Meta-Analysis in Clinical Trials*

Rebecca DerSimonian and Nan Laird

ABSTRACT: This paper examines eight published reviews each reporting results from several related trials. Each review pools the results from the relevant trials in order to evaluate the efficacy of a certain treatment for a specified medical condition. These reviews lack consistent assessment of homogeneity of treatment effect before pooling. We discuss a random effects approach to combining evidence from a series of experiments comparing two treatments. This approach incorporates the heterogeneity of effects in the analysis of the overall treatment efficacy. The model can be extended to include relevant covariates which would reduce the heterogeneity and allow for more specific therapeutic recommendations. We suggest a simple noniterative procedure for characterizing the distribution of treatment effects in a series of studies.

KEY WORDS: *random effects model, heterogeneity of treatment effects, distribution of treatment effects, covariate information*

INTRODUCTION

Meta-analysis is defined here as the statistical analysis of a collection of analytic results for the purpose of integrating the findings. Such analyses are becoming increasingly popular in medical research where information on efficacy of a treatment is available from a number of clinical studies with similar treatment protocols. If considered separately, any one study may be either too small or too limited in scope to come to unequivocable or generalizable conclusions about the effect of treatment. Combining the findings across such studies represents an attractive alternative to strengthen the evidence about the treatment efficacy.

The main difficulty in integrating the results from various studies stems from the sometimes diverse nature of the studies, both in terms of design and methods employed. Some are carefully controlled randomized experi-

Yale University, School of Medicine, New Haven, Connecticut (R.D.); Harvard University, School of Public Health, Boston, Massachusetts (N.L.)

*This research was supported by grant CA09424-03 from the National Cancer Institute and grant GM-29745 from the National Institute of Health. We are grateful to Frederick Mosteller, Tom Louis, and Katherine Halvorsen for critical readings of various drafts, encouragement, and advice.

Address reprint request to: Rebecca DerSimonian, Yale University School of Medicine, P.O. Box 3333, New Haven, CT 06510.

Received March 25, 1986; accepted April 7, 1986.

52 Vanderbilt Ave., New York, New York 10017

0197-2456/86/$3.50

ments while others are less well controlled. Because of differing sample sizes and patient populations, each study has a different level of sampling error as well. Thus one problem in combining studies for integrative purposes is the assignment of weights that reflect the relative "value" of the information provided in a study. A more difficult issue in combining evidence is that one may be using incommensurable studies to answer the same question. Armitage [1] emphasizes the need for careful consideration of methods in drawing inferences from heterogeneous but logically related studies. In this setting, the use of a regression analysis to characterize differences in study outcomes may be more appropriate [2].

This paper discusses an approach to meta-analysis which addresses these two problems. In this approach, we assume that there is a distribution of treatment effects and utilize the observed effects from individual studies to estimate this distribution. The approach allows for treatment effects to vary across studies and provides an objective method for weighting that can be made progressively more general by incorporating study characteristics into the analysis. We illustrate the use of this model in several examples, and based on the empirical evidence, suggest a simple noniterative procedure for testing and estimation.

DATABASE

In a systematic search of the first ten issues published in 1982 of each of four weekly journals (*NEJM, JAMA, BMJ,* and *Lancet*), Halvorsen [3] found only one article (out of 589) that considered combining results using formal statistical methods. Our data consist of an ad hoc collection of such articles from the medical literature found through references provided by colleagues and through bibliographic references in articles already located [4–11]. The method we propose applies to several additional articles that have come to our attention since our original analyses [12–14].

We examine in detail eight review articles each reporting results from several related trials. Each review pools the results from the relevant trials in order to evaluate the efficacy of a certain treatment for a specified medical condition. In most of these reviews the original investigators pool the results from the relevant trials and estimate an overall treatment effect without first checking whether the treatment effect across the trials is constant. Others exclude some trials and combine the results only from trials that are similar in design and implementation. The investigators who do check for homogeneity of treatment effect before pooling use different criteria to assess this homogeneity. With two exceptions [6,11], the reviews consider randomized trials only. The two reviews that include nonrandomized studies analyze the data from the two groups of studies (randomized and nonrandomized) separately. In this study, we restrict our attention to the results of randomized trials only. We first describe the eight reviews identifying each by its first author, and in Table 1 summarize the methods used in each review:

Winship: A review of eight trials that compare the healing rates in duodenal ulcer patients treated with cimetidine or placebo therapy [4].

