*Statistical Science*
1992, Vol. 7, No. 2, 227-236

# Some Progress and Problems in Meta-Analysis of Clinical Trials

**Frederick Mosteller and Thomas C. Chalmers**

*Abstract.* We report progress on some methodological issues in meta-analysis. Evidence continues to accumulate that randomized trials show smaller gains than nonrandomized trials when innovations are compared to standard therapies. Quality scores for randomized clinical trials show that reporting has improved about 27% in three decades, to a quality level slightly over 50%. Although quality scoring could be useful, in principle, for adjusting estimates of gain from innovations, a substantial study has not found a statistical relation between gains and quality. We describe a method of blinding papers to reduce the bias of readers doing meta-analyses. For combining data for fixed effects, Greenland and Salvan recommend using Mantel-Haensztel, weighted least squares or maximum likelihood methods. For random effects, Larholt, Tsiatis and Gelber have improvements for the DerSimonian and Laird method. Eddy and his colleagues have prepared software and book-length works on Bayesian methods for technology assessment using meta-analysis. Louis has a valuable review article on Bayesian approaches. The annoying difficulties in combining $2 \times 2$ tables when some cells have zeros has been largely overcome by exact calculation methods. From diagnostic data acquired from several independent investigations, new methods have appeared for estimating receiver operating characteristic curves. An update on meta-analyses of randomized clinical trials shows about 16 meta-analyses per year in journals during 1983–1990. We expect much more methodologic work as new issues appear and findings point us toward fresh solutions.

*Key words and phrases:* Clinical trials, overviews, combining data, quality scores, Bayesian methods, exact methods, ROC curves, technology assessment.

## INTRODUCTION

A short review paper tends to lean in directions of special interest to the authors. We have restricted ourselves to issues of methodology rather than substantive advances, but even among these, many advances would require much space to present. In the field of statistical analysis, so many papers appeared in 1985–1990 that we could not handle all the advances. Ingram Olkin (editor) offers an overview by way of introduction to this material, and, in a separate paper, Lawrence Hedges treats publication bias.

We have concentrated, but not exclusively, on quantitative methodological issues flowing from an empirical base. Our topics include effects of study design on outcomes in clinical trials, findings from quality scoring in randomized clinical trials (RCTs), an illustration from an historically important meta-analysis by Beecher, problems of reliability in recording results, issues of small versus large clinical trials, issues in combining data from collections of comparative studies and of diagnostic studies, the problem of multiple looks in cumulative meta-analysis of sequences of clinical trials on the same topic and an update on presentation of meta-analyses through 1990.

*Frederick Mosteller is Director and Thomas C. Chalmers is Associate Director of the Technology Assessment Group at the Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115.*

## PART I. PRIMARY DATA FOR META-ANALYSIS

### Effects of Study Design on Outcomes in Assessing New Treatments in Medicine and Surgery

Colditz, Miller and Mosteller (1989) and Miller, Colditz and Mosteller (1989) carried out substantial analyses relating study design to magnitude of gain of an innovative therapy over the standard therapy with the general intent of developing adjustments for the effects of study design. Thought of in terms of bias, the idea would be to produce a consideration for adjustment such as a report to the investigator that, "You have observed innovative treatment A to be 10% better than standard treatment B using method of investigation M. Method M seems to give innovations an average of 15% improvement, and so until stronger data come along, you might consider as serious the possibility that with only a 10% gain the innovation actually gives a 5% loss." To get the data, these studies used two readers and analyzed 113 medical reports published in a sample of medical journals in 1980 and 221 surgical reports published in six leading surgical journals in 1983.

**Medical Comparisons.** In the medical area, Colditz, Miller and Mosteller (1988, 1989) made comparisons between the innovation and the standard by using the Mann-Whitney (M-W) statistic. It tells the proportion of the time that a random individual drawn from the innovation treatment group has a better outcome than a random individual drawn from the standard treatment group (or control group), where ties count 0.5. Thus, an M-W of 1 means the innovation always performed better than the standard, while a score of 0.5 means that the two treatments performed equally well.

TABLE 1
*Mann-Whitney statistic and the rating of the authors' conclusion among a sample of evaluations of medical therapies*

| Study design | Number of studies | Mann-Whitney statistic mean | Rating of authors' conclusions* |
|---|---|---|---|
| Randomized control trials (parallel) | 36 | 0.61 | 4.4 |
| Randomized control trials (crossover) | 29 | 0.63 | 4.6 |
| Nonrandom comparisons | 46 | 0.81 | 4.9 |
| Refractory | 12 | 0.94 | 4.9 |
| Other | 34 | 0.76 | 4.9 |

* Average score on a scale of 1 to 6, running from 1 (*innovation much worse than standard*) to 6 (*innovation much better than the standard*); 3 and 4 mean that treatments perform about equally, but 3 means the innovation is a disappointment (for various reasons—cost, training required, side effects), while 4 means it is a welcome additional choice of treatment.

*Source:* Colditz, Miller and Mosteller (1989). Adapted and reprinted by permission of John Wiley and Sons, Ltd.

