Draft of 9.83
Printed 19.11.86

EVALUATING A SERIES OF CLINICAL TRIALS OF THE SAME TREATMENT

Douglas G. Altman

Division of Medical Statistics

MRC Clinical Research Centre

Watford Road

Harrow, Middlesex HA1 3UJ

(01)-864 5311 ext 2705

EVALUATING A SERIES OF CLINICAL TRIALS OF THE SAME TREATMENT

SUMMARY

Often several independent clinical trials are carried out to evaluate the same therapy or closely similar ones. In many cases there is considerable variation in the findings of the various trials, resulting in uncertainty about whether the therapy is beneficial or not.

This paper first describes the expected degree of random variation between the results of a series of identical trials, and outlines several other possible reasons for between-study variability in results. Arguments for and against statistically combining the results from several studies are presented, and it is concluded that such an approach is desirable under certain conditions.

Various methods of combining trial results are discussed, and a worked example illustrating several methods is given as an appendix. The Mantel-Haenszel and Woolf (logit) methods are especially suitable. Other aspects discussed include between-study homogeneity of results, ethical aspects, and publication bias.

KEYWORDS

Clinical trials, Combining results, Publication bias, Between-study variability, Mantel-Haenszel test, Logistic regression, Meta-analysis.

1.0 INTRODUCTION

When several clinical trials of the same treatment are carried out the results often vary considerably. Unless the true treatment effect is very large, which is rare, there may well be uncertainty as to whether the treatment is beneficial or not. Apart from the expected random variation there are many factors that may contribute to the heterogeneity of study findings, some of which directly to the validity of the results of individual trials. Even with several studies which are all reasonably reliable, there is the further problem of how to combine the results statistically to get the overall picture. An important problem is the possibility that studies which are published are a biased subset of the studies actually carried out. Each of these aspects will be considered in this paper.

In many trials the outcome of interests is dichotomous, such as death/survival or improved/not improved, and we are interested in the proportion of individuals with a particular outcome, which is equivalent to the probability of that outcome for an individual. This paper concentrates on such trials because they are more suitable for considering between-study variability. The emphasis will be on trials where the outcome of interest is whether or not a patient survives a given length of time. The same principles also apply to studies with continuous outcome measures, but it is harder to generalise for such trials because the results depend on the between-individual variability, so the continuous case will be referred to only briefly.

2.0 BETWEEN-STUDY VARIABILITY

2.1 RANDOM VARIATION

No medical treatment, whether effective or not, will produce identical responses from all patients. For example, hypotensive drugs will not produce exactly the same reduction in blood pressure for all patients, and may not show any benefit at all for some people. Likewise the death rate observed in a clinical trial may not accurately reflect the true risk of dying because of random variation.

If the 'true' proportions surviving in the control and treatment groups are p_C and p respectively, and each trial involves n subjects in each group, then it is possible to calculate the range within which the results of a series of identical trials might vary. There are several ways of measuring the effect of treatment, for example as p_C-p_T, as $100(p_C-p_T)/p_C$, or as $p_C(1-p_T)/[p_T(1-p_C)]$. The last of these is the odds ratio, which is approximately the relative risk of dying in the two treatment groups. Table I shows some examples of the expected 90% range for the odds ratio for plausible values of p_C, p_T and n. Clearly we should expect the observed results from a series of trials to show considerable variability around the true difference unless they were all extremely large.

Closely-related calculations are used when designing studies to obtain appropriate sample sizes to give a high probability (power) of obtaining a significant result when the true difference

is of a given magnitude. Although such calculations ought to be made at the planning stage of any trial, this is very rare to judge from reviews of published trials [1-4]. Thus many trials are too small to be able to detect quite large treatment effects. Whilst not wishing to underplay the serious weakness of such trials [5], several non-significant results which point in the same direction will usually be seen as more convincing than a less consistent set of results even if the overall significance level (to be discussed later) is the same. So it is worth considering how likely it is to obtain a result that is 'opposite' to the truth; that is, how often will the new treatment appear worse when it is in fact better? (This concept is less strong than the 'error of the third kind' (gamma) described by Schwartz *et al* [6], which is the probability of finding *significance* in the wrong direction.) These probabilities can be derived in a similar manner to sample size calculations. Table II shows such probabilities for the same values of p_C, p_T and n as were used in Table I. For the size of study often reported, the risk of observing $p_T > p_C$, when p_T is truly less than p_C, is quite considerable. Mosteller *et al* [2], in a survey of 285 clinical trials of treatment for breast cancer, found a median sample size of less than 50 per treatment. Studies as small as this have a large probability of failing to detect clinically important effects, or even of observing the treatment difference in the 'right' direction, as has been demonstrated.

Table III shows the chances of observing a statistically significant effect, again for the same combinations of p_C, p_T and n.

These figures show that if there is a real but modest benefit of treatment a series of small studies would be expected to yield a fairly wide range of results, and relatively few of them would show differences that were statistically significant. This, in itself, should not be taken as an indication of incompatibility of results, nor that there may not be an important real benefit of treatment. For cases where the effect of treatment is small, say with a reduction in mortality of less than 20%, even very large studies will often fail to detect the true effect (that is, the benefit of treatment will not be statistically significant). A good example of this phenomenon is given by recent trials of beta-blockade after myocardial infarction [7].

2.2 OTHER SOURCES OF BETWEEN-STUDY VARIATION IN RESULTS

The preceding remarks all relate to the unrealistic situation where it is possible to carry out several identical trials. In reality there are numerous other possible sources of between-study variability as well as random variation, the most important of which are examined below. Again, the discussion largely relates to trials where the outcome is death or survival but applies generally.

(a) Entry criteria

If the treatment in question is not equally effective for all types of patient, then the different entry criteria for patients in the various studies may contribute to between-study variation in results. Common important factors are age and severity of disease (the classification of which may also vary between studies); the effects of these may even interact. A more relevant consideration perhaps is diagnostic variation, which was observed in the series of case-control studies of reserpine and breast cancer [8] and trials of anticoagulants after myocardial infarction [9].

If subjects are allocated at random to treatment and control group without stratification, as is usual, then there may also be imbalance within a study. This additional form of random variation will again be potentially most harmful in small studies [10].

(b) Study populations

Differences in the populations from which study samples are drawn may affect the results. This may be a particular problem in studies involving hospital in-patients, since, for example, different hospitals will attract patients with varying cross-sections of disease states (apart from other differences).

(c) Randomisation

Most medical researchers are probably now aware of many of the most important statistical 'errors', such as not randomising, that can be made when carrying out a clinical trial. That there are still many poor studies performed (and published) may be partly due to less awareness of the potential magnitude of the influence on the findings of such deviations from good methodology. Series of trials of the same treatment offer an opportunity to assess some of these.

Perhaps the best known series is that of over thirty trials of anticoagulant therapy for myocardial infarction patients [9,11-13]. Examination of the individual results shows that the trials employing historical controls found much larger treatment effects than the randomised trials, and the between-study variability was much greater in the trials using historical controls than in the randomised trials.

