandom sequential trials, patients admitted to a study after failing on standard therapy (ie, patients refractory to standard therapy) had a very high probability of improving on an innovative therapy. The average Mann–Whitney statistic for these trials with refractory patients was 0.94, whereas for trials with patients who were not refractory, it was 0.76. Thus, even after removing the trials with refractory patients, nonrandom sequential studies had a significantly higher likelihood of patients succeeding on the innovation than did randomized controlled trials ($p=0.004$).

Blinding. The relationship between blinding and the size of gain was evaluated within the randomized controlled trials by classifying studies as double-blind, and non-double-blind (Table 2). The average Mann–Whitney statistic for double-blind trials was 0.58, and 0.69 for non-double-blind studies ($p=0.02$).

Placebo control. The innovation in therapy was compared to placebo in 15 of the 65 randomized controlled trials. The likelihood of performing better on the innovation was significantly higher when the comparison was to placebo therapy. The mean Mann–Whitney statistic was 0.72

TABLE 1
Mann–Whitney Statistic Among a Sample of Evaluations of Medical Therapy Reported in 1980

Study design	Number of studies	Mann–Whitney statistic* Mean	SD†
Randomized controlled trials (parallel)	36	0.61	0.14
Randomized controlled trials (cross-over)	29	0.63	0.14
Nonrandom parallel comparisons	3	0.56	0.07
Nonrandom sequential comparisons	46	0.81	0.15
refractory	12	0.94	0.09
other	34	0.76	0.16
External controls	5	0.65	0.10
Observational studies	9	0.57	0.04

*Estimated probability of a random patient performing better on the innovation than a random patient on the standard therapy; a null result is 0.50.
†Standard deviation of the individual measurements, not the standard deviation of the mean.

for placebo controlled trials and 0.61 for nonplacebo trials ($p=0.04$).

Gains for Surgery

Study types. For primary treatments, the average difference in the proportions of treatment successes for randomized controlled trials was 11.9%, smaller than for all other study types (Table 3). Differences in gains between randomized controlled trials and other study types were generally not statistically significant. The power of these comparisons was limited by the small number of studies involved.

The results for nonrandomized sequential comparisons, however, with an average difference of 56.5%, were significantly different from randomized controlled trials ($p<0.0001$). Due to the small number of nonrandomized sequential comparisons (six), we did not differentiate between nonrandomized sequential comparisons with refractory and nonrefractory patients.

Overall, results for secondary treatments were similar to those for primary treatments: the average difference in the proportions of treatment successes for randomized controlled trials (6.0%) was

TABLE 2
The Use of Blinding in the Evaluation of Medical Therapies, by Study Design

	Number of studies	Double-blind	Non-double-blind
Randomized controlled trials (parallel)	36	21	15
Randomized controlled trials (cross-over)	29	13	16
Nonrandom parallel comparisons	3	0	3
Nonrandom sequential comparisons	46	1	45
External controls	5	NA*	9
Observational studies	9	NA*	5

*Not applicable.
Note: A further categorization shows that 11 randomized controlled trials (cross-over) used a single-blind design as did 2 nonrandom sequential comparisons.

TABLE 3
Gains for Primary Treatments Among a Sample of Evaluations of Surgical Treatments Reported in 1983 (Difference in Percentage of Treatment Successes and Associated Mann–Whitney Statistic)

	Number of comparisons	Difference in percentage of treatment successes Mean	SD†	Equivalent Mann–Whitney statistic* Mean
Randomized controlled trials (parallel)	20	11.9%	15.6%	0.56
Randomized controlled trials (cross-over)	0	—	—	—
Nonrandom parallel comparisons	4	24.8%	24.8%	0.62
Nonrandom sequential comparisons	6	56.5%	15.4%	0.78
External controls	19	25.6%	24.4%	0.63
Observational studies	73	14.5%	29.0%	0.57

*Estimated probability of a random patient performing better on the innovation than a random patient on the standard therapy.
†Standard deviation of the individual measurements, not the standard deviation of the mean.

smaller than for all other study types (Table 4). The largest difference in the proportion of treatment successes for the innovation over the standard therapy, 80.0%, was produced by a nonrandomized sequential study. Nonrandomized controlled trials had the next largest gain, 10.3%. This average gain was not significantly larger than the average for randomized controlled trials of 6.3% (p=0.41).