Table 1 The Methods and Outcome Measures Used in the Original Reviews

	Outcome measure	Overall effect estimate	Test of homogeneity
Winship	difference in proportions	pooled raw data	———
Conn	difference in proportions	pooled raw data	———
Miao	difference in proportions	weighted average	Chi-Square Test (Q_w)
DeSilva	relative risk	Mantel–Haenszel estimate for pooled relative risk	———
Stjernsward	difference in proportions	minimum and maximum	———
Baum	difference in proportions	pooled raw data	Gilbert, McPeek, and Mosteller estimate of variance [15]
Peto	difference in proportions	———	Mentions lack of heterogeneity
Chalmers	difference in proportions	unweighted average	———

Conn: A review of nine trials that compare the survival rates in alcoholic hepatitis patients with steroids or control therapy [5].

Miao: A review of six trials that compare gastric with sham freezing in the treatment of duodenal ulcer [6]. In addition, this review considers 14 observational and two controlled but nonrandomized studies.

DeSilva: A review of six trials that evaluate the effect of lignocaine on the incidence of ventricular fibrillation in acute myocardial infarction [7]. This study originally considered 15 trials but because the trials vary widely in treatment schedules and doses, some criteria for adequacy of treatment are established and only six trials that fulfill these requirements are analyzed.

Stjernsward: A review of five trials that compare the 5-year survival rates of patients with cancer of the breast treated with surgery plus radiotherapy or surgery alone [8].

Baum: A review of 26 trials that evaluate the efficacy of antibiotics in the prevention of wound infection following colon surgery [9].

Peto: A review of six trials that evaluate the efficacy of aspirin in the prevention of secondary mortality in persons recovered from myocardial infarction [10].

Chalmers: A review of a number of trials that evaluate the efficacy of anticoagulants in the treatment of acute myocardial infarction [11]. Data from 18 surveys employing historical controls (HCT), eight studies employing alternately assigned controls (ACT), and six randomized controlled trials (RCT) are given. Three endpoints, total case fatality rates, case fatality rates excluding early deaths, and thromboembolism rates are considered, although not all studies report all three endpoints. Here we consider thromboembolism and total case fatality rates in the RCTs only. The results from randomized trials are compared to those of nonrandomized ones (HCTs and ACTs) in Laird and DerSimonian [16].

METHODS

We consider the problem of combining information from a series of k comparative clinical trials, where the data from each trial consist of the number of patients in treatment and control groups, n_T and n_C, and the proportion of patients with some event in each of the groups, r_T and r_C. Letting i index the trials, we assume that the numbers of patients with the event in each of the study groups are independent binomial random variables with associated probabilities p_{Ti} and p_{Ci}, $i = 1, \ldots, k$. The basic idea of the random effects approach is to parcel out some measure of the observed treatment effect in each study, say y_i, into two additive components: the true treatment effect, θ_i, and the sampling error, e_i. The variance of e_i is the sample variance, s_i^2, and is usually calculated from the data of the ith observed sample. The true treatment effect associated with each trial will be influenced by several factors, including patient characteristics as well as design and execution of the study. To explicitly account for the variation in the true effects, the model assumes

$$\theta_i = \mu + \delta_i$$

where θ_i is the true treatment effect in the ith study, μ is the mean effect for a population of possible treatment evaluations, and δ_i is the deviation of the ith study's effect from the population mean. We regard the trials considered as a sample from this population and use the observed effects to estimate μ as well as the population variance [$\text{var}(\delta) = \Delta^2$]. Here, Δ^2 represents both the degree to which treatment effects vary across experiments as well as the degree to which individual studies give biased assessments of treatment effects.

