STATISTICS IN MEDICINE, VOL. 8, 441–454 (1989)

HOW STUDY DESIGN AFFECTS OUTCOMES IN COMPARISONS OF THERAPY. I: MEDICAL

GRAHAM A. COLDITZ*

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

JAMES N. MILLER

Center for Science and International Affairs, John F. Kennedy School of Government, Harvard University,
Cambridge, MA 02138, 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 113 reports published in 1980 in a sample of medical journals to relate features of study design to the magnitude of gains attributed to new therapies over old. Overall we rated 87 per cent of new therapies as improvements over standard therapies. The mean gain (measured by the Mann–Whitney statistic) was relatively constant across study designs, except for non-randomized controlled trials with sequential assignment to therapy, which showed a significantly higher likelihood that a patient would do better on the innovation than on standard therapy ($p=0{\cdot}004$). Randomized controlled trials that did not use a double-blind design had a higher likelihood of showing a gain for the innovation than did double-blind trials ($p=0{\cdot}02$). Any evaluation of an innovation may include both bias and the true efficacy of the new therapy, therefore we may consider making adjustments for the average bias associated with a study design. When interpreting an evaluation of a new therapy, readers should consider the impact of the following average adjustments to the Mann–Whitney statistic: for trials with non-random sequential assignment a decrease of 0·15, for non-double-blind randomized controlled trials a decrease of 0·11.

KEY WORDS Research design Gains Evaluation of therapy Bias

INTRODUCTION

Investigators use a variety of study designs to assess the effectiveness of innovations. The Institute of Medicine's Book *Assessing Medical Technologies*[1] describes many of these. This paper (Part I) on medicine, and its companion (Part II) on surgery, examine gains observed in comparative clinical studies to see the effect of study design on outcome.

When investigators compare the performance of a new treatment in medicine or surgery to standard practice, the average gain attributed to the innovation is often found to be greater in non-randomized studies than in randomized studies.[2–6] Not all investigations have shown this result,[7]

* Reprint requests to Dr. G. Colditz, Channing Laboratory, 180 Longwood Avenue, Boston, MA 02115, U.S.A.

0277–6715/89/040441–14$07.00

Received December 1987
Revised August 1988

but this is a typical finding. These findings raise the question of how to interpret and combine the outcomes from non-randomized studies with those of randomized trials. We consider some options.

Possibilities range from combining all studies giving each equal weight, to ignoring all studies except randomized controlled trials. The evidence from the literature suggests that equal weights may not be appropriate. Setting all the non-randomized information aside seems a terrible waste. Nevertheless, Sackett (Reference 8, page 148) suggests that the busy clinician trying to 'keep up with the clinical literature' (rather than searching the clinical literature to decide how to treat a specific patient) discard at once all articles on therapy that are not randomized trials.

Less weight could be given to studies that seemed more biased, and this would reduce the amount of the bias, but would not directly address the issue. Cochran (Reference 9, Page 14) in his book *Planning and Analysis of Observational Studies* says that investigators will find it 'worthwhile to think hard about what biases are most likely, and to think seriously about their sources, directions, and even their plausible magnitudes'. Elsewhere (Reference 9, page 58) he suggests that if an investigator knows the size of the bias 'fairly well', that an adjustment for the bias be made. Once the magnitude is assessed, one could offer an adjustment for consideration,[10] and this is the direction we have taken.

Because this is the first of two papers dealing with such differences related to study design, the issues leading to these papers and some of the methodological points for both papers are contained in this paper and are not repeated in the second on surgery.

METHODS

Article identification

We chose four medical disciplines, cardiology, neurology, psychiatry, and respiratory medicine, to represent a range of contemporary medical therapies evaluated and reported in the literature. After ranking journals listed under these disciplines in Index Medicus by their impact factor[11] as defined and reported by the Science Citation Index, we drew a stratified random sample of journals within each discipline. Impact factor measures the average frequency of citation of an article in a particular year. The impact factor is a ratio between citations and citable items published. It helps to evaluate the absolute citation frequency by partially discounting the advantage of large journals over smaller ones.

Within each of the four medical disciplines, we ranked journals into quintiles of impact factor and we selected one journal from each quintile by the use of random digit tables. We then checked each selected journal to assure that it contained at least some studies, written in English, of medical therapy on humans with the response to therapy as the outcome measure. We excluded journals that failed to meet these criteria and replaced them with other randomly selected journals from the same quintiles of impact factor.