More recently Sacks *et al* [14] have compared the results of randomised clinical trials and trials using historical controls for six different therapies, including anticoagulant therapy after myocardial infarction. They found that in each case the use of historical controls tended to give an optimistic result in favour of the treatment compared with the results of the randomised trials.

Another example of the danger of using historical controls in a single study related to the use of computed tomography (C-T) scanning for stroke victims [15]. Here again changes over time

were largely to blame for a misleading association between C-T scanning and three-month survival rates.

Trials using historical controls are clearly likely to give misleading results, and ought to be excluded from any exercise combining the results of a series of trials.

(d) Blindness

The arguments in favour of blindness, especially double blindness, in controlled trials are similar in some respects to those relating to randomisation. The idea is to prevent the introduction of biases, both conscious and unconscious, caused by the patient and/or the investigator(s) knowing which treatment that patient is receiving. Unconscious biases may be related to preconceptions about the value of the treatment. Conscious bias on the part of the investigator may even be so strong as to lead to interference with the randomisation in order to improve the likelihood of obtaining the 'desired' result. Such a phenomenon is rarely reported for obvious reasons; it was described, however, in one of the studies of anticoagulant therapy already discussed [16]. The magnitude of such biases is uncertain, but variation in the degree of blindness in studies with concurrent controls would contribute to between-study variability.

In some circumstances it is not possible to perform double blind trials, for example when evaluating surgical procedures or regimens of physical exercise.

(e) Variation in protocols

The dose regimen is obviously an important factor in the observed success (or otherwise) of a treatment, but there are many ways in which this may vary between studies. In particular, different doses may be used in different studies, especially in the early stages of evaluation, which might affect the absolute success rate in the treated group, and thus the observed 'treatment effect' in comparison with the control group. Also the control groups in different studies may not receive the same standard treatment or placebo.

Another factor of great importance is length of follow-up. Differences between treatments will not be equally apparent at all stages of follow-up - clearly, in the extreme, all treatments will have a zero survival rate. Length of follow-up may vary between studies, as may the decision about including 'early' events, such as deaths within a short period of entering the study. Variability of this type is clearly seen in the well-documented series of studies of anticoagulant therapy for myocardial infarction [12]. Even in a single study this issue can cause considerable controversy, as in the Anturane reinfarction trial [17,18].

(𝄞) Deviation from protocol

Subjects drop out of clinical trials for a variety of reasons. Some of these will be unrelated to the trial, others clearly related, whilst for many the cause may be uncertain. What should be done about these, and about non-compliance - failure to

adhere to the study protocol (including accidental non-compliance)? There are some powerful reasons for including all cases in the final analysis, ignoring deviations from the protocol, and considering all dropouts as failures (i.e. subjects where the treatment was unsuccessful) irrespective of whether the cause of dropping out is obviously related to the nature of the trial or not.

Whether subjects who drop out of a trial are included or excluded from the analysis can have a marked effect on the results [19]. Variation between studies in the way such cases are dealt with may thus contribute to between-study variability in trial results. Some studies do not give a clear indication of their policy in this respect. There is a notable tendency for current opinion to favour inclusion of all cases in the analysis [20].

(g) Baseline differences between treatment groups.

Despite randomisation it is quite possible that in a particular trial the two treatment groups are unbalanced with respect to baseline characteristics. Such imbalances can be allowed for in the analyses of individual trials, but cannot be incorporated into a pooled analysis. The chance of such imbalance is greater for small trials and for thos using simple randomisation.

2.3 SUMMARY: BETWEEN-STUDY VARIABILITY

Because the effects of different independent sources of variability are additive, the cumulative effect of all the above factors will exceed the variability due to random variation alone. It is likely, though, that most of these sources of variation will have small effects in comparison with the large expected amount of random variation. Such considerations are important because they have a direct bearing on the validity of the statistical combination procedure discussed below.

We should not be surprised when clinical trials give differing results [21], except perhaps when the sample sizes are all large, and the statistical methodology is not only sound, but also consistent between all studies. Indeed, we ought to be suspicious of a series of small studies producing closely similar results. This argument applies even more strongly to the analysis of subgroups. For example, inconsistency in the observed effect of beta-blockers in different age groups [22] is only to be expected. Similarly any attempt to distinguish between different beta-blockers will inevitably be fruitless [7].

3.0 COMBINING RESULTS FROM SEVERAL TRIALS

If there are several trials which are felt to be reliable, then it is clearly desirable to combine their results statistically to obtain an overall estimate of the efficacy of the treatment. This will be particularly useful where the trials do not appear to give compatible quantitative results. We would not

necessarily expect compatibility of significance levels because these are related to sample size. Yet far too many studies yielding non-significant differences are taken as indicative of no treatment effect, although such results may only appear incompatible because of their low power arising from inadequate sample size. (The same problem might in some cases arise in the analysis of a single multi-centre trial, for example if the entry criteria varied, or in an international study in which there are many possible differences between centres.)

Goldman and Feinstein [9] have suggested criteria for the validity of an analysis which pools the results from several clinical trials. They felt that such an analysis 'can be valid only if the component studies contain patients who are similar in diagnosis, clinical severity, principal treatment, and outcome events.' They also pointed out that some older trial results may no longer be applicable due to changes in medical practice. Their discussion is helpful, but whether or not the criterion of similarity is met may well be a matter of opinion.

Perhaps unfortunately there are several different approaches to combining individual trial results, and these may lead to different overall findings. Broadly speaking there are three main possibilities, involving the combination of (a) the data; (b) the test statistics; or (c) the probabilities.

3.1 COMBINING THE DATA

It seems obvious that the best approach would be to combine in some way the raw data from the individual studies. If the outcome measure is dichotomous (e.g. improvement on treatment, survival a given length of time), then this ought to be straightforward. The published reports of such studies ought always to give the numbers in each group and the numbers or proportions with a specific outcome, so that getting at the raw data poses no problems. The statistical problem is one of estimating the relative risk by combining the results from several contingency tables. It is first worth noting that simply adding the corresponding cells in the different tables yields a very unreliable estimate of the relative risk [23].

Since the main purpose of this paper is to discuss the whole issue of whether or not to combine trials rather than to carry out a comparison of the available methods, not all of the possible statistical methods will be described. Fleiss [24] reviews the literature in some detail.

The best-known method is probably that due to Mantel and Haenszel [25]. This method involves computing an 'observed minus expected' type of statistic for each table and combining them. It leads to an overall estimate of relative risk as well as giving an overall probability for the difference between treated and control patients, and has the advantage of being very easy to compute.

An alternative approach, suggested by Woolf [26], is to calculate the approximate relative risk (or odds ratio) for each study, and to combine the logarithms of these by weighting each estimate by its variance. Since, for a single 2 x 2 table, the logarithm of the relative risk is equal to the difference between the logits of the proportions with an unfavourable outcome in the two groups (see Appendix), the overall approximate relative risk is easily estimated using logistic regression. This method is particularly suitable for certain extensions to be suggested later. The two methods yield very similar results in published examples [24,27,28], and in several other cases I have studied. Breslow and Day [28] and Fleiss [24], who discuss the methods in detail, observe that the Woolf method is not as good as the Mantel-Haenszel method for combining the results from small studies. A worked example using these two methods is given in the Appendix.