The model just described can thus be characterized by two distinct sampling stages. First we sample a study from a population of possible studies with mean treatment effect μ and variance in treatment effects of Δ^2. Then we sample observations in the ith study with underlying treatment effect θ_i.

One issue which deserves some attention is the specification of treatment effect, θ_i. Three commonly used measures are the risk difference, $p_{Ti} - p_{Ci}$, the relative risk, p_{Ti}/p_{Ci}, and the relative odds, $[p_{Ti}/(1 - p_{Ti})/p_{Ci}/(1 - p_{Ci})]$. The relative odds is popular because of its suitability in retrospective or case control studies, and because it has some interesting mathematical properties. In particular, if we assume a constant relative odds ($\theta_i = \mu$ or $\Delta^2 = 0$), then the Mantel–Haenszel statistic is optimal for testing H_0: $\mu = 1$, and there is considerable literature on efficient estimates of μ and on methods for testing H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$. Despite these advantages, the relative odds (and the closely related relative risk) suffers in interpretability. By far the most intuitively appealing measure for trials of clinical efficacy is the risk difference, since it measures actual gains which can be expected in terms of percentages of patients treated. Besides relevance of the measure and statistical efficiency, it is also desirable to choose a measure which is nearly constant over studies, so that the effect of heterogeneity is minimized. Unless there is no treatment effect at all ($p_{Ti} = p_{Ci}$ for all i), constancy of treatment effect in one scale (say $p_{Ti}/p_{Ci} = A$ for all i) implies variation across studies in another ($p_{Ti} - p_{Ci}$, say). Thus it is conceivable that the wrong choice of scale could imply heterogeneity in treatment effects which would not exist if a different measure were used. However, this is not likely to happen in practice unless there is a very wide range in the control rates (p_{Ci}) or all the rates are very close to zero (or one). In such cases, one might want to do the analysis in both the relative odds and risk difference scales.

HOMOGENEITY OF TREATMENT EFFECT

To evaluate constancy of treatment effect across strata, we use a large sample test based on the statistic $Q = \sum_i w_i(y_i - \bar{y}_w)^2$, where y_i is the ith treatment effect estimate, $\bar{y}_w = \sum_i w_i y_i / \sum_i w_i$ is the weighted estimator of treatment effect, and w_i is the inverse of the ith sampling variance. The test statistic Q is the sum of squares of the treatment effect about the mean where the ith square is weighted by the reciprocal of the estimated variance. Under the null hypothesis, Q is approximately a χ^2 statistic with $k - 1$ degrees of freedom; thus, when each study has a large sample size relative to the number of strata, Q may be used to test H_0: $\Delta^2 = 0$.

When y_i is a difference in proportions, $r_{Ti} - r_{Ci}$, we estimate the sampling variance in the ith study, s_i^2, by

$$s_i^2 = r_{Ti}(1 - r_{Ti})/n_{Ti} + r_{Ci}(1 - r_{Ci})/n_{Ci}, \tag{1}$$

and use $Q_w = \sum_i w_i(y_i - \bar{y}_w)^2$ to test constancy of treatment effect.

The weights in Q may vary according to the assumptions made about the sampling variances. For instance, when the sampling variances can be assumed to be equal, then $w_i, i = 1, \ldots, k$, is the inverse of a common sampling variance s^2. One review [9], which includes a qualitative assessment of homogeneity of treatment effect, uses the method of Gilbert et al. [15] to estimate the magnitude of the variation across the differences in proportions. Since the method of Gilbert et al. for estimating the variation in treatment effects assumes a common sampling variance, we calculate Q_u, the analogue of Q_w, assuming equal sampling variances. Here, the treatment effect is again the difference in proportions, but

$$w_i = s^{-2}, \qquad i = 1, \ldots, k,$$

where

$$s^2 = \sum_i s_i^2/k,$$

and s_i^2 is defined in equation (1).