The journals selected for this study and listed in order of decreasing impact factor were, for Cardiology: *American Journal of Cardiology*; *American Heart Journal*; *British Heart Journal*; *Japanese Heart Journal*; *Acta Cardiologica*; Neurology: *Annals of Neurology*; *Paraplegia*; *Canadian Journal of Neurological Sciences*; *Acta-Neurological Scandavica*; *Clinical Neurology and Neurosurgery*; Psychiatry: *American Journal of Psychiatry*; *Journal of Nervous and Mental Disease*; *Hospital Community Psychiatry*; *Social Psychiatry*; *Australian and New Zealand Journal of Psychiatry*; and Respiratory Medicine: *American Review of Respiratory Diseases*; *Chest*; *Thorax*; *European Journal of Respiratory Disease*; *Respiration.*

We identified eligible articles by systematic review of each journal issue with a publication date in 1980. To be eligible, an article had to report on an evaluation of medical therapy with the response to therapy as the outcome measure. At least ten subjects had to be included in the study (five for cross-over studies), and the outcome had to be reported for both standard therapy and the innovation. We subjected the identified articles to a second level of reading to determine final eligibility for this study. We included 113 articles that contained a total of 128 comparisons of a standard therapy with an innovation.

Data extraction

Two readers with training in statistical methods independently read each article. These readers first underwent a training programme to ensure uniform understanding and application of the definitions for data they extracted. We assigned articles randomly to pairs of readers, each of whom recorded the study design in the reported article and then completed a checklist. We used the following classification into six study designs, based on a scheme devised by Bailar *et al.*:[12] randomized controlled trial with parallel control groups; randomized controlled trial with sequential control (cross-over study); non-randomized controlled trial with parallel control groups; non-randomized controlled trial with sequential control (non-random cross-over); externally controlled trial (often comparing results of a series of patients to results previously reported by others and then classified as a comparison with 'historical controls'); and observational study. Observational studies use retrospective record reviews, and often included additional follow-up of patients. Such investigations differ fundamentally from the other study types, which are pre-planned for at least one of the treatments considered. A third reader independently adjudicated and resolved any differences in the classification of study design or data recorded on checklists.

Each reader indicated on the checklist whether or not the article covered each of a number of items. We modified the items on the checklist from those used by DerSimonian *et al.*[13] to represent more closely strength of study design, execution, and analysis rather than the quality of reporting. The items included:

1. *Eligibility criteria*: information explaining the criteria for admission of patients to the trial of the therapy.
2. *Completeness of admission*: the completeness of admission of patients who met eligibility criteria into the trial.
3. *Admission before allocation*: information indicating whether eligibility criteria were applied in the absence of knowledge about how the patient, if admitted, was to be treated.
4. *Random allocation*: information detailing the actual mechanism used to generate the random assignment, or for non-random studies, information on the method of allocation.
5. *Patients' blindness to treatment*: information about whether patients knew which treatment they had received.
6. *Blind assessment of outcome*: information about whether the person assessing the outcome knew which treatment had been given.
7. *Loss to follow-up*: information about the numbers of patients lost or otherwise omitted in follow-up and the reasons why they were lost.
8. *Statistical analyses*: analyses going beyond the computation of means, percentages, or standard deviations.
9. *Statistical methods*: names of tests, techniques, or computer programs used for specified statistical analyses.

10. The reported outcome in the group receiving the innovation in therapy.
11. The reported outcome in the group receiving standard therapy.
12. The authors' conclusion regarding the value of the new therapy.

Readers rated the authors' conclusion on a six point scale previously used by Gilbert *et al.*[6] to score the authors' conclusion as to the value of the innovation evaluated. The rating scale is set out below:

6 Innovation highly preferred to standard.
5 Innovation preferred to standard.
4 About equal, innovation a success.
3 About equal, innovation a disappointment.
2 Standard preferred to innovation.
1 Standard highly preferred to innovation.

Readers recorded the *p*-value (usually two-sided) for differences in performance between innovation and standard therapy when reported by the investigators.

Measures of gain

We employed two measures to estimate the magnitude of outcomes for comparisons of new with standard therapy. 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). One can calculate the Mann–Whitney statistic from many different statistical measures often reported in published comparisons of medical treatments, for example, proportion surviving, mean change in blood pressure, and frequency of side effects. Second, readers rated the conclusion reported by the authors regarding the relative overall merit of the innovation compared to standard therapy (see item 12). For example, the authors' conclusion 'the average blood pressure of the patients under active treatment was significantly lowered from the first week of the trial onwards . . .'[14] had a rating of 6, and 'the failure to demonstrate a significant difference in PEF between our study groups raises serious concerns about [the innovation]'[15] had a rating of 2.