Another method, suggested by Cochran [29], uses the weighted sum of the differences between the proportions in each group. The use of this method involves the assumption of a constant difference in the survival rates in the two groups. This might not be a valid assumption when combining the results from trials with differing lengths of follow-up. For example, several beta-blocker trial reports have included plots of survival by length of follow-up which suggested that the relative risk was more constant than the absolute difference.

All of these methods involve some approximations, but a small degree of statistical innaccuracy is surely acceptable in this context as combining results from different studies inevitably involves some assumptions that will not be met exactly.

For studies where the outcome measure is continuous (such as the change in some physical measurement) and the means and standard deviations are available for the treated and control groups, the data could be combined in a fairly straightforward manner by an unbalanced two-way analysis of variance.

3.2 COMBINING TEST STATISTICS

If the raw data cannot be combined, perhaps because they are unavailable or the experimental designs were incompatible, then it may be possible to combine test statistics.

For continuous variables, m individual t values derived from different studies (of reasonable size) can be combined by calculating

$$z_1 = \sum_{i=1}^{m} t_i \Big/ \sum_{i=1}^{m} [f_i/(f_i-2)]$$

where the ith t value is on f_i degrees of freedom, and z_1 is a standardised Normal deviate [30].

For categorical data χ^2 values resulting from the comparison of proportions can be combined into a standardised Normal deviate by calculating

$$z_2 = \sum_{i=1}^{m} s_i \sqrt{\chi_i^2} \Big/ \sqrt{m}$$

where s_i = 1 or -1 according to the direction of the difference, and z_2 is a standardised Normal deviate [31].

Whenever possible the more powerful Mantel-Haenszel and Woolf methods are preferable as they lead to a pooled estimate of the effect of interest rather than just a significance test.

For studies comparing survival times Peto *et al* [20] have pointed out that different trials could be combined using the logrank test if the relevant observed and expected logrank statistics were available for each study.

3.3 COMBINING PROBABILITIES

There are several methods of combining probabilities [32], the most familiar being due to Fisher [33]. The probabilities (P_i) associated with m tests of significance are combined by calculating

$$q = -2 \sum_{i=1}^{m} \ln P_i$$

where q is a χ^2 variate on 2m degrees of freedom. For the purposes of this calculation one-tailed probabilities must be used, a point which is not always made clear. The resulting pooled P value can be doubled to give an overall two-sided test. Gill [34] has pointed out that if the average of the P_i is greater than about 0.3 then combined significance can not be achieved - 'strong evidence can not be obtained by combining bits

of rather weak evidence.'

This method does not appear to incorporate any weighting to take account of reliability, so that a result with P=0.10 from a sample of 20 would have the same influence as one based on 2000. However, this is probably reasonable since sample size is implicit in the calculation of P, and the evidence from the larger study would in a sense not be as strong.

As already indicated, the theory underlying Fisher's formula requires one-tailed P values, but some tests, notably χ^2 tests, are inherently two-sided, so that P_i needs to be replaced by $s_iP_i/2$ in Fisher's formula. (This is exactly equivalent to taking the one-tailed probability associated with the Normal deviate obtained from the alternative formulation for testing the difference between two proportions.)

Chalmers *et al* [12] used the Mantel-Haenszel and Fisher methods to combine the results from the six randomised trials of anticoagulant therapy also analysed in the Appendix, and obtained P=0.015 and P=0.03 respectively, although it is not clear how they applied Fisher's formula. Fisher's method is generally felt to be less sensitive than using the raw data [20]. The method described in this section is also illustrated in the Appendix.

3.4 AN EXTENSION

As described earlier, for categorical data the results can be combined using linear model analysis (logistic regression), which allows other factors to be incorporated in the analysis.

Firstly, it is possible to weight each study not only to take account of its size, but also to consider its statistical quality. A scoring system was recently proposed [35] which could be the basis for such a method. Horwitz and Feinstein [21] argued against this approach for case-control studies since a score for methodological standards need not be related to the magnitude of bias introduced by the use of inferior methods, but some weighting may be preferable to none. We should certainly be suspicious if there was a noticeable correlation between the results of different studies and their weighting scores.

Secondly, and more importantly, account can sometimes be taken of various differences between the studies in the types of patient, treatment, follow-up period, and so on. I have used this approach to analyse the results from 30 controlled trials of imipramine for depression [36]. For each trial information was available on whether the depression was chronic or acute, endogenous or neurotic, and whether patients were in-patients or out-patients. These factors, together with year of publication (in 3 broad groups), were added to the regression model. The drug was clearly considerably superior to placebo ($t > 10$) although only 10 of the 30 studies had given $P<0.05$. This result was unaffected by any of the other factors, except that there was a significant association between drug-placebo difference and type

of depression - neurotic or endogenous. This effect was noted by Rogers and Clay [36], but is much more clearly seen (and estimated) from an analysis of this sort, which may have the added advantage of suggesting areas for further research which would probably not be apparent from the findings of a single study.

The same approach might be used to correct for imbalance in important baseline characteristics, which can have an important effect on the results of a trial whether or not the imbalance is statistically significant [10], and to investigate possible time trends.

For the case of continuous outcome measures extra variables can be incorporated in a similar way as covariates in the analysis of variance.

3.5 HOMOGENEITY OF STUDY RESULTS

The combination of the results of several studies to give an overall impression of the effectiveness of a treatment should incorporate an assessment of the homogeneity of the various results. Because the expected amount of random variability between studies can be calculated, such an assessment poses no difficulties. An appropriate significance test can be incorporated into both the Mantel-Haenszel and logistic regression analyses (see Appendix). In both cases the null hypothesis is that the true odds ratio is constant for all the trials.

Note that the homogeneity of odds ratios neither implies nor requires homogeneity of the contributory proportions. This is particularly relevant when considering studies with varying lengths of follow-up, if it can be assumed that the odds ratio stays reasonably constant with time. (For some studies the information might be provided, allowing assessment of how reasonable this assumption is.)

The end product of the combination of the results of several trials would carry greater weight if the results were shown to be homogeneous, and the procedure is arguably not valid otherwise. The analysis of homogeneity may itself not be very powerful, however, if it involves the comparison of several small studies with low power, nor if there is only a small number of trials.

In most cases there is no statistical evidence of heterogeneity, such as with the data analysed in the appendix. An interesting example of (possible) heterogeneity is given by the seven trials of platelet-active drugs discussed by May et al [46]. Six trials showed modest benefits, with odds ratios ranging from 1.22 to 1.79 but none was statistically significant. The seventh trial, which was the largest, found an adverse effect with an odds ratio of 1.13. The homegeneity test gives χ^2 = 11.2 and 6df (P=0.08), and pooling also gives an equivocal result with 95% confidence limits for the odds ratio of 1.00 to 1.26 (P=0.05).