We also use the Q statistic for testing homogeneity in the relative odds scale. In this scale,

$$Q_L = \sum_i w_i (y_i - \bar{y}_w)^2$$

where

$$y_i = \ln [r_{Ti}(1 - r_{Ci}) / r_{Ci}(1 - r_{Ti})],$$

$$w_i = s_i^{-2}, \qquad \bar{y}_w = \sum_i w_i y_i / \sum_i w_i,$$

and

$$s_i^2 = [n_{Ti} r_{Ti}(1 - r_{Ti})]^{-1} + [n_{Ci} r_{Ci}(1 - r_{Ci})]^{-1}.$$

In the large sample case, Q_L is analogous to the goodness-of-fit test in logistic models [17]. An alternate test statistic for assessing homogeneity is the likelihood ratio test which is computationally more cumbersome than the Q statistic used here [18].

ESTIMATION AND COMPUTATION

Most of the reviews consider the differences in proportions as a measure of treatment effect (Table 1). For estimating μ and Δ^2 we also restrict our attention to this scale.

When $\Delta^2 \neq 0$, Q_w is used to derive a noniterative estimate of Δ^2 by equating the sample statistic with the corresponding expected value. This yields a weighted estimator

$$\Delta_w^2 = \max \{0, \{Q_w - (k-1)\} / [\sum_i w_i - (\sum_i w_i^2 / \sum_i w_i)]\},$$

where Q_w, $\bar{y}_w$, w_i are as described above. The weighted least squares or Cochran's [19] semiweighted estimator of μ is

$$\mu_w = \sum_i w_i^* y_i / \sum_i w_i^*, \tag{2}$$

where

$$w_i^* = (w_i^{-1} + \Delta_w^2)^{-1}. \tag{3}$$

The asymptotic standard error of μ_w is

$$\text{s.e. } (\mu_w) = (\sum_i w_i^*)^{-1/2}. \tag{4}$$

When the sampling variances are assumed to be equal, these equations reduce to:

$$\Delta_u^2 = \max [0, \{\sum_i (y_i - \bar{y})^2 / (k-1)\} - s^2],$$

$$\mu_u = \bar{y},$$

and

$$\text{s.e. } (\mu_u) = [(s^2 + \Delta_u^2) / k]^{1/2},$$

where

$$\bar{y} = \sum_i y_i / k \quad \text{and} \quad s^2 = \sum_i s_i^2 / k.$$

Rao et al. [20] derive Δ_u^2 from an unweighted sum of squares procedure and show that it is also the Minque estimator when the sampling variances are all equal. The unweighted mean, μ_u, is equivalent to the estimate of the treatment effect in reviews that use the average difference in proportions to assess the overall treatment efficacy.

With an additional assumption that y_i is $N(\theta_i, s_i^2)$ and θ_i is $N(\mu, \Delta^2)$, we also compute maximum likelihood (ML) and restricted maximum likelihood (REML) estimates and compare them to the noniterative ones. The maximum likelihood estimates of the unknown parameters are those values that maximize the probability density function of the data. In REML estimation, the likelihood to be maximized is slightly modified to adjust for μ and Δ^2 being estimated from the same data. The REML estimators are the iterative equivalents of the weighted estimators above. Both ML and REML estimates of μ and its s.e. take the form given in equations (2) and (4) with weights given in (3), but differ in the way Δ^2 is estimated.

The ML estimating equations are given in Rao et al. [20] and the REML equations are reviewed by Harville [21]. For implementing the ML or REML procedures, we use the EM algorithm [22] which is an iterative procedure for computing maximum likelihood estimates appropriate when the observations can be viewed as incomplete data.

RESULTS

Homogeneity of Treatment Effect

We present the statistics for testing homogeneity of treatment effect in Table 2. For these reviews Q_w, the weighted statistic in the difference scale, and Q_L, the analogous statistic in the log odds scale, imply similar conclusions about the constancy of treatment effect. The assumption of homogeneity holds in the reviews by DeSilva [7], Stjernsward [8], Peto [10], and in Chalmers' [11] randomized controlled trials (case fatality rates). In the remaining five sets of trials, the evidence suggests heterogeneity of treatment effect irrespective of the scale of measurement.