Quality score

For each randomized controlled trial and each non-random sequential study, we calculated a quality score for the study as the number of items reported by the investigators from the following seven (defined above):

1. eligibility criteria
2. completeness of admission
3. admission before allocation
4. random allocation
5. loss to follow-up
6. statistical analyses
7. statistical methods.

Although our focus was the quality of design, the quality of reporting may influence this quality score.

Table I. Study design and study size for comparisons of innovations in medical therapy

	Number of studies	Mean number of subjects receiving innovation	Median number of subjects receiving innovation
Randomized controlled trial (parallel)	36	33	19·5
Non-random parallel comparisons	3	19·3	18
Randomized controlled trials (cross-over)	29	15·8	15
Non-random sequential comparisons	46	20·3	14
External controls	5	31·8	24
Observational studies	9	113·8	78

Analysis

We computed average gains within each study design, with use of the Mann–Whitney statistic and the rating of the authors' conclusions. We tested for differences between the mean gains among different study designs with use of Student's *t*-test. Our prior hypothesis was that 'higher quality' studies that employ randomization and blinding would tend to find smaller gains for new therapies than studies that were less well controlled.

We used Spearman's rank correlation coefficient to consider the relationship 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.

RESULTS

The 128 comparisons of an innovation with standard therapy were from the four disciplines in the following numbers: cardiology 39; neurology 11; respiratory medicine 62, and psychiatry 16. Among the disciplines, respiratory medicine had the highest proportion of randomized controlled trials (67 per cent) followed by cardiology and neurology (45 per cent), while psychiatry had no comparisons that used this design. Non-random sequential studies represented 45 per cent of comparisons in cardiology and neurology, 24 per cent of comparisons in respiratory medicine and 50 per cent of psychiatry. Within the study types, the median number of patients receiving the innovation in each comparison did not differ greatly, except for observational studies which had larger numbers of patients (see Table I). Because we had relatively few observational studies, we do not come to strong conclusions about them, although we have carried them in the tables.

We calculated the overall mean across studies for each of the measures of gain. Table II presents the average gains by study design. Regardless of the measure of gain used or the study design, we rated most new therapies as improvements over standard therapy. For randomized controlled trials that used 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 that we included, there was a 61 per cent chance that a random patient who received the innovation in therapy would fare better than a random patient on standard therapy. The mean rating of the authors' conclusion for the value of the innovation was 4·4, about midway between the ratings 'about equal, innovation a success' and 'innovation is preferred to standard'.

The average Mann–Whitney statistic was relatively constant for all but one study design, the non-randomized controlled trials, that used sequential assignment to therapy. This study design

Table II. Mann–Whitney statistic, and the rating of the authors' conclusion among a sample of evaluations of medical therapy reported in 1980

Study design	Number of studies	Mann–Whitney* statistic Mean	SD‡	Rating of authors' conclusion† Mean	SD
Randomized controlled trials (parallel)	36	0·61	0·14	4·4	1·0
Non-random parallel comparisons	3	0·56	0·07	4·7	1·5
Randomized controlled trials (cross-over)	29	0·63	0·14	4·6	0·8
Non-random sequential comparisons	46	0·81	0·15	4·9	0·8
(a) refractory	12	0·94	0·09	4·9	0·7
(b) other	34	0·76	0·16	4·9	0·9
External controls	5	0·65	0·10	5·6	0·9
Observational studies	9	0·57	0·04	4·4	0·8

* Estimated probability of a random patient performing better on the innovation than a random patient on the standard therapy

† Authors' conclusion scored from 'innovation highly preferred to standard' =6 to 'standard highly preferred to innovation' =1. An average score of 3·5 would correspond to no gain

‡ Standard deviation of the individual measurements, not the standard deviation of the mean

showed a substantial increase in the likelihood that a patient would do better on the innovation that on the standard therapy according to the Mann–Whitney statistic (Table II).

Within the non-random sequential trials, patients admitted to a study after having failed on standard therapy (that is, patients refractory to standard therapy), had, as might be expected, a very high probability of improvement with an innovative therapy. The average Mann–Whitney statistic for these trials with refractory patients was 0·94, whereas for trials with patients not refractory it was 0·76. Thus, even after removal of the trials with refractory patients, non-random sequential studies had a significantly higher likelihood of patients who succeeded on the innovation than did the randomized controlled trials ($t=2{\cdot}93$, $p=0{\cdot}004$).

