STATISTICS IN MEDICINE, VOL. 8, 455–466 (1989)

# HOW STUDY DESIGN AFFECTS OUTCOMES IN COMPARISONS OF THERAPY. II: SURGICAL

JAMES N. MILLER

*Center for Science and International Affairs, John F. Kennedy School of Government, Harvard University, Cambridge, MA 02138, U.S.A.*

GRAHAM A. COLDITZ*

*Channing Laboratory, Department of Medicine, Harvard Medical School and Brigham Women's Hospital, 180 Longwood Avenue, Boston, MA 02115, U.S.A.*

AND

FREDERICK MOSTELLER

*Technology Assessment Group, Department of Health Policy and Management, Harvard University School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.*

## SUMMARY

We analysed the results of 221 comparisons of an innovation with a standard treatment in surgery published in six leading surgery journals in 1983 to relate features of study design to the magnitude of gain. For each comparison we measured the gain attributed to the innovation over the standard therapy by the Mann–Whitney statistic and the difference in proportion of treatment successes. For primary treatments (aimed at curing or ameliorating a patient's principal disease), an average gain of 0·56 was produced by 20 randomized controlled trials. This was less than the 0.62 average for four non-randomized controlled trials, 0·63 for 19 externally controlled trials, and 0·57 for 73 record reviews (0·50 represents a toss-up between innovation and standard). For secondary therapies (used to prevent or treat complications of therapy), the average gain was 0·53 for 61 randomized controlled trials, 0·58 for eleven non-randomized controlled trials, 0·54 for eight externally controlled trials, and 0·55 for 18 record reviews. Readers of studies evaluating new treatments, particularly for primary treatments, may consider adjustment of the gain according to the study type.

KEY WORDS Research design Bias Evaluation of therapy

## INTRODUCTION

Study designs used in evaluations of new surgical therapies include randomized and non-randomized controlled trials, the comparison of a series of patients with results from the literature, and reviews of patient records. To interpret evaluations of new surgical therapies, we need to understand how reported gains may depend on study design. This paper is a companion to one on medicine (Part I: Medical). The introduction to the first paper serves also as a fuller introduction to this one. The issues relating to study design and the size of gains for new treatments discussed in the companion paper on medical therapies also apply to consideration of surgical therapies (see Part I: Medical), and are not repeated in this paper. To quantify the relationship between study type and the magnitude of gains reported for new surgical therapies, we examined studies from six

* Reprint requests to Dr G. Colditz.

0277–6715/89/040455–12$06.00

*Received December 1987*
*Revised August 1988*

surgery journals. In addition, we explored the relation between blinding and gains reported, and the type of observational design used. With these data we provide for consideration quantitative adjustments to the improvements reported from studies that used weaker designs. In this paper we do not repeat the description of methods and the discussion of results which applied to both the medical and surgical study; rather we refer the reader to Part I: Medical.

## MATERIALS AND METHODS

We reviewed all articles published during 1983 in six leading surgery journals: *American Journal of Surgery*; *Annals of Surgery*; *Archives of Surgery*; *British Journal of Surgery*; *Surgery*; and *Surgery, Gynecology and Obstetrics*. Two readers independently read each article to decide whether it qualified for inclusion. All readers participated in several hours of training. We initially accepted an article for further review if either of the two readers believed that it qualified for inclusion.

### Eligibility

To qualify for the surgical study, we used the criteria described for medicine (see Part I: Medical) and also required the report of outcomes in terms of the proportions of treatment successes for the innovation and the standard therapy (this allowed the computation of a quantitative measure of gain for comparison across study types).

People with training in statistical methods independently read each article initially selected. For each article, these readers identified the innovation and the standard, defined the study type, filled out a checklist on study design, and extracted the gains produced by each study (see below). Usually another person adjudicated any differences between the first two readers and made the final decision about article inclusion. Owing to personnel changes, only two people read the final 20 articles (out of 188 articles included). For these articles, the second reader effectively acted also as adjudicator after completion of the independent reading.