TABLE 2
*Percentage of reporting individual items in the medical study designs of the original investigations*

| | RCT (parallel), $n = 36$ | RCT (crossover), $n = 29$ | Nonrandom, $n = 46$ |
|---|---|---|---|
| Eligibility criteria | 86% | 86% | 95% |
| Completeness of admission | 22 | 11 | 12 |
| Admission before allocation | 61 | 50 | 49 |
| Methods of allocation | 17 | 11 | 37 |
| Loss to follow-up | 94 | 46 | 72 |
| Statistical analysis | 95 | 79 | 88 |
| Statistical methods | 78 | 75 | 70 |

*Source:* Colditz, Miller and Mosteller (1989). Reprinted by permission of John Wiley and Sons, Ltd.

Table 1 shows results for study designs in medicine based on at least 10 studies. The randomized studies, whether cross-over or parallel, gave very similar results. But studies where the choice of treatment was left open to the investigator showed very different M-W statistics. When the patient had already failed with a standard (refractory patient), the nonrandom innovation scored a very high M-W, namely, 0.94. For trials where the patient was not refractory, the comparisons still gave innovations high M-W's: 0.76, or 0.14 higher than the average for the two randomized sets of trials. Thus, the nonrandomized trials in medicine showed substantial evidence of bias, highest when patients were refractory.

**Reporting.** Seven issues of reporting were rated as shown in Table 2. The rates of reporting on these issues for randomized and nonrandomized trials was not substantially different. Two areas needing better reports were the issue of admission to the study before allocation and explanation of the method of allocation, the latter especially in the randomized trials.

**Blinding.** Table 3 shows use of blinding in various kinds of studies. For double-blind trials the M-W statistic was 0.58, and for non-double-blind trials it was 0.69, a significant difference favoring the innovations.

TABLE 3
*Use of blinding in the design of medical therapies*

| | Number of studies | Double blind | Not double blind |
|---|---|---|---|
| RCT (parallel) | 36 | 21 | 15 |
| RCT (cross-over) | 29 | 13 | 16 |
| Nonrandom | 46 | 1 | 45 |
| External controls | 9 | 0 | 9 |
| Observational studies | 5 | 0 | 5 |

*Source:* Colditz, Miller and Mosteller (1989). Adapted and reprinted by permission of John Wiley and Sons, Ltd.

**Placebo.** When a placebo control was used the M-W was 0.72, and for nonplacebo control 0.61. This probably means it is easier for a treatment to do better than a placebo than to outperform an active standard treatment; confusion arises here because many studies without a placebo still have a no-treatment or standard treatment control group. The issue is presumably whether the study compares two active treatments, or one thought to be active and one thought to be less active. In RCTs placebos are used to disguise treatments, not to withhold them.

**Bias.** Any adjustment depends on assumptions about the comparability of studies. Such adjustments must be used with caution, but for refractory patients included in nonrandom studies an M-W might be reduced by 0.33. For nonrandom unrefractory patients, reduce by 0.15, and for non-double-blind, reduce by 0.11. The idea is not so much to make the correction and believe in it as to temper enthusiasm by considering the implication of such an adjustment.

**Surgical Comparisons.** Nonrandomized studies tended to report larger gains than did randomized studies, consistent with our medical results and with those of other studies.

Earlier results by Gilbert, McPeek and Mosteller (1975) had found very small improvements on average for innovations in surgery, 1.3% for primary and 0.4% for secondary studies (based on treatments intended to ameliorate the effects of the surgery itself), whereas in Miller, Colditz and Mosteller (1989) the gains averaged 12.5% and 6.0%, respectively. [The samples drawn by Gilbert, McPeek and Mosteller (1975) came from the National Library of Medicine Literature Retrieval System for papers published before 1977, while those from Miller, Colditz and Mosteller (1989) came from searches in six surgical journals in 1983.]

### Quality Scores for Original Clinical Studies

Chalmers et al. (1981) have developed methods for scoring the quality of clinical trials that use randomization. These assessments deal with three major areas regarded as important for good quality: study design, implementation and analysis. Two readers make the ratings in blind fashion—blind to authors, source, results and discussion—in order to mask readers from potential sources of bias. When a particular item could not be carried out (e.g., blinding to amputation), the points allocated to it were set aside and the final percentage score was based on the reduced number of points.

If the quality scores could be related to the effects found in the studies, then they might be used to adjust the findings of the primary studies, possibly by giving the better quality studies more weight or reducing the observed treatment effect of weaker studies. To investigate this possibility, Emerson et al. (1990) reviewed quality scores previously assigned in seven published meta-analyses based on 107 randomized clinical trials that compared pairs of treatments, and they related the scores to the sizes of treatment differences. They found no statistically significant relations

- between treatment differences and overall quality scores
- between quality score and variation in treatment difference
- between treatment differences and components of quality scores

But they did find a statistically significant increase in quality scores of 9% per decade for three decades, averaging 0.51 on a scale from 0 to 1 in the 1980s. This still leaves lots of room for improvement.

Naturally these findings may not hold in some other areas of medicine or if new forms of quality scores are created. Nor do they suggest that poorly carried out studies are just as good as well-executed ones. We note that all these studies were judged strong enough and usable enough to be included in a meta-analysis and that each meta-analysis was published as well. Consequently some screening hurdles of the publication system have been passed by the studies included in these meta-analyses. If quality scores useful for adjusting the findings of the meta-analyses are found, they could be very useful. Meanwhile, they serve as a measure of progress in design and reporting, and they facilitate sensitivity analyses if the individual studies are presented to the reader in the order of estimated quality.