The average rating for the original authors' evaluations of the innovation relative to standard therapy varied little by study design except for external controls (Table III). Among the randomized evaluations, only four of 65 studies (6·2 per cent) yielded ratings that show the standard was 'preferred' or 'highly preferred' to the innovation, and for the non-random comparisons one of 49 (2 per cent) studies scored so negatively. For 34 of 65 randomized studies (52 per cent) the innovation's rating was 'highly preferred' or 'preferred' to standard therapy, and for 40 of 49 non-random comparisons (82 per cent) the innovation was so rated. The difference in the average rating of the authors' conclusion between the randomized controlled trials (4·6) and the non-random comparisons (4·9) did not reach statistical significance ($t=1{\cdot}68$, $p=0{\cdot}1$). The average rating of the authors' conclusion was 4·9 for both non-random sequential trials that included refractory patients and for those that did not.

To examine the relationship between our two measures of gain we correlated the authors' conclusion with the Mann–Whitney statistic. The two measures of gain resulted in a high correlation; the Spearman rank correlation between the rating of authors' conclusion and the Mann–Whitney statistic ranged from 0·5 for non-random sequential studies to 0·7 for randomized controlled trials that used parallel control groups ($p<0{\cdot}01$ for all study designs).

We examined the relation between study size and the gain reported. Overall, the larger the study, the smaller the gain. The Spearman rank correlation between study size and gain was

Table III. Study design by authors' conclusion of the value of the innovation in therapy versus standard therapy

Study design	Total score	Innovation highly preferred 6	Innovation preferred 5	About equal, innovation a success 4	About equal, innovation a disappointment 3	Standard preferred 2	Standard highly preferred 1
Randomized controlled* trial (parallel)	36	7	10	10	5	2	2
Non-random parallel comparisons	3	1	1	0	1	0	0
Randomized controlled trial (cross-over)	29	11	6	11	1	0	0
Non-random sequential comparisons	46	9	29	5	2	1	0
(a) refractory	12	2	7	3	0	0	0
(b) other	34	7	22	2	2	1	0
External controls	5	4	0	1	0	0	0
Observational studies	9	2	3	2	1	1	0

* t-test comparing randomized controlled trials to non-random comparisons (parallel and sequential, excluding studies with refractory patients, external controls and observational studies) $t = 1{\cdot}68$, $p = 0{\cdot}1$

Table IV. Mean and standard deviation of quality score for design, implementation and reporting of evaluations by study design

	Number of studies	Mean	SD
Randomized controlled trials (parallel)	36	4·5	1·3
Randomized controlled trials (sequential)	29	3·6	1·7
Non-random sequential comparisons	46	4·2	1·3

Table V. Percentage of reporting individual items included in the quality score by study design

	RCT (P) (number = 36)	RCT (S) (number = 29)	Non-random sequential (number = 46)
Eligibility criteria	86%	86%	95%
Completeness of admission	22	11	12
Admission before allocation	61	50	49
Method of allocation	17	11	37
Loss to follow-up	94	46	72
Statistical analyses	95	79	88
Statistical methods	78	75	70

RCT(P), randomized controlled trial with parallel control group
RCT(S), randomized controlled trial with sequential control group

$-0{\cdot}25$ ($p=0{\cdot}14$) for the 36 randomized controlled trials, 1·0 for the three non-random parallel comparisons, 0·07 ($p=0{\cdot}69$) for the 29 randomized controlled trials (cross-over), $-0{\cdot}20$ ($p=0{\cdot}17$) for the 46 non-random sequential studies, $-0{\cdot}6$ ($p=0{\cdot}28$) for the five studies using external controls, and $-0{\cdot}5$ ($p=0{\cdot}19$) for the nine observational studies.

Quality of study design

The mean quality score calculated for each study design varied from 4·5 for the randomized controlled trials (parallel), to 3·6 for the randomized controlled trials (sequential), and 4·2 for the non-randomized sequential trials (see Table IV). The frequency of reporting for each item in the quality score appears in Table V. Within randomized controlled trials this quality score did not correlate with the score for the authors' conclusion (Spearmann $r=0{\cdot}04$), nor with the Mann–Whitney statistic (Spearman $r=-0{\cdot}16$, $p=0{\cdot}18$). The quality score, however, was correlated strongly with the impact factor score, (Spearman $r=0{\cdot}50$, $p=0{\cdot}0001$).