### Measuring gain

We measured the gain attributed to the innovation over the standard therapy in two ways. First, we used the Mann–Whitney statistic to estimate the probability that a randomly selected patient will perform better given the innovation than a randomly selected patient given the standard treatment (see Appendix to Part I: Medical). Second, we found the difference in the proportion of each group considered as treatment successes; positive values favouring the innovation. We also assigned a score to the conclusion reached by the authors regarding the relative merit of the innovation versus the standard therapy (see Part I: Medical).

We used the criteria defined in Part I to categorize studies by type. To allow consideration of the strength of design within randomized and non-randomized controlled trials, we modified a checklist used by DerSimonian, *et al.*[1] to represent more closely the strength of study design, execution, and analysis rather than the quality of reporting (see Part I: Medical).

### Analysis

We computed average gains within each study type, with use of the Mann–Whitney statistic, the difference in proportions of treatment successes, and the rating of authors' conclusions. To sharpen comparisons of the average gains produced by different study types, we stratified all studies by whether evaluation concerned primary or secondary treatments. 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 anaesthesia or postoperative care.'[2]

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 distribution.

We used Spearman's rank correlation coefficient to consider the relation between the reporting/quality score and the gains produced by randomized and non-randomized controlled trials. In computing *p*-values, we made no adjustment for multiple tests.

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

## RESULTS

The 188 articles that qualified for inclusion provided 221 comparisons of an innovation with a standard therapy in surgery. The *British Journal of Surgery* provided the highest proportion of randomized controlled trials, 62 per cent of all comparisons from that journal, followed by *Surgery* with 48 per cent, and *Surgery, Gynecology and Obstetrics* with 40 per cent. *American Journal of Surgery*, *Annals of Surgery*, and *Archives of Surgery* all had fewer randomized controlled trials than the other three journals: 20, 29, and 26 per cent of all comparisons, respectively (Table I). Overall, 37 per cent of comparisons used the randomized controlled trial, 7 per cent used non-randomized controlled trials, 12 per cent externally controlled, 41 per cent observational study, and 3 per cent pre/post comparisons. In Table I, the usual chi-square test for independence of rows and columns gives a significant departure ($p \sim 0{\cdot}01$). An exploratory analysis of the chi-values $(X_{ij}-\mu_{ij})/\sqrt{\mu_{ij}}$, where $X_{ij}$ is the count in the cell and $\mu_{ij}$ the usual expected value as presented in Mosteller and Parunak (Reference 3, pp. 212–213) suggests that the *American Journal of Surgery* has a high number of externally controlled trials and the *British Journal of Surgery* a high number of randomized controlled trials.

The distribution of treatment successes by study design was similar for primary and secondary therapies (see Table II).

### Gains for primary treatments

The mean of the Mann–Whitney statistic was 0·56 for randomized controlled trials that evaluated primary treatments and the average difference in the proportions of treatment successes was 11·9 per cent, smaller values than for all other study types; randomized controlled trials also had a smaller average rating of authors' conclusions than all other study types except observational studies (Table III). When we compared the gains in different study types to the gains for the randomized controlled trials, we found these differences generally not statistically significant, in part due to the limited sample sizes involved.

The results for pre/post comparisons, however, did differ significantly from randomized controlled trials. Pre/post comparisons had an average Mann–Whitney statistic of 0·78, an average difference in the proportions of treatment successes of 56·5 per cent and an average rating