The review by Peto [10] mentions lack of heterogeneity in treatment effects across trials. For this review, the Q_u statistic supports the homogeneity assumption (p value = 0.65), whereas both Q_w and Q_L support that assumption only marginally (p value = 0.12). Baum et al. [9] estimate the variability in treatment differences using the method of Gilbert et al. [15] and conclude that relative to within study variation (assumed equal for all studies), between study variation is negligible. This qualitative assessment is not consistent with the results of Table 2 where a common treatment effect across the trials in this review does not seem to hold in either scale of measurement. The third review which includes a test of homogeneity before pooling the results [6] uses a slightly modified version of Q_w to test this hypothesis and the conclusion of lack of homogeneity agrees with the result in Table 2.

As in the review by Peto [10], Q_u and Q_w imply different conclusions about the homogeneity of treatment effect in the review by Winship [4]. In this review also, the method assuming equal weights (Q_u) implies homogeneity of effect while the weighted one implies the opposite.

These results emphasize that the variation in the treatment effect across several trials is often not negligible and should be incorporated into the anal-

Table 2 Test of Homogeneity[a]

	df[c]	Q_w[d]	Q_L[e]	Q_u[f]
Winship	7	15.2[b]	15.6[b]	7.9
Conn	8	15.6[b]	20.6[b]	19.3[b]
Miao	5	21.7[b]	18.8[b]	20.9[b]
DeSilva	5	9.1	4.4	7.3
Stjernsward	4	2.1	2.2	2.7
Baum	25	40.4[b]	35.2[b]	35.3[b]
Peto	5	9.0	9.2	3.5
Chalmers				
Thromboembolism	5	12.3[b]	10.3[b]	10.6[b]
Case fatality rates	5	3.5	2.4	1.8

[a]Figures in Tables 2–4 are based on data available at the time of review publication.

[b]p value <0.10.

[c]Degrees of freedom.

[d]Q statistic in difference scale (unequal weights).

[e]Q statistic in log odds scale (unequal weights).

[f]Q statistic in difference scale (equal weights).

ysis of the overall treatment efficacy. Lack of homogeneity holds both when the treatment effect is the difference in proportions and when it is the log odds. The unweighted statistic which assigns an equal weight to each study may not be appropriate for testing homogeneity when differences in sample sizes and/or underlying proportions across studies are large.

Estimation

For all four methods of estimation we present estimates of μ and its s.e. in Table 3, and estimates of Δ^2 in Table 4. The estimates of Δ^2, and s.e. ($\hat{\mu}$) are quite similar in the weighted noniterative method, maximum likelihood, and restricted maximum likelihood procedures. The Δ^2s from these three methods are zero or nearly so in the reviews by DeSilva [7], Stjernsward [8], Peto [10], and Chalmers' [11] randomized trials (case fatality rates). These same reviews have Q statistics that are small relative to their degrees of freedom (Table 2). The weighted method and the REML estimation procedures consistently yield slightly higher values of Δ^2 than the ML procedure. This is because both these procedures adjust for μ and Δ^2 being estimated from the same data whereas the ML procedure does not. The estimates of μ and its s.e. from these three procedures are expected to be similar since the estimates of Δ^2 are almost equal.

Comparing the unweighted method of moments with the other three methods, we find that the estimates for Δ^2 from this method differ, and sometimes differ widely, from the estimates of the other three methods but without any consistent pattern. The estimates of μ and its s.e. from the unweighted method also differ from the estimates of the other three methods and these differences are not necessarily due to the differences in Δ^2s. In Chalmers' [11] randomized trials (case fatality rates), for instance, even when Δ^2 is zero for all four methods, the estimate of μ is 0.042 (s.e. = 0.024) for the unweighted method while it is 0.029 (s.e. = 0.012) for the other three methods. The original reviewers report the unweighted average of the observed rate differences

Table 3 Estimated Overall Effects and Their Standard Errors[a]