Blind evaluation of therapy

We evaluated the relationship between blinding and the size of gain within the randomized controlled trials by classifying the studies as double-blind, and non-double-blind (Table VI). The average Mann–Whitney statistic for double-blind trials was 0·58 and 0·69 for non-double-blind studies ($t=2{\cdot}4$, $p=0{\cdot}02$). The average rating of authors' conclusions showed a similar significant increase in reported value of therapies for non-double-blind studies (4·2 for double-blind studies and 4·9 for non-double-blind studies). We found the average quality score similar for both double-

Table VI. The use of blinding in the evaluation of medical therapies, by study design

	Number of studies	Double-blind	Not double-blind
Randomized controlled trials (parallel)	36	21	15
Non-random parallel comparisons	3	0	3
Randomized controlled trials (cross-over)	29	13	16
Non-random sequential comparisons	46	1	45
External controls	9	0	9
Observational studies	5	0	5

A further categorization shows that eleven randomized controlled trials (cross-over) used a single blind design as did two non-random sequential comparisons

blind (4·3) and non-double blind trials (4·2). The mean impact factor score for the journal in which each article was published was slightly but not significantly higher for non-double-blind (2·1) compared with double-blind studies (1·9).

Placebo control

The innovation in therapy involved comparison with placebo in 15 of the 65 randomized controlled trials. The likelihood of better performance on the innovation was significantly higher when the comparison was to placebo therapy (the mean Mann–Whitney statistic was 0·72 for placebo controlled trials and 0·61 for non placebo controlled trials; $t=2{\cdot}1$, $p=0{\cdot}04$). The authors' conclusions showed a similar directional change, although the difference was not statistically significant (4·9 for placebo controlled trials and 4·5 for trials that did not use a placebo, $p=0{\cdot}21$).

Tests of significance

The original articles did not always contain tests of significance and p-values. Within the randomized controlled trials, 18 of 65 studies (28 per cent) did not assess statistical significance. Of the 47 trials that did consider statistical significance, 21 reported results as 'not significant' or stated that p-values exceeded $p=0{\cdot}05$. For the non-random comparisons (parallel and sequential), 16 of 49 studies (33 per cent) did not report tests of significance. Of the remaining 33 studies, 28 reported a p-value of less than 0·05.

Impact factor

The Science Citation Index Impact Factor of the journals included in this study ranged from 6·1 for the *American Journal of Cardiology* to 0·3 for *Paraplegia.* Overall, the impact factor did not correlate with the measures of gain. For the Mann–Whitney statistic the Spearman correlation was 0·04 ($p=0{\cdot}79$) and for the authors' conclusion it was 0·02 ($p=0{\cdot}84$). The quality score, however, correlated significantly with the impact factor for the pooled randomized and non-randomized controlled trials (Spearman $r=0{\cdot}38$, $p=0{\cdot}0001$).

Because a journal article may contain more than one comparison of an innovation with a standard therapy we repeated analyses in which we used the journal article as the unit of measure rather than the individual comparisons. For each article we look the average gain across all comparisons from that article. These analyses produced practically unchanged results for all measures of gain. Similarly, we repeated analyses controlling for the four disciplines included in this study. In these analyses the results remained unchanged both for the overall findings and for those analyses limited to randomized controlled trials.

DISCUSSION

In this study of 128 evaluations reported in journals selected from four medical disciplines, half of the investigations used a randomized design. The overall likelihood of success reported from studies, of innovations over standard therapy as measured by the Mann–Whitney statistics, was substantially greater for the non-random studies than for the randomized studies. Within the randomized trials, those that did not use a double-blind design reported a significantly greater likelihood of success for innovations on average, than did those that used double-blinding. Also, those trials that compared an innovation in therapy to a placebo control showed significantly greater likelihood of success for the innovation than did trials that compared an innovation to some other form of standard therapy.

These results are consistent with previous work reported by Gilbert, McPeek and Mosteller[6] who observed greater gains for the innovation among surgical trials that used a non-random study design compared with randomized trials. Chalmers *et al.*[3] have also reported this association from their study of anticoagulants in acute myocardial infarction. In their review of the medical literature, these authors identified six randomized controlled trials, eight trials that used non-random allocation to treatment groups, and 18 reports of therapy compared to historical controls. They observed greater gains reported for anticoagulants among less well controlled studies. Wortman and Yeatman[5] observed a similar association in their meta-analysis of the efficacy of coronary bypass graft surgery. Shaikh *et al.*[16] reviewed studies that evaluated the efficacy of tonsillectomy and adenoidectomy, and, after they scored each article for the quality of design and reporting, they concluded that studies with a lower quality score more likely favoured tonsillectomy.