Table I. Number (and percentage*) of comparisons from each journal by study type

| Study type | Journal | | | | | | |
|---|---|---|---|---|---|---|---|
| | *American Journal of Surgery* | *Annals of Surgery* | *Archives of Surgery* | *British Journal of Surgery* | *Surgery* | *Surgery, Gynecology and Obstetrics* | Total |
| Randomized controlled trial | 10 (20%) | 10 (29%) | 8 (26%) | 24 (62%) | 14 (48%) | 15 (40%) | 81 (37%) |
| Non-randomized controlled trial | 3 (6%) | 2 (6%) | 4 (13%) | 2 (5%) | 1 (4%) | 3 (8%) | 15 (7%) |
| Externally controlled trial | 14 (28%) | 3 (9%) | 3 (10%) | 0 (0%) | 3 (10%) | 4 (10%) | 27 (12%) |
| Observational study | 22 (44%) | 18 (53%) | 15 (49%) | 13 (33%) | 9 (31%) | 14 (37%) | 91 (41%) |
| Pre/post comparison | 1 (2%) | 1 (3%) | 0 (3%) | 2 (0%) | 2 (7%) | 7 (5%) | (3%) |
| Total | 50 | 34 | 31 | 39 | 29 | 38 | 221 |

* Percentages shown are based on column totals. For example, 20 per cent of the comparisons from *American Journal of Surgery* came from randomized controlled trials

Table II. Study design by author's conclusion of the value of the innovation in therapy versus standard therapy

| | Primary | | | | Secondary | | | |
|---|---|---|---|---|---|---|---|---|
| Study design | number | better | equal | worse | number | better | equal | worse |
| Randomized controlled trial | 20 | 50% | 45% | 5% | 61 | 57% | 44% | 5% |
| Non-randomized controlled trial | 4 | 50% | 50% | — | 11 | 64% | 36% | 1 |
| Externally controlled trial | 19 | 79% | 10·5% | 10·5% | 8 | 75% | 25% | — |
| Observational study | 73 | 64% | 21% | 15% | 18 | 67% | 22% | 11% |
| Pre/post | 6 | 100% | — | — | 1 | 100% | — | — |

of authors' conclusions of 5·33 (between 'innovation preferred' and 'innovation greatly preferred').

We explored the relation between study size and the size of the gain reported for the innovation compared to standard therapy. The Spearman rank correlation was $-0{\cdot}4$ ($p=0{\cdot}08$) for the 20 randomized controlled trials, $-0{\cdot}6$ ($p=0.4$) for the non-randomized controlled trials, $-0.38$ ($p=0{\cdot}11$) for the 19 externally controlled trials, $-0{\cdot}17$ ($p=0{\cdot}15$) for 73 observational studies, and 0·23 ($p=0{\cdot}65$) for the six pre/post comparisons.

### Gains for secondary treatments

The overall results for secondary treatments were similar to those for primary treatments: the average Mann–Whitney statistic was 0·53 and the average difference in the proportions of treatment successes for randomized controlled trials (6.0 per cent) were both smaller than for all other study types, as was the average rating of authors' conclusions of 4·51 (Table IV).

We found the greatest gain for the innovation over the standard therapy among non-randomized controlled trials, although this was not significantly larger than the average for randomized controlled trials ($p=0{\cdot}41$).

We explored the relation between study size and the gain observed. The Spearman rank correlation was 0·15 ($p=0{\cdot}25$) for the 61 randomized controlled trials, $-0{\cdot}30$ ($p=0{\cdot}37$) for the eleven non-randomized controlled trials, $-0{\cdot}16$ ($p=0{\cdot}69$) for the eight externally controlled trials, and $-0{\cdot}53$ ($p=0.03$) for the 18 observational studies.

### Primary versus secondary treatments

The secondary studies in surgery had more subjects receiving standard therapy for each study design and a lower proportion of treatment successes on standard therapy than the primary studies. For randomized controlled trials the mean number of subjects for the 20 primary studies was 44·7 compared to 92 for 61 secondary studies; and the mean proportion of successes on standard therapy was 0·70 for primary studies compared to 0·82 for secondary studies. For observational studies the mean proportion of successes on standard therapy was 0·57 for 73 primary studies and 0·79 for 18 secondary studies.