In their monumental two-volume collection of some hundreds of meta-analyses in obstetrics, Chalmers, Enkin and Keirse (1989) present the outcomes for each study included in a meta-analysis. When they judged that the studies were of similar methodological quality, they listed the studies by year of publication, the most recent at the top; otherwise they list in order of descending quality assessed by their own methods (the Chalmers here is Iain rather than Thomas C.). Thus a reader could readily consider what the effect might be of setting aside the weaker studies.

The hope is that similar volumes will appear in other areas of medicine.

### The Powerful Placebo

Among the early meta-analyses in modern medicine was a methodological study by the anesthesiologist Henry K. Beecher of Massachusetts General Hospital in Boston. He made many contributions to medicine and led many studies of analgesics and anesthetics. He led the first American Committee to help decide when a patient was brain dead.

Beecher's 1955 paper was entitled "The Powerful Placebo." We include this example for more than its historical interest because it is seemingly simple and

yet provides an illustration of several important issues often brought up in meta-analyses: the apples and oranges problem, the issue of heterogeneity versus homogeneity and the question of appropriate models such as fixed or random effects. It also illustrates an attempt to make a broader generalization than the more usual comparison of efficacies of two treatments for a disease, and thus brings up the apples and oranges problem in a more severe form than usual. The usual complaint is that every study is not carried out in exactly the same way. Here we deal with entirely different sources of pain. Beecher was impressed with how often and how substantial the contribution of placebos was to the relief of patients' symptoms in a variety of circumstances. To show this, he gathered all the data he could find on this subject and assembled it into a table to give a notion of the magnitude of the effect of placebos. We thought these findings given in Table 4 would be relevant, not only for their own sake, but because they bear on the use of placebos in medical investigations.

Beecher (1955) himself says, "It is evident that placebos have a high degree of therapeutic effectiveness in treating subjective responses, decided improvement . . . being produced in 35.2 ± 2.2% of cases. This is shown in over 1,000 patients in 15 studies covering a wide variety of areas: wound pain, the pain of angina pectoris, headache, nausea, phenomena related to cough and to drug-induced mood changes, anxiety and tension, and finally the common cold, a wide spread of human ailments where subjective factors enter" (p. 1606). We have abbreviated his table to show his findings.

He also says that "whenever judgment is a component of appraisal of a drug or a technique, and this is often the case, conscious or unconscious bias must be eliminated by the procedures just mentioned [which were, among others, the use of a placebo, double-blindness, and randomization]" (p. 1606). We think Beecher's paper led investigators to increased use of placebos for controls.

One of the present authors (Mosteller) recalls that in preparing the data for the placebo responses, Beecher was uneasy about taking an average of the percent relieved across the several conditions, essentially because he thought critics would complain about an "apples and oranges" problem. One way to think about the issue is to generalize the question asked in a specific set of disorders to an overall placebo effort in the relief of pain – or even more generally.

Some conditions in Table 1 lead to percent relief in the 30's, others in the 50's; thus we can readily imagine that there may be inhomogeneity in placebo effectiveness from one condition to another. We can readily believe that the placebo has positive effectiveness because of the consistency and size of the effects, but we may wish a better summary of the variation of results across conditions.

### Reliability of Assessing Studies in Meta-Analyses

Because meta-analysis and reanalysis of papers on whether corticosteroid drugs cause peptic ulcers led to differing conclusions in separate analyses (Chalmers, 1987), Chalmers et al. (1981) developed a method for strengthening the reliability of the analyses. In brief, they believed they found a tendency toward bias in the selection of papers for inclusion in the meta-analyses and tried to create procedures that would reduce such effects. Table 5 outlines their program. In brief, they select papers liberally for consideration, blind the readers to source and findings to reduce bias, rate the quality, use two readers to decide inclusion and adjudicate the final quality score.

More than half the disagreements between raters occur because one of the raters did not find facts in the paper that the other one found. For this reason, we ask raters to record exactly where in the paper they found the basis for their rating. This record speeds up final adjudication.

TABLE 4
*Placebo effectiveness (after Beecher)*

| Condition | Number of studies | Relieved (%) |
|---|---|---|
| Severe postoperative wound pain | 5 | 32 |
| Cough | 1 | 40 |
| Mood changes | 1 | 30 |
| Angina pectoris | 3 | 34 |
| Headache | 1 | 52 |
| Seasickness | 1 | 58 |
| Anxiety and tension | 1 | 30 |
| Common cold | 1 | 35 |
| Average | | 35 |

*Source:* Adapted from Beecher (1955).

TABLE 5
*Ratings and choices for papers included in a meta-analysis*

1. Choose papers liberally for possible inclusion in the final meta-analysis.
2. Blind the readers to source information and findings including treatment and control.
3. Two readers independently decide on inclusion or exclusion of paper, rate the quality of the methods section according to a protocol and extract the blinded results.
4. The results section is scored for quality from the unblinded paper, and the extracted results are identified.
5. The readers consult to agree on a final quality score, starting with the blinded papers and consulting the unblinded when necessary.