The measures of gain available for an innovation as compared to a standard therapy are fundamental to our analysis. While we estimated these measures of gain, our investigation is limited by the quality of data reported in the evaluations of therapy that we identified and used. We could calculate the Mann–Whitney statistic and the rating of authors' conclusion for all trials, while we could use *p*-values only when reported. Thus our most complete analyses used the Mann–Whitney statistic, which estimates the likelihood that a randomly selected patient would perform better on the innovation than on standard therapy. (A value of 0·5 represents a toss-up between the two therapies.) A discussion of the merits and drawbacks of this statistic is published elsewhere.[17]

A strength of our analysis is the mixing together of evaluations from different areas of medicine so that our findings have wide application. A potential weakness is that studies that use different types of design may not evaluate innovations with the same underlying distribution of gain. Bailar has suggested,[18] for example, that investigators often undertake randomized trials to confirm observations made from methodologically weaker studies undertaken without full understanding of relevant study design factors. If this is the case, then differences between average gains of different study designs may not result from differences in study design (or random variation) but from systematic differences in the underlying distribution of gain. Another potential weakness is the mixing of study designs and disciplines; however, when we repeated analyses that controlled for the disciplines included in our study, the results remained unchanged. Within the studies identified, 87 per cent evaluated drugs and 13 per cent evaluated other interventions (including physical training, group psychotherapy, drug delivery devices, and physical therapy). This distribution differed only slightly and not significantly across disciplines.

We examined the relation between study size and gain to explore the possible effect of this feature of design. The number of subjects in studies was relatively constant, except for observational studies which included more subjects.

We excluded reports whose authors failed to make a comparison with standard therapy in any way other than to report that the new therapy was better. To calculate a measure of gain, we required some quantification of the efficacy of standard therapy. The criteria for inclusion, however, remain explicit: authors must report the outcomes observed for patients on standard therapy and on the innovation. With such data reported, randomized studies show considerably smaller gain than do the results from non-random sequential studies. Even when we exclude studies that used patients who were previous treatment failures, the probability of success on new compared to old therapy, 0·76 for non-randomized studies, was significantly higher than the 0·63 observed for the randomized studies.

Studies that use external controls or an observational design occur rarely in the evaluation of medical therapies. This may reflect, in part, the requirements of the U.S. Food and Drug Administration (FDA) that evaluations of new therapies require randomized controlled trials. The small number of studies in these two categories of design preclude any firm conclusion regarding possible biases encountered with them. In contrast, when we examined with a similar approach the relationship between study design and bias in surgery, 52 per cent of the comparisons identified from six leading journals published in 1983 used external controls or were observational studies (see Part II. Surgery). FDA approval does not generally pertain to surgical procedures, although it may apply to adjuvant therapies, especially pharmacological.

The quality score we used did not include randomization or blindness. Rather, we stratified our analyses on these major features of research design. As a result we have not attained as large a difference in quality scores as we might otherwise obtain. Although this quality score may in part be influenced by the quality of the report, it nevertheless reflects the quality of the investigation as reported. A more precise measure of quality may only be obtained through more thorough investigation of this issue, such as review of study protocols rather than just the final report of the research. We observed, however, a modest correlation between the quality score as measured by the reporting of items that may bear on the quality of the original work, and the impact factor score for the source journals. Our results agree with a report by Bruer,[19] who observed a significant correlation of 0·27 between a methodologic score for journal articles and the impact factor. This is consistent with the notion that studies with stronger designs generally attain publication in more widely read journals, and consequently receive more frequent citations. Because this impact factor did not correlate with the size of the gain reported from the randomized studies, but did correlate with the quality score, the quality of the design rather than the size of the gain reported for the new therapy may influence the review process and so help determine the journal for publication of the article.

CONCLUSION

Fineberg[20] 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 do not notably have more success than weaker forms in the shaping of medical practice. For example, although many cancer researchers hold the opinion that randomized compared with non-randomized trials generally have had more influence in development of medical therapies, accepted treatments for acute leukaemia have more commonly derived from non-randomized rather than randomized trials,[21] and clinical trials have had limited impact on the length of stay for myocardial infarction.[22]

We observed that serveral features of study design influence the likelihood of a report that patients perform better on the innovation than on standard therapy. Those features included randomization, blinding, the use of placebo and the inclusion of patients refractory to standard therapies.