4.0 PUBLICATION BIAS

The unusual is more likely to be published than the routine. One little-discussed aspect of this relates to disputes in statistical methodology, and is evidenced by the disproportionately high amount of space given to the minority views on the acceptability of historical controls in clinical trials. A similar phenomenon has been seen in the publicity given to anyone disagreeing with the widespread belief that smoking causes lung cancer.

Of more relevance in this paper, however, is consideration of the possible biases relating to which studies are published. It is extremely difficult to get hard evidence of this, but there are two likely sources of bias:

(a) Researchers are more likely to submit their results for publication if they have achieved a positive (i.e. significant) result (or possibly an unusual result).

(b) Journals are more likely to publish papers that demonstrate a positive (or unusual) result.

Clearly these two possibilities are related. For example, one rejection may be more likely to deter the authors from resubmitting the paper elsewhere if their results were 'negative'.

These suggestions are a mixture of speculation and anecdote, although many believe that such biases exist [13,37-39]. What evidence is there to support the idea of publication bias? If we consider trials with a dichotomous outcome measure, then we would expect the proportions of successes in the treated and control groups observed in the various subjects to vary around their true population values. Clearly the magnitude of these deviations will be potentially greater for small studies as the variance of the observed proportions is greater, but we would not expect small studies to achieve different results on average.

Because of their greater power, larger studies ought to produce more significant results, although the magnitude of the difference between treatments should not be related to sample size. Two series of published trials show a relationship between study size and treatment effect, suggesting a publication bias of the kind postulated. Peto [13], discussing the results of various studies of rapid 5-fluorouracil injection for advanced colorectal cancer, observed that the treatment effect was half as large again in the smaller studies as in the bigger studies. Also, the 30 trials of imipramine [36] show a clear relationship between statistical significance and study size (in the opposite direction to that which might be expected), suggesting that perhaps small studies tend to be published only if significant, whereas larger studies are respectable enough to be published whether the results are significant or not. Such a situation, if it exists, is bound to lead to bias in the results of published papers, in favour of the treatment. If this sort of bias were stronger for a new treatment we would expect to see the effectiveness of treatment

appear to diminish with time.

Chalmers et al [40] have written about the "understandable tendency of clinicians to report unusual rather than expected phenomena". They suggest that an unusual result in a small sample, possibly just due to biological or sampling variability, would be more likely to appear in print than more ordinary results. Since unusual results may be in either direction they postulated that the observed between-study distribution of results would be flattened and spread out compared to what ought to be seen in an unbiased selection of studies. Publication bias has also been discussed in relation to the studies of anticoagulant therapy for myocardial infarction [12]. Maxwell [38], discussing the imipramine series, has suggested that non-significant results should be published by title only so that others are aware of such studies, but this is surely totally impracticable.

It is sometimes suggested that 5% of reported clinical trials will show false positive findings, but this would be true only if there were no publication bias, and if no treatment were effective. Since the latter condition is certainly not true, and the former is probably not, such a figure is clearly wrong - in the absence of the very knowledge that the trials are attempting to obtain, it is hard to see how a proper estimate of this sort can be obtained, although it is reasonable to assume that many false positive findings are published, especially among smaller trials.

The medical literature may not be as bad in this respect as the psychology literature. Even in 1959 a survey of 362 papers in four psychology journals found that of 294 reports using tests of significance only eight did not rejct their main null hypothesis (or the majority of null hypotheses tested). [ ]. More recently the editor of one journal has attempted to justify his policy of only accepted papers in which the test of significance of the major hypothesis yielded P<0.01[ ].

Although the problem of possible publication bias may appear to be a major restriction on the validity of combining the results from several trials, it is important to realise that any such bias also applies to the interpretation of individual studies, although this is always ignored and each study's results taken at face value. The non-publication of results is much more likely with small trials - the publicity given to large trials usually ensures that the results are published [41].

It is, however, easy to calculate the effect of adding in a further trial that showed no treatment effect. For example, adding to the six trials analysed in the appendix a further (fictitious) trial with 500 subjects in each group and an observed death rate of 20% in each group changes ψ_W from 1.29 to 1.21 and the significance level for the test of the null hypotheses that $\psi = 1$ from P = 0.004 to P = 0.013.

In some circumstances it may be possible to obtain information about unpublished studies, however, which greatly increases the credibility of any combined analysis. Peto [42] used this approach to combine the results of all known studies of

several treatments where there is much controversy over their efficacy, often with clear findings of either benefit or no benefit.

5.0 ETHICAL CONSIDERATIONS

In a clinical trial patients ought to be recruited in sufficient numbers to give a high probability of detecting an important therapeutic benefit if it exists. Often studies are much too small to meet this criterion, and may be thought to be unethical as a consequence of the low probability of providing useful results [5]. Similarly a study that is too large may be considered unethical, but this is hardly a widespread problem.

It can be argued that if the results from several statistically acceptable and scientifically similar studies are combined and yield a very highly significant result then it would be unethical to carry out a further study even if none of the contributing results was highly significant. Similarly a combined result that is not nearly significant could be strong grounds for abandoning the treatment. In such cases it is quite possible that the last contributing trial can be shown to have been unnecessary, which has ethical implications too. This argument depends strongly on the acceptability of the statistical combination procedure. Chalmers *et al* [12] used the above methods to consider six randomised trials of anticoagulant therapy after myocardial infarction (see Appendix too), and concluded that the positive combined result made further studies unethical. But

Goldman and Feinstein [9] challenged this view on account of the considerable differences between the studies with regard to important diagnostic and methodological features. In this case the ethical position is unclear. Indeed the ethical implications are likely to be uncertain unless the combined result was unequivocal and the studies could be shown to be very similar in design and execution. This is a stronger requirement than that of showing homogeneity of results.

Nevertheless it is important to consider whether the results of a series of studies of the same treatment should be accumulated on a regular basis in order to monitor the current state of knowledge about those treatments. Further trials might then be dependent on the combined significance of already completed trials, but using a stricter level of statistical significance (say $P<0.001$) than is usually applied in single trials. Even without such information clinical trials should perhaps not be given ethical committee approval unless the researchers had analysed the results of published trials in the way suggested in order to demonstrate that there was still uncertainty about the efficacy of the treatment, and the range of uncertainty encompassed clinically relevant benefit. Further, power calculations for a new trial could be conditional on the results of published trials.

The 95% (or 99%) confidence interval for the combined estimate of relative risk (or percentage improvement) can be calculated after each additional trial is published. Baum *et al* [43] used this method retrospectively to show that several

studies of antibiotic prophylaxis in colon surgery used no-treatment controls after such confidence intervals were well clear of the no-effect point.