## PART II. STATISTICAL CONSIDERATIONS

### Meta-Analysis of Many Small RCTs Versus Very Large Studies—Which Is the Gold Standard?

Arguments abound as to whether clinical decisions should be based more on meta-analysis of many small studies (Chalmers et al., 1987a) or rather on one or more very large studies (Yusuf, Collins and Peto, 1984). Arguments can be made for either, and it may be that one should strive for both. Small studies carried out in many clinics will have a heterogeneity that has some benefits of reality, because the situation mimics what is encountered in day-to-day practice. Large studies require adherence to one protocol and may therefore produce a more reliable answer to the question of interest, but that question may be too narrow, or restrictions to the protocol may result in the admission of a very small percentage of patients presenting themselves for treatment. (Very large studies are actually cooperative participations by different clinics, and the methods of analysis should take that into account rather than rely solely on crude pooling, though this allowance for variation between clinics has usually been ignored.)

An empirical approach to the problem of many small versus one or more large studies would examine those situations whose answers have been obtained both by the meta-analysis and big study techniques. We know of only four such instances:

1. The comparison of streptokinase with a placebo or no fibrinolytic treatment in patients with acute myocardial infarction (AMI) has been carried out in several small studies, combined in three different meta-analyses (Stampfer et al., 1982; Yusuf et al., 1985; Chalmers et al., 1987b) and in two very big studies (GISSI, 1986; ISIS-2, 1988), and the results are almost identical, a highly significant reduction in hospital mortality of over 20%.

2. The comparison of mixed beta-blocking drugs with no blocking of sympathetic receptors in AMI patients revealed the same reduction in death rate in two meta-analyses (Yusuf et al., 1985; Chalmers et al., 1987b) as in two large studies (MIAMI, 1985; ISIS-1, 1986). The difference was that the confidence intervals were narrower in the big studies, and the observed difference became statistically significant. (Of course, a meta-analysis ultimately would include both the small and the large studies.)

3. Meta-analysis of multiple small RCTs of photography for neonatal hyperbilirubinemia revealed it to be highly effective for both treatment and prevention (Chalmers et al., 1988) long before a big trial begun by the National Institutes of Health found effects of the same size.

4. A possible current exception to agreement is a comparison of the new and expensive genetically derived thrombolytic drug tPA with streptokinase in AMI. Four small RCTs all favor the new drug, but the difference is not quite significant, and reports from two big trials of over 10,000 patients each made it extremely unlikely that any clinically important difference could exist. This difference may be an example of publication bias or some other factor not yet established such as the difference in timing of accompanying heparin therapy.

The bottom line is that we need many more comparisons of results for pooled small studies to results from large studies.

### Methods of Combining Data

**Fixed Effects.** Many meta-analyses compare the performance of two treatments or that of a treatment to that of a placebo. Frequently the outcome for a single patient is regarded as a success or a failure, and the outcomes for a study then essentially form a $2 \times 2$ contingency table. Each treatment has a number of successes and failures, and the two independent treatment groups are regarded as equivalent because of random assignment to treatment group. Each study may then produce a separate $2 \times 2$ table, and the outcomes are to be combined by one or another method depending upon the assumptions being used.

In summarizing data from several $2 \times 2$ contingency tables, the Mantel-Haenszel method and the one-step Peto method might be used in fixed-effects situations. Greenland and Salvan (1990) have compared these methods. They report that in many common circumstances the Peto method turns out to have substantial biases. This difficulty occurred when each $2 \times 2$ table was badly balanced. It also occurred when the effects were substantial. In some instances the Peto statistic fell outside the 95% range of the Mantel-Haenszel method. The latter gives results close to those of conditional maximum likelihood.

**Random Effects.** New work on combining information from several sources for meta-analysis looks at the effects of standard methods. In her doctoral dissertation, Larholt (1989), Larholt and Gelber (1989) and Larholt, Tsiatis and Gelber (1989) reviewed methods of computing confidence limits in the fixed-effects model and random-effects model by examining the results of simulations. They considered especially binomial situations for both treatment and control groups, analyzing the log of the odds ratio. In the fixed-effects model, they used a version of Mantel-Haenszel given by Yusuf et al. (1985) (Method I), and for the random-effects model they used approaches related to DerSimonian and Laird (1986) (Method II). One reason for concern arises because the DerSimonian and Laird approach assumes that the weights for the studies are known, whereas they actually have to be estimated from the data in the studies. The resulting uncertainty

in the weights could and, they find, does affect the coverage of the confidence intervals.

As a first step, Larholt and Gelber (1989) compare the coverage achieved with Method I and Method II when the actual effects were fixed or random. For fixed-effect situations Method I gave approximately 95% coverage with equal sample sizes even with heterogeneous but fixed effects. When sample sizes were unequal, running from 50 to 200 in seven studies, the Method I coverage still ran close to 95% when the weighted average (by sample sizes) of the log odds ratio was taken as the true parameter (but not when the parameter was taken as the equally weighted average).

When Method II was applied to the fixed effects situation with equal sample sizes, the coverage averaged 96.5% instead of 95%. As the fixed effects became more heterogeneous, the coverage by Method II increased even more, as it also did for unequal sample sizes.