One purpose of this paper was to help readers evaluate findings from studies of various designs. Although not all studies using a given design may have a fixed level of bias, in general we may wish to adjust for, or consider the effect of an adjustment for, the average bias associated with a given design. Using the results of our investigation for this purpose requires some assumptions of comparability that on one hand may be incorrect but on the other may be adequate to give the reader some caution in interpretation of these findings of non-randomized evaluations. Therefore, we have tried to summarize the adjustments one might make when such assumptions are correct. In that spirit we computed the following average adjustments if they are numerally possible. Inclusion of patients refractory to standard therapy increases the likelihood of a positive response to therapy. For studies that use this entry criterion one might reduce the Mann–Whitney statistic by 0·33, and the authors' rating by 0·5. For other non-random sequential studies one might reduce the Mann–Whitney statistic by 0·15. Within the randomized controlled trials, failure to use a double-blind design had an associated increase of 0·11 for the Mann–Whitney statistic; one might reduce the statistic for these studies by this amount. One can consider the effect of adjustment for each of these features when evaluating a report of a new therapy and its possible application in clinical practice. Although we are unsure of the appropriateness of these reductions, merely considering the effect may suitably temper enthusiasm for results based on weaker designs.

One way to think about these adjustments is to ask what features of a published study might make it lead to a smaller bias than the average estimated here for its type. For example, were some special precautions taken that would tend to reduce bias (for example, every patient in a series included or only a subset of the patients, or were the external control data derived from a set of patients who had a similar age range and severity of illness?). For the reader, too, it may be useful to note that in presenting an average bias, we are not usually indicating the worst that can happen. One does not have to believe that the bias reported here is close to its true value for a study in order that its consideration be useful. Just noting that an 8 per cent gain, for example, is less than a baseline bias of 11 per cent offers the practitioner a degree of caution.

APPENDIX: THE MANN–WHITNEY STATISTIC

The Mann–Whitney statistic, denoted schematically by $P(I > S)$, is an estimate of the probability that a randomly selected patient would perform better on the innovation than a randomly selected patient on the standard treatment. Probably the simplest situation to illustrate is for pre/post data. For example, if ten patients on standard therapy were changed to treatment with the innovation and eight improved while two got worse, $P(I > S)$ is estimated as 0·80 (8/10). With matched pairs of patients, the computation is analogous: if eight of ten pairs showed the innovation preferable and two pairs showed the standard preferable, then $P(I > S)$ is again estimated as 0·80. Ties are divided evenly between the innovation and the standard, as if one-half of a patient (or pair of patients) had found the innovation preferable, and one-half had found the standard treatment preferable.

To estimate $P(I > S)$ directly from data for two groups, we count the proportion of all possible comparisons between outcomes on the innovation and on the standard that favour the innovation. If M patients received the innovation and N patients received the standard, there would be $M \times N$ such comparisons. The same computation method may be used to compute $P(I > S)$ from a survival graph.

The formula for estimating $P(I > S)$ from the proportions of treatment failures is $0{\cdot}5 + 0{\cdot}5\,(p_S - p_I)$, where p_S and p_I are proportions of treatment failures on the standard and on the innovation, respectively.

Continuous data require a different approach. If we let '*d*' equal the difference between a randomly selected patient's score on the innovation and a randomly selected patient's score on the standard, then $P(I > S)$ is an estimate of the probability that d is greater than zero. $P(I > S)$, then, is computed by using as a normal deviate the standard score which is the difference in average scores for the innovation and the standard divided by the standard deviation of this difference. For matched pairs samples, the standard deviation is estimated from the differences in scores for each pair. $P(I > S)$ is then read from a standard normal table as the probability of a deviate larger than the observed standard. For independent samples, the standard deviation is computed as the square root of the sum of the variances of means of scores for the innovation and for the standard.

For example, if scores for both the innovation and the standard treatment were independently drawn from normal populations with variance equal to 1, the following expected values of $P(I > S)$ would correspond to stated differences between the means of the distributions.

Difference in means	$P(I > S)$
0	0·50
0·5	0·64
1	0·76
1·5	0·86
2	0·92
3	0·98

The first line shows that if the difference between the means of the distributions were zero, then the standard and the innovation are equally likely to be the better performer. As the difference in means of innovation over the standard increases, the probability increases that a patient would do better on the innovation than on the standard treatment.

ACKNOWLEDGEMENTS

The study Group on Quantitative Methods in the Evaluation of Medical Technology provided valuable input to the work reported here and in Part II. P. Lavori assisted in design of this study and the development of the classification of study designs. The contribution of the readers who participated in this study is gratefully acknowledged: Anna Angelos; Andrew L. Avins; Jesse Berlin; Karen Bierstein; Christina Braun; Ruth Cogan; Timothy R. Cote; Katherine M. Coughlin; G. Wade Davis; Mary Ettling; Joy R. Esterlitz; Manuel Friere; Paul Gompers; Joan Gopin; David Hunter; Katheryn Lasch; Ann Levin; Mark McClellan; Alesia Maltz; Rama Peri; Laura Rosen; Heather Sacks; Lory Stapsy; Ellen Velie; Margaret Stassen provided invaluable assistance in the final compilation of adjudicated data.