6.0 DISCUSSION

Many clinical trials have very little chance of detecting the size of effect that the investigators are interested in because the sample sizes are much too small [1]. Indeed for such trials to produce a statistically significant result they must, by chance, observe a greatly overestimated treatment benefit [ ]. Where the magnitude of treatment benefit may be relatively small (say, a 25% reduction in mortality, or less) it is only to be expected that a series of several trials of the same treatment, most of which have low power, will not lead to a clear idea about whether the treatment is beneficial or not.

It is customary to interpret the results of each trial in isolation, but if there are several trials of the same treatment one would in principle get a more reliable assessment of treatment efficacy by combining in some way the results from all the trials. In such circumstances it may even seem perverse not to pool the evidence. This may be especially appropriate as it could well be that the later trials in a series were performed as a direct consequence of the failure of the earlier trials to resolve the issue. As has been shown, good statistical methods are available, but ought they to be used in practice?

Several possible problems concerning the validity of the statistical combination of results from different trials have been suggested: publication bias, differences in treatment, differences in study design, and incompatible results. Publication bias has already been discussed at some length. Direct evidence and subjective impressions both suggest that smaller trials have a greater chance of being published if the overall result is statistically significant, although the magnitude of this bias is impossible to quanitfy. Of course many small trials with non-significant results are published, but it would seem prudent to omit smaller trials from the statistical combination of different trials. If there are many trials being combined it would be sensible to investigate the relation between outcome and study size. In some areas of research it may be practical to obtain the results from unpublished studies to give information about all the trials of a particular treatment. In this case the methods described will be of much greater value, although it may still be difficult to produce results that will be convincing.

Whether or not the treatment is the precisely the same in a group of trials is an important consideration, but not really a statistical problem. The review paper by May et al [46] considers several groups of trials in the field of secondary prevention after myocardial infarction. In each case the treatment was not the same in all the trials. For example, in the seven trials of platelet-active drugs previously discussed six used aspirin in doses from 300 to 1500 mg per day and one used sulphinpyrazone. (The trial with the 'inconsistent' finding was one of the aspirin

trials.) It is unlikely ever to be possible to differentiate between the degrees of effectiveness of closely similar treatments. It thus seems reasonable to combine them to get an overall assessment of benefit which might apply to that group of similar treatments - the β-blocker trials are a good example of this [41].

Clinical trials can differ in numerous ways. In the present context what matters is whether there are differences between trials that would have had a direct, and non-negligible, effect on the observed trial results. If, as is advisable, consideration is restricted to randomised double-blind trials, the differences most likely to be important would relate to treatment (especially the exact treatment given, dose, and length of treatment), entry criteria, and length of follow-up. If trial results are to be combined using the Mantel-Haenszel or Woolf methods, the question is whether it can be assumed that between-trial differences in such factors did not have an important influence on the observed relative risks from each trial. (This is closely related to the requirements for the validity of the proportional hazards regression models for analysing survival times.) Such an assumption cannot readily be verified, but it is less demanding than the requirement that the outcomes (e.g. death rates) in the treatment and control groups were unaffected by differences in study design. An important example relates to length of follow-up, which can vary considerably between trials. The observed incidences of adverse outcome will obviously be affected greatly by variation in follow-up time, but the observed relative risks will be much more consistent, as is demonstrated by the

plots of deaths by length of follow-up in two of the beta-blocker trials [44,45]. Differences in trial design ought also to be judged against the background of the large degree of inherent random variability, as shown in Table 1.

In general the validity of statistical combination must be a matter of subjective opinion. For example the combination [12] of the six anticoagulant trials (discussed in the Appendix) was criticised [9] because of differences between the trials. In a recent valuable review, May et al [46] have considered six groups of studies relating to long-term secondary prevention after myocardial infarction. They cite differences in design as the main reason for not combining the results of different trials, but additionally point to the variation in mortality in the different control groups, which was discussed above. They further suggest that 'the trend of the results of most of the individual trials should be in the same direction before pooling is contemplated'. The results in the first section of the present paper show that this is not a necessary requirement. Indeed, it is when a series of trials do not all give results in the same direction that it is most desirable to carry out statistical pooling. The decision on whether or not to pool should not be affected by whether the results are apparently consistent, a possibility that can be tested by the appropriate test of homogeneity. This test is not especially powerful, however, so that genuine differences may well be missed, particularly when the number of trials is not great. Published examples of statistical combination usually show tests of homogeneity with large P values.

What should be done if there is significant heterogeneity, or when large differences in trial design make it undesirable to combine differences? A good approach is to produce for all the trials in question a careful description of the main design features and key results. May et al [46] used this format and also provided a graph showing for each trial 95% confidence limits for the relative benefit of treatment. Such comprehensive presentation, which is also desirable when statistical combination is used, will probably lead to some subjective combination of findings, however. It might sometimes be possible to separate the trials into categories related to the design. This has been done for the trials of beta-blockers according to whether subjects were entered into the trial as soon as possible after hospital admission, or somewhat later, with markedly different findings for the two groups [41]. In other cases differences in design may be incorporated into the statistical model. If it is possible to incorporate information about patient characteristics, differences in therapy and so on, such analyses may provide additional information and suggest hypotheses for further investigation.

A much more sensitive analysis could be performed if it were possible to get the raw data for every subject in each trial, including baseline factors of prognostic importance. Further, where the endpoint is death or some other fixed event, it would be better to use some form of survival analysis based on the exact times that the events occurred (although individual trials are not always analysed in this way). Unfortunately, as Chalmers [47] has described, it is likely to be excessively difficult to gain the necessary cooperation of other researchers in such an exercise.

In view of the non-statistical problems in the combination of results from different trials, the choice of statistical method is unlikely to matter greatly, but methods which make use of the raw data are definitely preferable to the combination of probabilities. The pooled estimate of relative risk should be presented with its confidence interval. A statistically significant overall treatment effect should be interpreted cautiously if only moderate significance (perhaps $0.05>P>0.005$) is obtained, although this must partly depend on the number of trials being combined. The implications of a highly significant result should, however, be considered very seriously.

The advantages of combining trial results probably outweigh the disadvantages if the trials included are of good quality, and this approach is becoming more common as more replicated trials are performed [7,41,43,47-50]. Indeed the use of this technique means that small trials which might otherwise be felt to have little or no scientific value may contribute, if not greatly, to treatment evaluation. In no sense, however, can this be taken as adequate justification for carrying out trials which are too small to be independently valuable [51]. One or two large studies will be much more valuable than a dozen small studies. In particular the findings from a large carefully-executed (possibly multi-centre) trial using a standardised protocol will be much more reliable than the combined result obtained from several small studies using different protocols.

7.0 APPENDIX: COMBINING THE RESULTS OF SEVERAL CLINICAL TRIALS - AN ILLUSTRATIVE EXAMPLE

Some of the methods described in the text will be illustrated in detail using the results of six randomised trials of the use of anticoagulants after myocardial infarction which were analysed by Chalmers *et al* [12]. The data for these six studies, together with various calculated quantities, are shown in tables IV and V. Other authors [24,28] have described these methods in much greater detail.