When the random-effects model was the true situation, as the between-studies variance component increased, Method I gave lower coverage, as one might expect, down to 57% instead of 95% in one example. Method II did not maintain its coverage of more than 95%, and as the between-variance component exceeded the within-variance component, the coverage by Method II gradually decreased to about 90% instead of 95% whereas Method I gave 57% coverage.

In a paper in the dissertation, Larholt, Tsiatis and Gelber (1989) developed methods that considerably improve the coverage of Method II by using a $t$-distribution with degrees of freedom based on an approximation suggested by Satterthwaite. The details are too extensive to give here. For large numbers of studies, say more than 20, they suggest that the special methods may not be needed.

**Fixed, Random and More General Effects.** For fixed effects, Greenland and Salvan (1990) recommended Mantel-Haenszel, weighted least squares, or maximum-likelihood methods when enough data are available.

Greenland and Salvan's (1990) view about fixed effects versus random effects is that, to the extent that study odds ratios (or risk differences) are alike, the fixed versus random analyses will be very similar. To the extent that they are heterogeneous, they think that we should be modeling the study differences instead of offering a single summary. Of course, it is an open question how we should do the modeling and summarizing.

If we have a well-established covariate, in the sense that the profession agrees on the basis of evidence that it should be used for adjustment, then using the covariate to remove variation seems desirable. If, however, we are working in an exploratory mode, running through many possible covariates in order to pick a few that explain variance, it is hard to say what the final answer means except that exploration has suggested a few possible useful variables for future verification.

Turning back to types of effects, from an analysis of variance point of view, situations are usually regarded either as having fixed effects or random effects, though these are not the only possibilities. We fear that some investigators prefer the fixed-effect approach because it gives narrower confidence limits rather than because they want to apply their inferences to the particular population sampled. Another way of speaking about this is to talk of getting a significant result sooner as one accumulates results from more and more studies. This sort of consideration, of course, should come up in the general context of the decision to declare one treatment superior to another. It could well be that the 5% significance level heavily used by the Food and Drug Administration (FDA) in the United States is not a good universal choice, though in the medical arena we often behave as if it is, perhaps because it seems better to have some standard than none. From the point of view of FDA, unless they have such a standard, the courts may regard them as not being fair to the various pharmaceutical companies.

The random-effects model uses a two-stage sampling idea, as if we sampled from a superpopulation of studies that might be carried out and then sampled patients within the studies. Of course, we almost never do such sampling. The real situation is more like a selection of studies that can be carried out. For example, we do not do studies in institutions that do not cooperate.

If we use the random effects model, we are presumably extending our inference to the superpopulation of which the studies are a sample rather than to the ideal superpopulation of all studies. For further discussion see Laird and Mosteller (1990). The main point is that the actual variability associated with our inference in real studies and meta-analyses is almost certainly larger than that given even by the usual random-effects models. This is a routine fact about science generally when we are not in position to list our populations and sample them. It is not a special feature of meta-analysis or medicine, but it is a feature of most sciences that often deal with the real world by using observational studies, such as geology, meteorology, astronomy and biology.

**Bayesian Approaches.** Eddy and his colleagues have developed a substantial Bayesian approach called the confidence profile method for carrying out meta-analyses for comparative clinical trials and other sorts of investigations where studies are to be combined. Their paper (Eddy, Hasselblad and Schachter, 1990) won first prize in the 1990–91 FHP Prize competition supported by the FHP Foundation and awarded by the International Society of Technology Assessment in

TABLE 6
*Schematic outcome for test outcome compared to actual patient state*

| Test | Patient state | |
|---|---|---|
| | Positive | Negative |
| + | $a_k$ | $b_k$ |
| − | $c_k$ | $d_k$ |

Cell entries are counts from the $k$th cut-off point, or from the $k$th study.

Health Care. They have written up their work in substantial volumes (Eddy, Hasselblad and Schachter, 1992; Eddy, 1992) and have prepared software (Eddy and Hasselblad, 1992) for carrying out this work.

Louis (1991) has provided a paper on empirical Bayes methods valuable both for its description of theory and illustrations and for its extensive references. An early application of these methods, to medical meta-analysis, appears in Gilbert, McPeek and Mosteller (1977).

**Exact Calculations.** In handling several $2 \times 2$ tables by approximate methods, difficulties often emerge because some cells are empty or full (contain all the counts in a row or column). Recent literature has provided practical solutions for such problems by giving exact computational methods that can be executed in reasonable amounts of time: for confidence intervals for the common odds ratio (Mehta, Patel and Gray, 1985), for exact logistic regression (Hirji, Mehta and Patel, 1987) and for matched case-control studies (Hirji, Mehta and Patel, 1988). Vollset, Hirji and Elashoff (1991) have suggested some further improvements.

## Receiver Operating Characteristic Curve

We have not had very good ways of combining information from investigations of a diagnostic test so as to produce a single receiver operating characteristic curve (ROC), but some new work has appeared.

Let us review briefly the notion of an ROC curve. When a test for a disease is applied to a population, we may find results as indicated in Table 6. The letters $a_k$, $b_k$, $c_k$ and $d_k$ indicate the numbers of cases falling into each cell of Table 6.