This work was supported by the National Center for Health Services Research and Health Care Technology Assessment Grant No HSR 1 R01 HS-05156-01 and HS-05936; and by the Macy Foundation and the Rockefeller Foundation.

REFERENCES

1. Committee for Evaluating Medical Technologies in Clinical Use. 'Methods of technology assessment', in *Assessing Medical Technologies*, National Academy Press, Washington D. C., 1985, pp. 70–175.
2. Chalmers, T. C., Block, J. B. and Lee, S. 'Controlled studies in clinical cancer research', *New England Journal of Medicine*, **287**, 75–78 (1972).
3. Chalmers, T. C., Matta, R. J., Smith, H. Jr. and Kunzler, A. M. 'Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction', *New England Journal of Medicine*, **297**, 1091–1096 (1977).
4. Sacks, H. S., Chalmers, T. C. and Smith, H., Jr. 'Randomized versus historical assignment in controlled clinical trials', *New England Journal of Medicine*, **309**, 1353–1361 (1983).
5. Wortman, P. M. and Yeatman, W. H. 'Synthesis of results in controlled trials of coronary artery bypass graft surgery', in Light, R. (ed) *Evaluation Studies Review Annual*, Vol. 8, Sage Publications, 1983, pp. 536–557.
6. 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.) *Cost, Risks, and Benefits of Surgery*, Oxford University Press, 1977, pp. 124–169.
7. Hovell, M. F. 'The experimental evidence for weight-loss treatment in essential hypertension: A critical review', *American Journal of Public Health*, **72**, 359–368 (1982).
8. Sackett, D. L. 'Evaluation: Requirements for clinical application', *in* Warren, K. S. (ed.) *Coping with the Biomedical Literature: A Primer for the Scientist and the Clinician*, Praeger, New York, 1981.
9. Cochran, W. G. *Planning and Analysis of Observational Studies*, Wiley, New York, 1983.
10. Colditz, G. A., Miller, N. J. and Mosteller, F. 'The effected study designs on gains in the evaluations of new treatments on medicine and surgery', *Drug Information Journal*, **22**, (1988) (in press).
11. Garfield. E. (ed). 'Journal Citation Reports: A bibliometric analysis of science journals on the 1st data base', Institute for Scientific Information, Philadelphia, 1980, vol. 14, p. 12A.
12. Bailar, 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).
13. 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).
14. Lochaya, S., Thongmilr, V. and Suvachittanont, O. 'Antihypertensive effect of BS100-141 a new central acting antihypertensive agent', *American Heart Journal*, **99**, 58–63 (1980).
15. Josephson, G. W., Kennedy, H. L., MaKenzie, E. J. and Gibson, G. 'Cardiac dysrhythmias during treatment of acute asthma. A comparison of two treatment regimens by a double-blind protocol', *Chest*, **78**, 429–435 (1980).
16. Shaikh, W., Vayda, E. and Feldman, W. 'A systematic review of the literature on evaluation studies of tonsillectomy and adenoidectomy', *Pediatrics*, **57**, 401–407 (1976).
17. Colditz, G. A., Miller, J. N. and Mosteller, F. 'Measuring gain in the evaluation of medical technology: The probability of a better outcome', *International Journal of Technology Assessment Health Care*, (1988) (in press).
18. Bailar, J. C. 'Research quality, methodologic rigor, citation counts, and impact', *American Journal of Public Health* **72**, 1103–1104 (1982).
19. Bruer, J. T. 'Methodologic rigor and citation frequency in patient compliance literature', *American Journal of Public Health*, **72**, 1119–1123 (1982).
20. Fineberg, H. V. 'Effects of clinical evaluation on the diffusion of medical technology', in *Assessing Medical Technologies*, National Academy Press, Washington, D.C., 1985, pp. 176–210.
21. Gehan, E. A. 'Progress in therapy in acute leukemia 1948–1981: Randomized versus non-randomized clinical trials', *Controlled Clinical Trials*, **3**, 199–208 (1982).
22. Chassin, M. P. 'Health Technology Case Study, 24: Variations in hospital length of stay: Their relationship to health outcomes', OTA-HCS 23, Washington, D.C., US Congress, Office of Technology Assessment.