7.1 MANTEL-HAENSZEL METHOD

A summary χ^2 statistic may be calculated for the overall relationship between treatment groups and survival. If the ith 2 x 2 table is of the following structure

	Dead	Survived	Total
Control	a_i	b_i	a_i+b_i
Treated	c_i	d_i	c_i+d_i
Total	a_i+c_i	b_i+d_i	N_i

then the expected number of cases in the first cell under the hypothesis of no association is

$$E(a_i) = (a_i+b_i)(a_i+c_i)/N_i$$

and its variance is

$$V(a_i) = (a_i+b_i)(a_i+c_i)(b_i+d_i)(c_i+d_i)/[N_i^2(N_i-1)].$$

The summary χ^2 statistic is given by

$$\chi^2 = \frac{(|\Sigma a_i - \Sigma E(a_i)| - 0.5)^2}{\Sigma V(a_i)}$$

so that, in this case,

$$\chi^2 = (|303 - 271.633| - 0.5)^2/121.936$$

$$= 7.814 \ (P = 0.005).$$

The uncorrected version of χ^2, used for a later calculation, is given by

$$\chi^2 = (303 - 271.633)^2/121.936$$

$$= 8.069.$$

The pooled estimate of the odds ratio (or approximate relative risk) ψ is given by

$$\psi_{MH} = \frac{\Sigma(a_i d_i/N_i)}{\Sigma(b_i c_i/N_i)}$$

$$= 139.464/108.099$$

$$= 1.290.$$

The simplest method of calculating an approximate confidence interval for this estimate is due to Miettinen [52]. The lower and upper limits of a 95% confidence interval are given by

$$\psi_L, \psi_U = \psi_{MH}^{(1 \pm 1.96/\chi)}$$

where χ is the square root of the summary χ^2 statistic without the continuity correction. We have

$$\chi = \sqrt{8.069}$$

$$= 2.841$$

and $1.96/\chi = 0.690$

giving $\psi_L, \psi_U = 1.290^{0.310}$ and $1.290^{1.690}$

so $\psi_L = 1.082$ and $\psi_U = 1.538$.

To test for homogeneity it is necessary to calculate for each study the expected number of subjects in the first cell, E(a), under the assumption that the pooled estimate of the odds ratio was the truen value for all studies. We have $\psi_{MH} = 1.29$, so if we let A = E(a) we can solve for A in each table by means of the equation

$$\frac{A(d-a+A)}{(a+b-A)(a+c-A)} = 1.29$$

(This is just defining AD/BC as the common odds ratio, where B, C and D are the expected values of b, c and d.) This is a quadratic equation in A, the solutions of which are

$$A = \frac{S \pm \sqrt{\{S^2 - 4\psi(\psi-1)(a+b)(a+c)\}}}{2(\psi-1)}$$

where $S = \psi(2a+b+c)+d-a$. Only one of the solutions will give sensible answers. The variance of this estimate is given by

$$V(a) = \left(\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}\right)^{-1}$$

$$= \left(\frac{1}{A} + \frac{1}{a+b-A} + \frac{1}{a+c-A} + \frac{1}{d-a+A}\right)^{-1}$$

The Mantel-Haenszel homogeneity χ^2 is given by

$$\sum \left\{ \frac{(a_i - A_i)^2}{V(a_i)} \right\}$$

which for these six studies gives $\chi^2 = 2.385$ on 5 d.f. (P=0.79) (see Table IV). There is thus no statistical reason for not combining the results of these six trials. This method is

considerably more complex than the equivalent test using the logit method described in the next section.

7.2 WOOLF (LOGIT) METHOD

The approximate relative risk is estimated for each study as

$$\psi_i = \frac{a_i d_i}{b_i c_i}$$

and its variance is given by

$$V(\ln\psi_i) = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

The weighted mean of the logarithms of these relative risks, calculated using weights $w_i = 1/V(\psi_i)$, is

$$\Sigma w_i \ln\psi_i / \Sigma w_i$$
$$= 31.693/123.655$$
$$= 0.256$$

so that the pooled estimate of ψ is

$$\psi_w = e^{0.256} = 1.292$$

which is almost exactly the same as the Mantel-Haenszel estimate. Its variance is given by

$1/\Sigma w_i = 0.00809$ and the normal deviate for testing the null hypothesis is thus 2.85 (P=0.004).

A test of homogeneity is given by the χ^2 statistic on 5 degrees of freedom

$$\Sigma w_i (\ln\psi_i)^2 - (\Sigma w_i \ln\psi_i)^2 / \Sigma w_i$$
$$= 10.508 - 31.693^2/123.655$$
$$= 2.385$$

for which P=0.79. This value is identical to that obtained by the Mantel-Haenszel test of homogeneity, but is computationally much simpler.

Approximate 95% confidence limits for $\log\psi_W$ (under the assumption of a constant relative risk across all the trials) are given by

$$\ln\psi_W \pm 1.96\sqrt{(1/\Sigma w_i)}$$
$$= 0.256 \pm 1.96/\sqrt{123.655}$$
$$= 0.080 \text{ and } 0.432.$$

The 95% confidence limits for ψ_W are thus 1.083 and 1.540, which again agree almost exactly with those derived by the Mantel-Haenszel method.

The Woolf method may easily be implemented using a program such as GLIM [53] to perform logistic regression, although the connection may not be immediately obvious. The logit of a proportion p is defined as ln[p/(1-p)], so that, for example, the logit of the proportion dying in the treated group is

$$\ln \frac{a/(a+b)}{b/(a+b)}$$
$$= \ln(a/b).$$

The log odds ratio, $\ln\frac{ad}{bc}$, can be rewritten as ln(a/b) - ln(c/d) which is the difference between the logits of the proportions dying in the treated and control groups. Analysis by GLIM using this approach yields answers identical to those shown above.

More generally one can use log-linear models (also possible with GLIM) to analyse the numbers in the cells of each 2 x 2 table. This method gives the same answers as the logistic regression, but has the advantage that it can be extended to the case where there are more than two outcome categories. Adena and Wilson [54] have illustrated this use of GLIM using data from Breslow and Day [12]. With either of these methods it is simple to incorporate other variables, as suggested in the main paper.

7.3 COMBINATION OF χ^2 VALUES AND P VALUES

The techniques for combining χ^2 or P values are more widely applicable but, in this situation, less powerful than the methods just illustrated. Also they only lead to an overall test of significance, without any estimate of the pooled relative risk. These techniques are shown for purposes of comparison - they would not normally be used when the more powerful methods could be used.

The individual (uncorrected) χ^2 values and corresponding P values are shown in table VI. All the trials favoured the treatment, but only one result was statistically significant. One of the trials was very small (with only four deaths in all), and might be felt to be receiving undue weight in these analyses.

The χ^2 values can be combined by calculating

$$z = \sum_{i=1}^{6} \sqrt{\chi_i^2} / \sqrt{6}$$

$= 2.34$

and a two-tailed P value of 0.019.