The proportion of true positives detected, $a_k/(a_k + c_k)$, is called the sensitivity of the test, and the proportion of true negatives detected, $d_k/(b_k + d_k)$, is called the specificity of the test.

When the outcome of testing is based on continuous measurement, or an ordered scale, a cut-off point may be used to define whether the test will be declared to be positive or negative. Each cut-off point may produce a different table of the form given in Table 6, and the subscript $k$ is intended to index such cut-off points. To summarize the outcome of these different cut-offs, investigators may use an ROC curve. It plots the true positive rate (sensitivity) against the false positive rate (1 − specificity) as the cut-off runs through its possible values. This leads to the kind of curve illustrated in Figure 1.

Suppose that several studies 1, 2, . . . , $K$ give us counts as shown in Table 6 relating the actual state of the patient according to some standard with the outcome of a diagnostic test. Then it would be desirable to estimate an ROC curve for the test based on the data from the $K$ $2 \times 2$ tables. To that end, Littenberg, Moses and Rabinowitz (1990) define slightly adjusted log odds (to avoid zeros as follows):

$$\hat{U}_k = \ln \frac{c_k + \frac{1}{2}}{d_k + \frac{1}{2}}$$

$$\hat{V}_k = \ln \frac{a_k + \frac{1}{2}}{b_k + \frac{1}{2}}$$

and regress

$$\hat{V}_k - \hat{U}_k \text{ on } \hat{V}_k + \hat{U}_k$$

They then fit a straight line to the points either with a resistantly fitted line or with weighted least squares. From this fitted line they retrieve an ROC curve.

They provide formulas for getting back to the ROC curve from the fitted line and give information about the uncertainty of the curve. Moses kindly called our attention to a paper by Kardaun and Kardaun (1990) that anticipates this work. Kardaun and Kardaun's instructive paper illustrates the method on practical problems and provides information about the estimation theory.

## An Issue in Accumulating Evidence: Multiple Looks

As comparative trials of the same procedure accumulate, the question arises whether anything should be done about the "multiple looks" phenomenon. In cumulative meta-analysis, we are in a situation rather

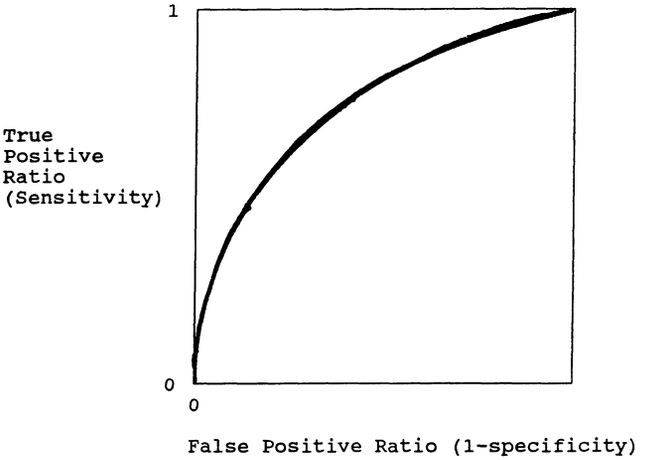


FIG. 1. *Example of ROC curve.*

different from that of a sequential clinical trial. In the sequential trial, we are deciding whether to stop the trial and declare an outcome in favor of one or another treatment. In cumulative meta-analysis, we have little control over the introduction of new trials and can merely record what they produce, adjoin their results to those already on hand and keep reporting the cumulative situation. Possibly, announcements made on the basis of these analyses can influence the start-up of future investigations, but the findings do not set policy like an FDA acceptance of a drug for use in treatment. Perhaps it is reasonable merely to accumulate evidence and report the size of the observed effect and its P value or some equivalent and not worry about the multiple looks phenomenon.

To have a multiple looks problem there has to be some sort of stopping rule or decision rule that depends on the findings, and to evaluate the effect of multiple looks we need the details of the rule. In the rather open-ended meta-analysis situation we do not have specific mathematical rules. One reasonable position is merely to report on the basis of the sampling done so far, and regard estimates of values as based on samples of information cumulated thus far (or of a suitably weighted and analyzed meta-analysis).

It is true that various groups may then be concerned about the information accumulated, such as the FDA (in the United States), an investigator considering starting a new trial or an institutional review board considering whether to approve an investigator's proposal for a further study. The investigator or the institutional board might want to consider the potential impact of the study when its findings are adjoined to those of all the studies done so far. The investigator or the board could be concerned about the value of one more look, though without a model of the rate of new looks and their possible consequences it will be hard for them to assess the future.

To promote discussion of the issue, we raise the question whether any special analysis is required in routine accumulation.

## PART III: FEATURES OF PUBLISHED META-ANALYSES

### Update of Meta-Analyses Based on Randomized Control Trials

Sacks et al. (1987) published a description of 86 meta-analyses based on RCTs. For the second edition of *Medical Uses of Statistics* (Bailar and Mosteller, 1992), they have prepared an update based on 78 additional meta-analyses published before December 1989. Their earliest paper was Beecher (1955) on the powerful placebo.