The average difference in the proportions of treatment successes was larger for

TABLE 4
Gains for Secondary Treatments Among a Sample of Evaluations of Surgical Treatments Reported in 1983 (Difference in Percentage of Treatment Successes and Associated Mann–Whitney Statistic)

	Number of comparisons	Difference in percentage of treatment successes Mean	SD†	Equivalent Mann–Whitney statistic* Mean
Randomized controlled trials (parallel)	61	6.0%	16.4%	0.53
Randomized controlled trials (cross-over)	0	—	—	—
Nonrandom parallel comparisons	11	10.3%	15.6%	0.56
Nonrandom sequential comparisons	1	80.0%	—	0.90
External controls	8	7.6%	9.6%	0.54
Observational studies	18	9.4%	16.1%	0.55

*Estimated probability of a random patient performing better on the innovation than a random patient on the standard therapy.
†Standard deviation of the individual measurements, not the standard deviation of the mean.

primary than for secondary treatments for all but two study types. (The two exceptions each had only one evaluation of a secondary treatment.) The average difference in the proportions of treatment successes was also less closely grouped around the gains for randomized controlled trials for primary than for secondary treatments. To show this quantitatively, the average absolute difference between the average gains for randomized controlled trials and those for other study types was 8.8% for primary treatments, and 2.9% for secondary treatments (excluding the two study types that each contributed only one comparison).

Blinding. The relation between blinding and gain was evaluated within randomized controlled trials. About half of all randomized controlled trials of secondary treatments (32 of 61) had either patients or assessors, or both, "blind" to the treatment received (Table 5). Contrary to the hypothesis that studies with weaker design tend to find larger gains, comparisons that were "double-blind" produced the largest average gains for secondary therapies, significantly larger than the average for comparisons that involved no blinding ($p=0.032$). The small number of comparisons precluded further analysis of blinding within the 20 primary trials.

Placebo control. One possible source of the disparity between the average gains found by double-blind and nonblind randomized controlled trials may be that placebos are often used in double-blind trials (and by definition cannot be used in studies where no blinding was used). We explored this possible effect of placebos and observed that the average difference in the proportion of treatment successes for 12 double-blind trials that used a placebo was 19.0%, compared to 4.7% for 10 double-blind trials that did not use a placebo. Another possible factor that we considered was whether a drug or other treatment ("nondrug") was evaluated. For both double-blind and non-double-blind studies evaluating secondary treatments, larger gains were associated with nondrug treatments than with drug regimens.

To attempt to hold constant all three factors at once (primary/secondary, placebo/nonplacebo, drug/nondrug) and make a more balanced comparison between double-blind and nonblinded studies, we examined the average gains of *secondary drug* treatments that were *not placebo controlled.* This group had by far the larg-

TABLE 5
Gains of Comparisons for Secondary Treatments in Surgery Evaluated by Randomized Controlled Trials, by Blinding of Patients and Assessors (Difference in Percentage of Treatment Successes and Associated Mann–Whitney Statistic)

	Number of comparisons	Difference in percentage of treatment successes Mean	SD†	Equivalent Mann–Whitney statistic* Mean
Double-blind comparisons	22	12.5%*	15.9%	0.56
Patients blind only	3	2.4%	2.6%	0.51
Assessors blind only	7	2.0%	11.9%	0.51
No blinding	29	2.4%	17.2%	0.51

*Difference in percentages of 19.0% for 12 placebo-controlled trials and 4.7% for 10 non-placebo-controlled trials.
†Standard deviation of the individual measurements, not the standard deviation of the mean.

est number of comparisons of any subgroups of these three variables. Within this subset, 10 double-blinded comparisons produced an average difference in the proportion of treatment successes of 4.7%, compared to −2.2% for 14 nonblinded comparisons ($p=0.80$). Thus, while not statistically significant, the tendency for double-blinded studies to find *larger* gains than studies that used no blinding (in terms of the difference in proportions of treatment successes) persisted when we held constant for whether the comparison was of a primary or secondary treatment, involved drugs or other therapy, and used a placebo control.

DISCUSSION

In this study we observed that nonrandomized studies tended to report larger gains than did randomized studies. This finding, although not striking, is generally consistent with previous investigations of study design and reported gains (2–6). For surgery, our results are consistent for primary and secondary treatments, and include in the nonrandomized studies investigations that use sequential comparisons, external controls, or observational designs.