The χ^2 test is a two-sided test, so that the P values need to be converted into one-sided values before Fisher's formula can be applied. This can be achieved by letting $P_i^* = P_i/2$ if $s_i=1$ and $1-P_i/2$ if $s_i=-1$. We calculate

$$\chi^2 \text{ (12 df)} = -2\Sigma \ln_e P_i^*$$

$$= 26.45 \text{ (P = 0.009)}.$$

Omitting the fifth study has little effect, the combination giving

$$\chi^2 \text{ (10 df)} = 25.01 \text{ (P = 0.005)}.$$

The column of weights in Table V shows that the fifth study had virtually no influence on the Woolf method, and its effect on the Mantel-Haenszel χ^2 was also negligible.

7.4 COMMENT

The results using the two main approaches - Mantel-Haenszel and logit - were almost identical throughout. The 95% confidence interval for the relative risk was 1.08 to 1.54, corresponding to between 8% and 54% more deaths in the control group, or, equivalently, a reduction in the treated group of between 7% and 35% in the death rate after myocardial infarction.

The results shown in this appendix are very similar to with those reported by Chalmers *et al* [12] but, for reasons which are not apparent, they do not agree.

REFERENCES

1 Freiman, J.A., Chalmers, T.C., Smith, H. and Kuebler, R. 'The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 negative trials.', New England Journal of Medicine, 299, 690-694 (1978).

2 Mosteller, F., Gilbert, J.P. and McPeek, B. 'Reporting standards and research strategies for controlled trials. Agenda for the editor', Controlled Clinical Trials, 1, 37-58 (1980).

3 Reed, J.F. and Slaichert, W. 'Statistical proof in inconclusive "negative" trials', Archives of Internal Medicine, 141, 1307-1310 (1981).

4 Hall, J.C. 'The other side of statistical significance: a review of type II errors in the Australian medical literature', Australian and New Zealand Journal of Medicine, 12, 7-9 (1982).

5 Altman, D.G. 'Statistics and ethics in medical research. III. How large a sample?' British Medical Journal, 281, 1336-1338 (1980).

6 Schwartz, D., Flamant, R. and Lellouch, J. Clinical Trials, Academic Press, London, 1980.

7 Anonymous. 'Long-term and short-term beta-blockade after myocardial infarction', Lancet, i, 1159-1161 (1982).

8 Labarthe, D.R. 'Methodologic variation in case-control studies of reserpine and breast cancer', Journal of Chronic Diseases, 32, 94-104 (1979).

9 Goldman, L. and Feinstein, A.R. 'Anticoagulants and myocardial infarction. The problems of pooling, drowning and floating', Annals of Internal Medicine, 90, 92-94 (1979).

10 Altman, D.G. Comparability of randomised groups. Statistician, 34, 125-136 (1985).

11 Gifford, R.H. and Feinstein, A.R. 'A critique of methodology in studies of anticoagulant therapy for acute myocardial infarction', New England Journal of Medicine, 280, 351-357 (1969).

12 Chalmers, T.C., Matta, R.J., Smith, H. 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).

13 Peto, R. 'Clinical trial methodology', Biomedicine (special issue), 28, 24-36 (1978).

14 Sacks, H., Chalmers, T.C. and Smith, H. 'Randomized versus historical controls for clinical trials', American Journal of Medicine, 72, 233-240 (1982).

15 Christie, D. 'Before-and-after comparisons: a cautionary tale', British Medical Journal, 2, 1629-1630 (1979).

16 Carleton, R.A., Sanders, C.A. and Burack, W.R. 'Heparin administration after acute myocardial infarctions', New England Journal of Medicine, 263, 1002-1005 (1960).

17 Anturane Reinfarction Trial Research Group, 'Sulfinpyrazone in the prevention of sudden death after myocardial infarction', New England Journal of Medicine, 302, 250-256 (1980).

18 Relman, A.S. 'Sulfinpyrazone after myocardial infarction: no decision yet', New England Journal of Medicine, 303, 1476-1477 (1980).

19 Sackett, D.L. and Gent, M. 'Controversy in counting and attributing events in clinical trials', New England Journal of Medicine, 301, 1410- 1412 (1979).

20 Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. and Smith, P.G. 'Design and analysis of randomised clinical trials requiring prolonged observation of each patient. II. Analysis and examples', British Journal of Cancer, 35, 1-39 (1977).

21 Horwitz, R.I. and Feinstein, A.R. 'Methodologic standards and contra- dictory results in case-control research', American Journal of Medicine, 66, 556-564 (1979).

22 Hampton, J.R. 'Now you see it, now you don't. Eccentricities and anomalies in the results of b-blocking trials', British Journal of Clinical Pharmacology, 14 (Suppl 1), 51S-55S (1982).

23 Gart, J.J. 'On the combination of relative risks', Biometrics, 18, 601-610 (1962).

24 Fleiss, J.L. Statistical Methods for Rates and Proportions. 2nd edition, Wiley, New York, 1981, p. 160.

25 Mantel, N. and Haenszel, W. 'Statistical aspects of the analysis of data from retrospective studies of disease', Journal of the National Cancer Institute, 22, 719-748 (1959).

26 Woolf, B. 'On estimating the relation between blood group and disease' Annals of Human Genetics, 19, 251-253 (1955).

27 Armitage, P. Statistical Methods in Medical Research, Blackwell, Oxford, 1971, p. 427.

28 Breslow, N.E. and Day, N.E. Statistical Methods in Cancer Research. Volume 1 - The Analysis of Case-control Studies, IARC, Lyon, 1980, p. 136 and p. 210.

29 Miller, R.G. 'Combining 2 x 2 contingency tables' In: Miller, R.G., Efron, B. et al (Eds.) Biostatistics Casebook, Wiley, New York, 1980. or Cochran, W.G. 'Some methods for strengthening the common χ^2 tests', Biometrics, 10, 417-451 (1954). [There is some confusion here!]

30 Winer, B.J. Statistical Principles in Experimental Design, 2nd edition, McGraw Hill, London, 1971, p. 49.

31 Armitage, P. Statistical Methods in Medical Research, Blackwell, Oxford, 1971, p. 373.

32 Rosenthal, R. 'Combining results of independent studies', Psychological Bulletin 85, 185-193 (1978).

33 Fisher, R.A. Statistical Methods for Research Workers. Revised 13th edition, Oliver and Boyd, Edinburgh, 1963, p.99.

34 Gill, J.L. Design and Analysis of Experiments in the Animal and Medical Sciences. Volume 1. Iowa State University Press, Ames, 1978, p. 75.

35 Chalmers, T.C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D. and Ambroz, A. 'A method for assessing the quality of a randomized control trial', Controlled Clinical Trials, 2, 31-49 (1981).

36 Rogers, S.C. and Clay, P.M. 'A Statistical review of controlled trials of imipramine and placebo in the treatment of depressive illness', British Journal of Psychiatry, 127, 599-603 (1975).

37 Gilbert, J.P., McPeek, B. and Mosteller, F. 'How frequently do innovations succeed in surgery and anaesthesia?' In: Tanur, J.M., Mosteller, F. et al (Eds.) Statistics: a guide to the biological and health sciences. Holder-Day, San Francisco, 1977.