The 164 meta-analyses found were scattered among over 50 journals, only 10 of which published more than two meta-analyses. The *American Journal of Medicine* had 14, *Journal of the American Medical Association* 12, *Lancet* 11, *New England Journal of Medicine* 8, *Gastroenterology* 5, *American Psychologist* 4 and three journals (*Australia and New Zealand Journal of Psychiatry, Annals of Internal Medicine* and *Cancer*) had 3 each.

TABLE 7
*Adequacy of selected quality features among meta-analyses*

| Features | 1955–1982 | 1983–1986 | 1987–1990 |
|---|---|---|---|
| Number of meta-analyses | 40 | 66 | 58 |
| | Percent adequate | | |
| Design | | | |
| Literature search | 25 | 36 | 69 |
| Treatment assignment | 95 | 26 | 79 |
| Combining | | | |
| Criteria for inclusion | 43 | 39 | 67 |
| Measurement | 13 | 26 | 47 |
| Statistical analysis | | | |
| Statistical method | 55 | 61 | 78 |
| Confidence interval | 35 | 41 | 84 |

*Source:* Sacks et al. (1992). Reproduced with permission from the *New England Journal of Medicine.*

Sacks et al. (1992) reviewed the adequacy of design, analysis and reporting procedures in considerable detail. They report the percent of meta-analyses regarded as giving adequate attention and reporting for each of 23 items. Table 7 shows their findings for a selected six of the 23 items. We know that computer searches alone still find less than two-thirds of the relevant trials, and so searches of reference lists and inquiries of experts are very helpful. How treatments were assigned can be important because historical controls frequently give biased results.

The criteria for including and excluding trials in the meta-analyses should be given in some detail as well as information about the measurements used to compare the effectiveness of treatment.

Sacks et al. (1992) required that some standard method be used to combine the data [usually Mantel-Haenszel (1959) was chosen], and they preferred confidence intervals to point estimates of sizes of effects because of the additional information supplied.

Our discussion of progress and problems in the meta-analysis of clinical trials will need extension because other issues will arise as further experience is gained. The technique is so important in the process of applying clinical research to practice that efforts to improve its execution and presentation are bound to continue at a high level.

## ACKNOWLEDGMENTS

We thank John Emerson, David Hoaglin, Stephen Stigler, Ingram Olkin and three other reviewers for their helpful comments. Preparation of this manuscript was supported in part by Grant HS 05936 from the Agency for Health Care Policy and Research to Harvard University.

## REFERENCES

BAILAR III, J. C. and MOSTELLER, F. (1992). *Medical Uses of Statistics*, 2nd ed. NEJM Books, Boston.

BEECHER, H. K. (1955). The powerful placebo. *Journal of the American Medical Association* **159** 1602–1606.

CHALMERS, I., ENKIN, M. and KEIRSE, M. J. N. C. (1989). *Effective Care in Pregnancy and Childbirth*. Oxford Univ. Press.

CHALMERS, T. C. (1987). Meta-analysis in clinical medicine. *Transactions of the American Clinical and Climatological Association* **99** 144–150.

CHALMERS, T. C., SMITH, H. JR., BLACKBURN, B., SILVERMAN, B., SCHROEDER, B., REITMAN, D. and AMBROZ, A. (1981). A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials* **2** 31–49.

CHALMERS, T. C., LEVIN, H., SACKS, H. S., REITMAN, D., BERRIER, J. and NAGALINGAM, R. (1987a). Meta-analysis of clinical trials as a scientific discipline. I: Control of bias and comparison with large cooperative trials. *Statistics in Medicine* **6** 315–328.

CHALMERS, T. C., BERRIER, J., SACKS, H. S., LEVIN, H., REITMAN, D. and NAGALINGAM, R. (1987b). Meta-analysis of clinical trials as a scientific discipline. II. Replicate variability and comparison of studies that agree and disagree. *Statistics in Medicine* **6** 733–744.

CHALMERS, F. T., CHALMERS, T. C., WRIGHT, E. C., BERLINE, J. A. and NAGALINGAM, R. (1988). Meta-analysis of small old studies of phototherapy for neonatal hyperbilirubinemia compared with a large cooperative trial. *Controlled Clinical Trials* **9** 250.

COLDITZ, G. A., MILLER, J. N. and MOSTELLER, F. (1988). The effect of study design on gain in evaluation of new treatments in medicine and surgery. *Drug Information Journal* **22** 343–352.

COLDITZ, G. A., MILLER, J. N. and MOSTELLER, F. (1989). How study design affects outcomes in comparisons of therapy. I: Medical. *Statistics in Medicine* **8** 441–454.

DERSIMONIAN, R. and LAIRD, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7** 177–188.

EDDY, D. M. (1992). *A Manual for Assessing Health Practices and Designing Practice Policies*. American College of Physicians, Philadelphia.

EDDY, D. M. and HASSELBLAD, V. (1992). *Fast Pro: Software for Meta-Analysis by the Confidence Profile Method*. Academic, San Diego.

EDDY, D. M., HASSELBLAD, V. and SCHACHTNER, R. D. (1990). Bayesian method for synthesizing evidence: The confidence profile method. *International Journal of Technology Assessment in Health Care* **6** 31–56.