These results are consistent with previous work reported by Gilbert, McPeek, and Mosteller (5), who observed that greater gains for the innovation are reported from surgical trails using a nonrandom study design than for randomized trials. Chalmers et al (3) have also reported this association from their study of anticoagulants in acute myocardial infarction. A similar association was observed by Wortman and Yeatman (6) in their meta-analysis of the efficacy of coronary bypass graft surgery. Shaikh et al (11) reviewed studies evaluating the efficacy of tonsillectomy and adenoidectomy and, after scoring each article for the quality of design and reporting, concluded that studies with a lower quality score were more likely to favor tonsillectomy than be against it.

A strength of our analysis is the mixing together of evaluations from different areas of medicine and surgery, because the findings would have wider application. A corresponding potential weakness is that studies using different types of design may not evaluate innovations with the same underlying distribution of gain. Bailar has suggested (12), for example, that randomized trials are often undertaken to confirm observations made from methodologically weaker studies, undertaken without full understanding of relevant study design factors. If so, then the difference between average gains of different study types may not be due to study design (or random variation), but to systematic differences in the underlying distribution of gain. Another potential weakness of the medicine study is the mixing of disciplines; however, when we repeated analyses controlling for the disciplines included in the analysis, the results remained unchanged.

We found no randomized cross-over studies in the survey of surgery journals, while 23% of comparisons of medical therapies were based on such a design. Further, we found that 51% of all evaluations in medicine came from a randomized controlled trial, compared to 37% for surgical treatments. This may reflect, in part, the requirements of the US Food and Drug Administration that new drugs and some therapeutic devices be evaluated through randomized controlled trials. (Surgical treatments are not generally subject to FDA approval.) Studies that used external controls or an observational design were rarely used in the evaluation of medical therapies, whereas they accounted for 52% of surgery comparisons.

CONCLUSION

Fineberg (13) has reviewed the association between the study design used in the evaluation of technology and subsequent clinical practice, and concluded that stronger forms of evaluation such as controlled

studies are not notably more successful than weaker forms in shaping medical practice. For example, although many cancer researchers hold the opinion that randomized trials have generally been more influential than nonrandomized trials in developing medical therapies, accepted treatments for acute leukemia have more commonly derived from nonrandomized trials than from randomized trials (14).

One purpose of our research was to help readers interpret findings from studies of various designs. This requires assumptions of the comparability of the underlying distribution of gain for different study types that on one hand may not be exactly correct, but on the other may be near enough to give the reader some caution in interpreting findings of nonrandomized evaluations. Therefore, we have attempted to summarize the adjustments that should be made when such assumptions are correct. By considering the reported value and an adjusted value, readers can quantitatively face the possible need for reductions in reported improvements from weaker designs. In that spirit, we offer the following average adjustments to be considered along with the original findings.

Inclusion of patients refractory to standard therapy increases the likelihood of a positive response to therapy. For medical studies using this entry criterion, the Mann–Whitney statistic may be reduced by 0.33. For other nonrandomized sequential studies, in both medical and surgical treatments, the Mann–Whitney statistic may be reduced by about 0.15 (for surgery, this corresponds to a reduction of 30% in the difference in proportions of treatment successes).

For medical therapies and secondary treatments in surgery, the results for nonrandomized parallel comparisons, external controls, and observational studies were not significantly different from the results for randomized controlled trials. We therefore offer no adjustment for these study types when used in medicine or for secondary treatments in surgery, noting that the statistical power of comparisons of these study types with randomized controlled trials was extremely limited due to the relatively small number of studies found.

For primary treatments in surgery, the Mann–Whitney statistic from nonrandom parallel comparisons and external controls should be reduced by 0.06 (or equivalently, the difference in proportions of treatment successes should be reduced by 13%). The results for observational studies evaluating primary treatments were sufficiently close to randomized controlled trials that no adjustment is offered.