	μ_u[b]	μ_w[c]	μ_M[d]	μ_R[e]
Winship	0.406 (0.046)	0.389 (0.058)	0.384 (0.053)	0.387 (0.056)
Conn	0.102 (0.092)	0.075 (0.072)	0.070 (0.063)	0.073 (0.069)
Miao	0.077 (0.125)	0.094 (0.111)	0.095 (0.106)	0.093 (0.118)
DeSilva	0.026 (0.019)	0.027 (0.019)	0.026 (0.017)	0.027 (0.019)
Stjernsward	0.046 (0.020)	0.041 (0.018)	0.041 (0.018)	0.041 (0.018)
Baum	0.203 (0.031)	0.208 (0.026)	0.208 (0.025)	0.208 (0.026)
Peto	0.018 (0.008)	0.015 (0.008)	0.014 (0.008)	0.015 (0.008)
Chalmers				
Thromboembolism	0.102 (0.036)	0.079 (0.020)	0.078 (0.017)	0.078 (0.020)
Case fatality rates	0.042 (0.024)	0.029 (0.012)	0.029 (0.012)	0.029 (0.012)

[a]Figures in parentheses represent the standard errors of the corresponding estimates.

[b]Noniterative estimates with equal weights.

[c]Noniterative estimates with weights to reflect unequal variances.

[d]Maximum likelihood estimates.

[e]Restricted maximum likelihood estimates.

Table 4 Estimated Variation in the True Effects

	Δ_u^{2a}	Δ_w^{2b}	Δ_M^{2c}	Δ_R^{2d}
Winship	0.0020	0.0137	0.0096	0.0117
Conn	0.0442	0.0208	0.0112	0.0176
Miao	0.0716	0.0540	0.0482	0.0638
DeSilva	0.0007	0.0009	0.0006	0.0009
Stjernsward	0	0	0	0
Baum	0.0072	0.0062	0.0049	0.0057
Peto	0	0.0002	0.0002	0.0002
Chalmers				
Thromboembolism	0.0041	0.0012	0.0007	0.0012
Case fatality rates	0	0	0	0

[a]Noniterative estimates with equal weights.

[b]Noniterative estimates with weights to reflect unequal variances.

[c]Maximum likelihood estimates.

[d]Restricted maximum likelihood estimates.

(0.042) as an estimate of overall treatment efficacy. The weighted estimate of the treatment effect which weights the observed effects in relation to sample size is lower than the unweighted average, since some of the larger studies have smaller estimated treatment effects.

DISCUSSION

We have used a simple random effects model for combining evidence, and applied it to characterize the distribution of treatment effects in a series of studies. The model is useful both in summarizing the data and in illustrating the different kinds of results which one obtains from randomized and nonrandomized studies. In general, studies with greater potential for bias, such as uncontrolled or nonrandomized ones, show greater treatment effect as well as greater heterogeneity [2,16].

One important finding that emerges from this investigation is that heterogeneity of treatment effects across studies is common and should be incorporated into the analysis. The random effects model incorporates this heterogeneity, however small, in the analysis of the overall efficacy of the treatment. The method estimates the magnitude of the heterogeneity, and assigns a greater variability to the estimate of overall treatment effect to account for this heterogeneity. In principle, we can extend the model to include pertinent covariate information [2]. Utilizing covariate information may substantially reduce the heterogeneity of effects and thus allow for more specific therapeutic recommendations. This is often difficult in practice, however, since covariate information may be missing for some studies. Improvement in publication standards for medical reporting and further methodological work for handling missing covariate information are needed to strengthen our ability to combine results from clinical studies.

For estimating the overall treatment effect and the variation of effects across studies, our results suggest that the weighted noniterative method is an attractive procedure because of the comparability of its estimates with those of the maximum likelihood methods and because of its relative simplicity. On

the other hand, the unweighted method which ignores differences in sample sizes yields estimates that often differ from the estimates of the other methods.

A problem in pooling data we have not addressed here is that of publication bias. This problem relates to studies being executed, but not reported, usually because treatment effect has not been found. Reviewers generally recount those studies that appear to be worthwhile and discount those that are unpublished or are not in agreement with a favored group of studies. The method we describe here represents a systematic, quantitative pooling of available data to resolve controversies about a treatment effect. With each individual controversy, unpublished information may be elicited and along with recent findings the method can be used to update the results.