The average difference in the proportions of treatment successes was larger for primary than for secondary treatments for all but two study types. (The two exceptions to this rule each had only

Table III. Gains for primary treatments. Average difference in proportions of treatment successes and average rating of authors' conclusions by study type, with associated standard error of mean and *p*-values for comparison with average gains of randomized controlled trials

| Study type | Total comparisons | Mann–Whitney statistic | | Difference in proportion of treatment successes | | | Ratings of authors' conclusions | | |
|---|---|---|---|---|---|---|---|---|---|
| | | average | s.e.* | average | s.e.* | *p*-value† | average | s.e.* | *p*-value† |
| Randomized controlled trial | 20 | 0·56 | 0·02 | 11·9% | 3·5% | — | 4·50 | 0·27 | — |
| Non-randomized controlled trial | 4 | 0·62 | 0·06 | 24·8% | 12·4% | 0·32 | 4·75 | 0·75 | 0·76 |
| Externally controlled trial | 19 | 0·63 | 0·03 | 25·6% | 5·6% | 0·11 | 5·00 | 0·32 | 0·23 |
| Observational study | 73 | 0·56 | 0·02 | 14·5% | 3·4% | 0·30 | 4·41 | 0·17 | 0·61 |
| Pre/post | 6 | 0·78 | 0·03 | 56·5% | 6·3% | 0·0001‡ | 5·33 | 0·21 | 0·014 |

* Standard error of the mean gain within each study type
† *p*-values for comparison of gains for specific study type with gains for randomized controlled trials
‡ *p*-value less than 0·0001

Table IV. Gains for secondary treatments. Average difference in proportions of treatment successes and average rating of authors' conclusions by study type, with associated standard error of mean and *p*-values for comparison with average gains of randomized controlled trials

| | Total comparisons | Mann–Whitney statistic | | Difference in proportion of treatment successes | | | Ratings of authors' conclusions | | |
|---|---|---|---|---|---|---|---|---|---|
| | | average | s.e.* | average | s.e.* | *p*-value† | average | s.e.* | *p*-value† |
| Randomized controlled trial | 61 | 0·54 | 0·01 | 6.0% | 2·1% | — | 4·51 | 0·17 | — |
| Non-randomized controlled trial | 11 | 0·55 | 0·02 | 10·3% | 4·7% | 0·41 | 4·94 | 0·72 | 0·48 |
| Externally controlled trial | 8 | 0·54 | 0·02 | 7·6% | 3·4% | 0·69 | 4·93 | 0·25 | 0·12 |
| Observational study | 18 | 0·55 | 0·02 | 9·4% | 3·8% | 0·22 | 4·67 | 0·32 | 0·43 |
| Pre/post comparisons | 1 | 0·90 | — | 80·0% | — | — | 5 | — | — |

* Standard error of the mean gain within each study type

† *p*-values for comparison of gains for specific study type with gains for randomized controlled trials.

Table V. Gains of comparisons for secondary therapies evaluated by randomized controlled trials, by blinding of patients and assessors. Average difference in proportions of treatment successes and average rating of authors' conclusions by study type, with associated standard error of mean

| Blinding | Total comparisons | Mann–Whitney statistic | | Difference in proportion of treatment successes | | Ratings of authors' conclusions | |
|---|---|---|---|---|---|---|---|
| | | average | s.e.* | average | s.e.* | average | s.e.* |
| Double-blind comparisons | 22 | 0·59 | 0·03 | 12·5% | 3·4% | 5·04 | 0·25 |
| Patients blind only | 3 | 0·51 | 0·01 | 2·37% | 1·45% | 3·33 | 0·33 |
| Assessors blind only | 7 | 0·53 | 0·01 | 2·00% | 4·54% | 4·29 | 0·52 |
| No blinding | 29 | 0·51 | 0·02 | 2·44% | 3·23% | 4·28 | 0·24 |

* Standard error of the mean gain within each study type

one evaluation of a secondary treatment.) The average difference in the proportion of treatment successes was also less closely grouped around the gains for randomized controlled trials for primary than for secondary treatments. The average absolute difference between the average gains for randomized controlled trials and those for other study types was 8·8 per cent for primary treatments, and 2·9 per cent for secondary treatments. (We exclude the two study types that each contributed only one comparison.)