Within randomized controlled trials, we found consistent results from medicine and surgery on the effects of using a placebo control. Placebo-controlled trials in medicine had an average Mann–Whitney statistic of 0.72, compared to 0.61 for trials not using a placebo. In surgery, the difference between placebo-controlled trials had an average Mann–Whitney statistic of 0.60, and nonplacebo trials a Mann–Whitney statistic of 0.52. It is possible that a placebo may be used when a standard treatment exists. If so, the efficacy of the standard therapy, rather than patient outcomes on placebo, would form a more appropriate baseline for evaluating the innovation. To account for the use of placebos in randomized controlled trials where a standard treatment is in fact available, the Mann–Whitney statistic may be reduced on the order of 0.10.

Within medicine, not using a double-blind design in randomized controlled trials was associated with an increase of 0.11 for the Mann–Whitney statistic; such studies might therefore have their average gain reduced by this amount. In surgery, however, we found that failure to use a double-blind design was associated with a statistically insignificant, but nonetheless surprising, *decrease* in the Mann–Whitney statistic of 0.03 (equivalent to a reduction in the difference of success proportions of 6.9%) for secondary drug treatments that

were not placebo controlled, the largest subset of such studies.

The features discussed above should be considered when evaluating the report of a new therapy and its possible application in clinical practice. Though we cannot be sure that these reductions are appropriate, considering them may suitably temper our enthusiasm for results based on weaker designs.

Acknowledgments—The contribution of readers who participated in this project is gratefully acknowledged: Anna Angelos, Andrew L. Avins, Jesse Berlin, Karen Biernstein, Christina Braun, Evridaki Chatzandreou, Ruth Cogan, Timothy R. Cote, Katherine M. Coughlin, G. Wade Davis, Mary Ettling, Joy R. Esterlitz, Manuel Friere, Paul Gompers, Joan Gopin, Jacquelyn Hedlund, David Hunter, Katheryn Lasch, Kihan Lee, Ann Levin, Mark McClellan, Alesia Maltz, Rama Peri, Laura Rosen, Heather Sacks, Roger Snow, Gail Speck, Lory Stapsy, Margaret Stassen, and Ellen Velie. We received many helpful comments from members of the Quantitative Evaluation Project at the Harvard School of Public Health. Phillip Lavori assisted in the development of the classification of study types.

REFERENCES

1. United States Institute of Medicine, Division of Health Sciences Policy. *Assessing medical technologies*. Washington, DC: National Academy Press; 1985.
2. Chalmers TC, Block JB, Lee S. Controlled trials in clinical cancer research. *N Engl J Med*. 1972;287:75-87.
3. Chalmers TC, Matta RJ, Smith H Jr, Kunzler AM. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med*. 1977;297:1091-1096.
4. Sacks HS, Chalmers TC, Smith H Jr. Randomized versus historical assignment in controlled clinical trials. *N Engl J Med* 1983;309: 1353-1361.
5. Gilbert JP, McPeek B, Mosteller F. Progress in surgery and anesthesia: Benefits and risks of innovative therapy. In: Bunker JP, Barnes BA, and Mosteller F. eds. *Costs, risks, and benefits of surgery*. New York: Oxford University Press; 1975:124-169.
6. Wortman PM, Yeatman WH. Synthesis of results in controlled trials of coronary bypass graft surgery. In: Light R, ed. *Evaluation studies review annual* vol. 8. New York: Sage Publications; 1983:536-557.
7. Straw RB. Deinstitutionalization in mental health: A meta-analysis. In Light R, ed. *Evaluation studies review annual* vol. 8. New York: Sage Publications; 1983:253-278.
8. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of medical therapies. (submitted for publication)
9. Miller JN, Colditz, GA, Mosteller F. How study design affects outcomes in comparisons of surgical treatments. (submitted for publication)
10. Bailar JC, Louis TA, Lavori PW, Polansky M. A classification tree for biomedical research reports. *N Engl J Med*. 1984;311:705-710.
11. Shaikh W, Vayda E, Feldman W. A systematic review of the literature on evaluation studies of tonsillectomy and adenoidectomy. *Pediatrics*. 1976;57:401-407.
12. Bailar JC. Research quality, methodologic rigor, citation counts, and impact. *Am J Public Health*. 1982;72:1103-1104.
13. Fineberg HV. Effects of clinical evaluation on the diffusion of technology. In *Assessing medical technologies*. Washington, DC: National Academy Press; 1985:176-210.
14. Gehan EA. Progress in therapy in acute leukemia 1948-1981: Randomized versus non-randomized clinical trials. *Controlled Clin Trials*. 1982;3: 199-208.