In all our work we assume that the sampling variances are known, although in reality we estimate them from the data. Further research needs to be done in this area as there are alternative estimators that might be preferable to the ones we use. For instance, if the sample sizes in each study are small, then sampling variances based on pooled estimates of the proportions in the treatment and control groups might be better than the ones based on estimates of proportions from the individual studies. Another alternative is to shrink the individual proportions towards a pooled estimate before calculating the variances. Further investigation is needed before one single method emerges as superior.

REFERENCES

1. Armitage P: Controversies and achievements in clinical trials. Controlled Clin Trials 5: 67–72, 1984
2. DerSimonian R, Laird N: Evaluating the effect of coaching on SAT Scores: a meta-analysis. Harvard Ed Rev 53: 1–15, 1983
3. Halvorsen K: Combining results from independent investigations: meta-analysis in medical research. In: Medical Uses of Statistics, Bailar JC, Mosteller F, Eds. Boston: New England Journal of Medicine (in press)
4. Winship D: Cimetidine in the treatment of duodenal ulcer. Gastroenterology 74: 402–406, 1978
5. Conn H: Steroid treatment of alcoholic hepatitis. Gastroenterology 74: 319–326, 1978
6. Miao L: Gastric freezing: an example of the evaluation of medical therapy by randomized clinical trials. In: Costs, Risks, and Benefits of Surgery, Bunker JP, Barnes BA, Mosteller F, Eds. New York: Oxford University Press, 1977, pp. 198–211
7. DeSilva RA, Hennekens CH, Lown B, Casscells W: Lignocaine prophylaxis in acute myocardial infarction: An evaluation of randomized trials. Lancet ii: 855–858, 1981
8. Stjernsward J: Decreased survival related to irradiation post-operatively in early operable breast cancer. Lancet ii: 1285–1286, 1974
9. Baum ML, Anish DS, Chalmers TC, Sacks HS, Smith H, Fagerstrom, RM: A survey of clinical trials of antibiotic prophylaxis in colon surgery: evidence against further use of no-treatment controls. N Engl J Med 305:795–799, 1981
10. Peto R: Aspirin after myocardial infarction. Lancet i: 1172–1173, 1980 (unsigned editorial)
11. Chalmers TC, Matta RJ, Smith H, Kunzler AM: Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. N Engl J Med 297: 1091–1096, 1977

12. Stampfer MJ, Goldhaber SZ, Yusuf S, Peto R, Hennekens CH: Effect of intravenous streptokinase on acute myocardial infarction: Pooled results from randomized trials. N Engl J Med 307: 1180–1182, 1982
13. Long-term and short-term beta-blockade after myocardial infarction. Lancet i: 1159–1161, 1982
14. Wortman PM, Yeaton WH: Synthesis of results in controlled trials of coronary artery bypass graft surgery. Evaluation Studies Review Annual 1983
15. Gilbert JP, McPeek B, Mosteller F: Progress in surgery and anesthesia: benefits and risks of innovative therapy. In: Costs, Risks, and Benefits of Surgery, Bunker JP, Barnes BA, Mosteller F, Eds. New York: Oxford University Press, 1977, pp. 124–169
16. Laird N, DerSimonian, R: Issues in combining evidence from several comparative trials of clinical therapy. In: Proceeding of the XIth International Biometric Conference. 1982, pp. 91–97
17. Breslow NE, Day NE: Statistical methods in cancer research. International Agency for Research on Cancer, 1980, pp. 136–146
18. Hedges LV, Olkin I: Statistical methods for meta-analysis. London: Academic Press, 1985, pp. 122–127
19. Cochran WG: Adjustments in analysis. In: Planning and Analysis of Observational Studies, Moses LE, Mosteller F, Eds. New York: Wiley, 1983, pp. 102–108
20. Rao PS, Kaplan J, Cochran WG: Estimators for the one-way random effects model with unequal error variances. J Am Stat Assoc 76: 89–97, 1981
21. Harville DA: Maximum likelihood approaches to variance component estimation and to related problems. J Am Stat Assoc 72: 320–338, 1977
22. Dempster AP, Laird NM, Rubin DB: Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39: 1–38, 1977