Primary therapies were evaluated in 25 per cent of randomized controlled trials and in 26 per cent of non-randomized controlled trials. In contrast, externally controlled trials, observational studies, and pre/post studies evaluated primary therapies in 73 per cent, 78 per cent and 83 per cent of all comparisons, respectively.

### Study design score

We used the study design score to assess the strength of design within the controlled clinical trials. The average study design score for the 81 randomized controlled trials was 4·94, not significantly different from the value of 5·00 for the 15 non-randomized controlled trials ($p=0·52$).

We found a weakly negative, and statistically insignificant, correlation between study design score and gain. (A negative correlation would mean that studies with small study design scores tended to be associated with large gains, and vice versa.) Spearman's rank correlation between the study design score and the difference in the proportion of treatment successes was $-0·04$ for randomized controlled trials ($p=0·70$), and $-0·27$ for non-randomized controlled trials ($p=0·37$). The rank correlation between the study design score and our scoring of authors' conclusions was $-0·09$ for randomized controlled trials ($p=0·40$) and $-0·43$ for non-randomized controlled trials ($p=0·14$). Thus, the tendency for studies with lower study design scores to find larger gains was not statistically significant at a $p$-value of 0·05.

### 'Blinding' of patients and assessors

We evaluated the relation between blinding and the size of gain within the randomized controlled trials. About half of all randomized controlled trials of secondary treatments (41 of 81) had either patients or assessors, or both, 'blind' to the treatment received (Table V). Contrary to the hypothesis that studies with weaker design tend to find larger gains, 'double-blind' comparisons

produced the largest average gains for secondary therapies, significantly larger than the average for comparisons that involved no blinding ($p=0{\cdot}032$ for difference in proportions of treatment successes; $p=0{\cdot}028$ for ratings of authors' conclusions). The small number of comparisons precluded further analysis of blinding within the 20 primary trials.

One possible source of the disparity between the average gains found by double-blind and non-blind randomized controlled trials may be that double-blind trials often use placebos (and by definition these are not applicable in studies where no blinding occurred). We explored this possible effect of placebos and observed that the average difference in the proportion of treatment successes for twelve double-blind trials that used a placebo was 19·0 per cent, compared to 4·7 per cent for ten double-blind trials that did *not* use a placebo. Another possible factor that we considered was whether the comparison involved a drug or some other treatment ('non-drug'). For both double-blind and non-double-blind studies that evaluated secondary treatments, we found larger gains associated with non-drug treatments than with drug regimens.

To attempt to hold constant all three factors simultaneously (primary/secondary, placebo/non-placebo, drug/non-drug) and make a more balanced comparison between double-blind and non-blinded studies, we examined the average gains of *secondary drug* treatments that were *not placebo controlled.* This group had by far the largest number of comparisons of any subgroups of these three variables. Within this subset, ten double-blind comparisons produced an average difference in the proportion of treatment successes of 4·7 per cent, compared to $-2.2$ per cent for 14 non-blind comparisons ($p=0{\cdot}80$). Thus, while not statistically significant, the tendency for double-blind 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.

**Study design and gain for observational studies**

To understand better the effect of study design on gain, we distinguished between four different kinds of observational studies:

(a) use of the innovation and the standard therapy during the same period;
(b) use of the innovation superseded that of the standard therapy;
(c) comparison of a record review for the innovation with external literature for the standard therapy;
(d) comparison of a record review for the standard therapy with external literature for the innovation.

These categories were chosen because they represent different temporal sequences in the use of the innovation and standard and each has a different prior likelihood of the innovation performing better (or worse) than the standard.