EDDY, D. M., HASSELBLAD, V. and SCHACHTNER, R. D. (1992). *Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence*. Academic, San Diego.

EMERSON, J. P., BURDICK, E., HOAGLIN, D. C., MOSTELLER, F. and CHALMERS, T. C. (1990). An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clinical Trials* **11** 339–352.

GILBERT, J. P., MCPEEK, B. and MOSTELLER, F. (1977). Progress in surgery and anesthesia: benefits and risks of innovative therapy. In *Costs, Risks, and Benefits of Surgery* (J. P. Bunker, B. A. Barnes and F. Mosteller, eds.) 124–169. Oxford Univ. Press.

GISSI (Gruppo italiano per lo studio della streptochinasi nell'infarto miocardico). (1986). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. *Lancet* **1** 397–402.

GREENLAND, S. and SALVAN, A. (1990). Bias in the one-step method for pooling study results. *Statistics in Medicine* **9** 247–252.

HIRJI, K. F., MEHTA, C. R. and PATEL, N. R. (1987). Computing distributions for the exact logistic regression. *J. Amer. Statist. Assoc.* **82** 1110–1117.

HIRJI, K. F., MEHTA, C. R. and PATEL, N. R. (1988). Exact inference for matched case-control studies. *Biometrics* **44** 803–814.

ISIS-1 (First International Study of Infarct Survival) (1986). Randomised trial of intravenous atenolol among 16,027 cases of suspected acute myocardial infarction: ISIS-1. *Lancet* **2** 57–65.

ISIS-2 (Second International Study of Infarct Survival) (1988). Randomized trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet* **2** 349–360.

KARDAUN, J. W. P. F. and KARDAUN, O. J. W. F. (1990). Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. *Methods of Information in Medicine* **29** 12–22.

LAIRD, N. M. and MOSTELLER, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care* **6–1** 5–30.

LARHOLT, K. M. (1989). Statistical methods and heterogeneity in meta-analysis. Ph.D. dissertation, Harvard School of Public Health.

LARHOLT, K. M. and GELBER, R. D. (1989). Heterogeneity in meta-analyses: A simulation study of fixed effect and random effect methods. In Larholt (1989), Ph.D. dissertation, Harvard School of Public Health.

LARHOLT, K. M., TSIATIS, A. A. and GELBER, R.D. (1989). Variability of coverage probability when applying a random effects methodology for meta-analysis. In Larholt (1989), Ph.D. dissertation, Harvard School of Public Health.

LITTENBERG, B., MOSES, L. and RABINOWITZ, D. (1990). Estimating diagnostic accuracy from multiple conflicting reports: A new meta-analytic method. Presented at American Federation for Clinical Research meetings, May. [Printed as an abstract in *Clinical Research* **38** 415A.]

LOUIS, T. A. (1991). Using empirical Bayes methods in biopharmaceutical research. *Statistics in Medicine* **10** 811–829.

MANTEL, N. and HAENSZEL, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22** 719–748.

MEHTA, C. R., PATEL, N. R. and GRAY, R. (1985). Computing an exact confidence interval for the common odds ratio in several $2 \times 2$ contingency tables. *J. Amer. Statist. Assoc.* **80** 969–973.

MIAMI (1985). Metoprolol in acute myocardial infarction (MIAMI). A randomised placebo-controlled international trial. The MIAMI Trial Research Group. *European Heart Journal* **6** 199–226.

MILLER, J. N., COLDITZ, G. A. and MOSTELLER, F. (1989). How study design affects outcomes in comparisons of therapy. II. Surgical. *Statistics in Medicine* **8** 455–466.

SACKS, H. S., BERRIER, J., REITMAN, D., ANCONA-BERK, V. A. and CHALMERS, T. C. (1987). Meta-analysis of randomized controlled trials. *New England Journal of Medicine* **316** 450–455.

SACKS, H. S., BERRIER, J., REITMAN, D., PAGANO, D. and CHALMERS, T. C. (1992). Meta-analyses of randomized control trials: an update. In *Medical Uses of Statistics*, 2nd ed. (J. C. Bailar III and F. Mosteller, eds.). NEJM Books, Boston.

SACKS, N., ROSSI, L., KUPELNICK, B. and CHALMERS, T. C. (1992). Meta-analyses of small and large trials comparing tPA with streptokinase in acute myocardial infarction. In preparation.

STAMPFER, M. J., GOLDHABER, S. Z., YUSUF, S., PETO, R. and HENNEKENS, C. H. (1982). Effect of intravenous streptokinase on acute myocardial infarction: Pooled results from randomized trials. *New England Journal of Medicine* **307** 1180–1182.

VOLLSET, S. E., HIRJI, K. F. and ELASHOFF, R. M. (1991). Fast computation of exact confidence limits for the common odds ratio in a series of $2 \times 2$ tables. *J. Amer. Statist. Assoc.* **86** 404–409.

YUSUF, S., COLLINS, R. and PETO, R. (1984). Why do we need some large, simple randomized trials? *Statistics in Medicine* **3** 409–20.

YUSUF, S., PETO, R., LEWIS, J., COLLINS, R. and SLEIGHT, P. (1985). Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Disease* **27** 335–71.