For primary treatments, observational studies with the innovation and standard therapy in use over the same period had a gain that was similar to that observed in the primary randomized controlled trials. Observational studies that involved comparison of a record review for the *standard* therapy (records usually had some association with the study's authors) to external literature for the *innovation* found the standard preferable to the innovation on the average (Table VI). Studies that involved a comparison of a record review for the *innovation* to external literature for the *standard*, found the innovation preferable to the standard on the average. In these studies, the authors used records that reflected their own experience to provide the data for the innovation.

Table VI. Gains for observational studies, by primary/secondary treatment. Average difference in proportions of treatment successes and average rating of authors' conclusions by study type, with associated standard error of mean

| | Total comparisons | Mann–Whitney statistic | | Difference in proportion of treatment success | | Ratings of authors' conclusions | |
|---|---|---|---|---|---|---|---|
| | | average | s.e.* | average | s.e.* | average | s.e.* |
| *Primary treatments* | | | | | | | |
| observational study (a) | 39 | 0·55 | 0·02 | 13·4% | 3·9% | 4·39 | 0·21 |
| observational study (b) | 16 | 0·57 | 0·04 | 16·1% | 8·1% | 4·75 | 0·36 |
| observational study (c) | 12 | 0·62 | 0·05 | 23·6% | 9·8% | 4·92 | 0·38 |
| observational study (d) | 6 | 0·47 | 0·09 | −0·5% | 17·1% | 2·83 | 0·75 |
| Total | 73 | 0·56 | 0·02 | 14·5% | 3·4% | 4·41 | 0·17 |
| *Secondary treatments* | | | | | | | |
| observational study (a) | 6 | 0·54 | 0·05 | 8·5% | 11·0% | 3·50 | 0·43 |
| observational study (b) | 7 | 0·53 | 0·01 | 6·1% | 1·5% | 5·43 | 0·30 |
| observational study (c) | 4 | 0·56 | 0·02 | 11·8% | 5·1% | 5·00 | 0·41 |
| observational study (d) | 1 | 0·64 | — | 27·5% | — | 5 | — |
| Total | 18 | 0·55 | 0·02 | 9·4% | 3·8% | 4·67 | 0·32 |

*Observational studies*
(a) innovation and standard therapy in use over same period
(b) innovation superseded standard
(c) record review for innovation; external literature for standard
(d) record review for standard; external literature for innovation

* Standard error of the mean gain within each study type

For secondary treatments, average authors' conclusions for randomized controlled trials were a full point *higher* than those for observational studies that involved use of the innovation and the standard over the same period ($p=0·13$); and a full point *lower* than for observational studies where the use of the innovation had superseded that of the standard ($p=0·007$). The disparity in the ratings of authors' conclusions between observational studies that involved use of the innovation and the standard over the same period, and those where the innovation superseded the standard, was statistically significant ($p=0·008$).

## DISCUSSION

In this study we observed that non-randomized studies tended to report larger gains than did the randomized studies. These results are generally consistent with previous investigations of study design and reported gains (see Part I: Medical). Our results are consistent for primary and secondary treatments, and include in the non-randomized studies investigations that use external controls, a pre-post design, and observational designs. The strengths and limitations of our approach to understanding the relation between study design and gain are discussed in the companion paper (Part I: Medical).

Gilbert *et al.*[3] investigated the gains attributed to innovations in surgery and reviewed randomized and non-randomized controlled trials. They found average differences in the proportion of treatment successes of 1·3 per cent for primary therapies and 0·4 per cent for

secondary therapies. These results contrast with corresponding gains of 12·5 per cent and 6·0 per cent in our study.

There are two notable differences between our results and those of Gilbert *et al.* First, we found larger gains, on the average, for both primary and secondary therapies, and for both randomized and non-randomized trials. Second, a much greater proportion of the randomized controlled trials in our study evaluated secondary rather than primary therapies: 60 of 81 comparisons, compared to 21 of 44 for Gilbert *et al.* A possible explanation for these differences is that the two studies drew their samples from different populations. Gilbert *et al.* used a search of the National Library of Medicine's Literature Retrieval System for papers published before 1977 while in the present study we surveyed all reports published in 1983 in six surgery journals.

We found no cross-over studies in this survey of surgery journals, though they appear commonly in medicine to evaluate therapies.[4] Indeed, when we examined the relation between study type and bias in medical interventions using a similar approach (see Part I: Medical), we found that 23 per cent of comparisons of medical therapies utilized a randomized cross-over design, and 27 per cent of comparisons involved a non-randomized sequential design. Further, we found that 37 per cent of all evaluations of innovations in surgery came from a randomized controlled trial, compared to 51 per cent for all new medical therapies. This may reflect, in part, the impact of the U.S. Food and Drug Administration on the evaluation of prescription medicines. The negative correlation relating study design score with average gains in the randomized controlled trials of surgical therapies was similar to that observed in the medical evaluations. In both surgical and medical randomized controlled trials, those studies using a placebo control had a greater gain than similar trials using an active therapy as the standard.

## Interpreting results from non-randomized studies

One purpose of our research was to help readers interpret findings from studies of different designs. Readers may use non-quantitative means to discount the findings of less well controlled studies. Our results may help in suggesting quantitatively how the reader might discount or consider discounted findings from studies with non-randomized designs. The magnitude of an adjustment for bias due to study design may vary even with a given study design, however, we have focused this analysis on the average for each design. By considering the reported gain and an adjusted value for this gain, readers can temper their views by facing quantitatively the possible need for reductions in the reported improvements from weaker designs.

For both primary and secondary therapies, the results from non-randomized studies seem to need discounting. Compared with randomized controlled trials, one might reduce the Mann–Whitney statistic for each of non-randomized controlled trials and externally controlled trials that address primary therapies by 0·06. Similarly, one could reduce the observational studies that evaluate an innovation in primary therapy and a standard therapy in use over the same time period by 0·5. Among the secondary therapies, the corresponding reductions for non-randomized designs are 0·02. Though we cannot assure the appropriateness of these reductions, their consideration may suitably temper one's enthusiasm for results based on weaker designs.

In our more exploratory analyses, we observed authors somewhat more enthusiastic about the innovation for observational studies based only on record reviews when the innovation had superseded the standard, and much less enthusiastic when the innovation and the standard were used over the same period. One might also consider these design features in evaluation of the report of a surgical therapy.

ACKNOWLEDGEMENTS

Supported by the National Center for Health Services Research and Health Care Technology Assessment Grant no. HSR 1 R01 HSO5156-01 and HS-05936, the Macy Foundation, and the Rockefeller Foundation.

## REFERENCES

1. DerSimonian, R., Charette, L. J., McPeek, B. and Mosteller, F. 'Reporting on methods in clinical trials', *New England Journal of Medicine*, **311**, 442–448 (1984).
2. Gilbert, J. P., McPeek, B. and Mosteller, F. 'Progress in surgery and anesthesia: benefits and risks of innovative therapy', *in* Bunker, J. P., Barnes, B. A. and Mosteller, F. (eds) *Costs, Risks, and Benefits of Surgery*, Oxford University Press, New York, 1975, Chapter 9.
3. Mosteller, F. and Parunak, A. 'Identifying extreme cells in a sizable contingency table: probabilistic and exploratory approaches', *in* Hoaglin, D.C., Mosteller, F. and Tukey, J. W. (eds), *Exploring Data Tables, Trends, and Shapes*, Wiley, New York, 1985, pp. 189–224.
4. Bailer, J. C., Louis, T. A., Lavori, P. W. and Polansky, M. 'A classification tree for biomedical research reports', *New England Journal of Medicine*, **311**, 705–710 (1984).