38 Maxwell, C. 'Clinical trials, reviews, and the journal of negative results', British Journal of Clinical Pharmacology, 13, 15-18 (1981).

39 Meier, P. 'Current research in statistical methodology for clinical trials' Biometrics, 28 (Suppl), 141-150 (1982).

40 Chalmers, T.C., Koff, R.S. and Grady, G.F. 'A note of fatality in serum hepatitis', Gastroenterology, 49, 22-26 (1965).

41 Lewis, J.A. 'β-blockade after myocardial infarction - a statistical view', British Journal of Clinical Pharmacology, 14 (Suppl), 15S-21S (1982).

42 Peto, R. 'Theoretical and practical difficulities in trials in cancer and vascular disease'. Paper presented at an RSS/RSM meeting on clinical trials, 1982.

43 Baum, M.L., Anish, D.S., Chalmers, T.C., Sacks, H.S., Smith, H. and Fagerstrom, R.M. 'A survey of clinical trials of antibiotic prophylaxis in colon surgery: evidence against further use of no-treatment controls' New England Journal of Medicine, 305, 795-799 (1981).

44 Beta-Blocker Heart Attack Trial Study Group. 'Beta-blocker heart attack trial. Preliminary report', Journal of the American Medical Association, 246, 2073-2074 (1981).

45 Norwegian Multicenter Study Group, 'Timolol-induced reduction in mortality and reinfarction in patients surviving acute myocardial infarction', New England Journal of Medicine, 304, 801-807 (1981).

46 May, G.S., Eberlein, K.A., Furberg, C.D., Passamani, E.R. and DeMets, D.L. 'Secondary prevention after myocardial infarction: a review of long-term trials', Progress in Cardiovascular Disease, 24, 331-352 (1981).

47 Chalmers, T.C. 'Combinations of data from randomized controlled trials', Biometrics, 38 (Suppl), 150-153 (1982).

48 Elashoff, J.D. 'Combining results of clinical trials', Gastroenterology, 75, 1170-1174 (1978).

49 DeSilva, R.A., Hennekens, C.H., Lown, B. and Cascells, W. 'Lignocaine prophylaxis in acute myocardial infarction: an evaluation of randomised trials', Lancet, ii, 855-858 (1981).

50 Baber, N.S. and Lewis, J.A. 'Confidence in results of beta-blocker post- infarction trials', British Medical Journal, 284, 1749-1750 (1982).

51 Altman, D.G. 'Size of clinical trials', British Medical Journal, 286, 1842-1843 (1983).

52 Miettinen, O.S. 'Estimability and estimation in case-referrent studies', American Journal of Epidemiology, 103, 226-235 (1976).

53 Baker, R.J. and Nelder, J.A. The GLIM System Release 3 Manual. Numerical Algorithms Group, Oxford, 1978.

54 Adena, M.A. and Wilson, S.R. Generalised Linear Models in Epidemiological Research: Case-control Studies, Intstat Foundation, Sydney, 1982.

Table I. Approximate 90% ranges(*) of observed difference in death rates in control group and treatment group for different true values of p.. and p.., and the number of subjects (n) in each group.

Difference	p..	p..	n (in each group) 40	75	100	250	500	1000	2500	True %
	0.5	0.25	8 to 42	12 to 38	14 to 36	18 to 32	20 to 30	22 to 28	23 to 27	25
p..=0.5p..	0.25	0.125	-2 to 27	2 to 23	4 to 21	7 to 18	8 to 17	10 to 15	11 to 14	12.5
	0.15	0.075	-4 to 19	-1 to 16	0 to 15	3 to 12	4 to 11	5 to 10	6 to 9	7.5
	0.5	0.375	-6 to 31	-1 to 26	1 to 24	5 to 20	7 to 18	9 to 16	10 to 15	12.5
p..=0.75p..	0.25	0.19	-9 to 21	-5 to 17	-3 to 16	0 to 12	2 to 11	3 to 9	4 to 8	6
	0.15	0.11	-9 to 16	-5 to 13	-4 to 12	-1 to 9	0 to 7	1 to 6	2 to 5	4
	0.5	0.45	-13 to 23	-8 to 18	-7 to 17	-2 to 12	0 to 10	1 to 9	3 to 7	5
p..=0.9p..	0.25	0.225	-13 to 18	-9 to 14	-7 to 12	-4 to 9	-2 to 7	-1 to 6	1 to 4	2.5
	0.15	0.135	-11 to 14	-8 to 11	-7 to 10	-4 to 7	-2 to 5	-1 to 4	0 to 3	1.5

* Calculated as p..-p..etc....

In IEATAB.TXT

Table II. Probability of observing $p_T > p_C$ when $p_T = 0.5p_C$, $0.75\ p_C$, and $0.9p_C$ for different sample sizes (n).

			n (in each group)						
	p_C	p_T	40	75	100	250	500	1000	2500
	0.5	0.25	1%	—	—	—	—	—	—
$p_T=0.5p_C$	0.25	0.125	7%	2%	1%	—	—	—	—
	0.15	0.075	14%	7%	5%	—	—	—	—
	0.5	0.375	13%	6%	4%	—	—	—	—
$p_T=0.75p_C$	0.25	0.19	25%	18%	14%	4%	1%	—	—
	0.15	0.11	31%	25%	21%	11%	4%	1%	—
	0.5	0.45	33%	27%	24%	13%	6%	1%	—
$p_T=0.9p_C$	0.25	0.225	40%	36%	34%	25%	18%	10%	2%
	0.15	0.135	42%	40%	38%	32%	25%	17%	6%

Table III. Chance(*) of observing a significant difference between groups (P<0.05) when p..= 0.5p... 0.75p... and 0.9p... for different sample sizes (n).

			n (in each group)						
	p...	p..	40	75	100	250	500	1000	25000
					(54.4)	(85.7)	(94.48)		
	0.5	0.25			54%	86%	95%	100%	100%	100%	100%
p..0.5p...	0.25	0.125			19%	41%	55%	94%	100%	100%	100%
	0.15	0.075			9%	21%	30%	71%	96%	100%	100%
					(13.7)	(27.6)	(37.1)		
	0.5	0.375			14%	27%	37%	78%	98%	100%	100%
p..0.75p...	0.25	0.19			5%	10%	14%	35%	64%	91%	100%
	0.15	0.11				5%	8%	19%	38%	68%	97%
	0.5	0.45							
p..0.9p...	0.25	0.225							
	0.15	0.135							

* Values presented are 100 (1-B) where B is the Type II error, for a two-sided test of significance at the 5% level.

Table IV. Raw data and Mantel-Haenszel calculations for the combination of results of six trials of anticoagulant therapy after myocardial infarction as discussed by Chalmers *et al* [12].

Study	Control Group D	Control Group S	Treated Group D	Treated Group S										
1	18	29	13	32	92	576	377	1.528	15.837	5.193	6.261	4.098	17.14	5.10	0.145
2	15	55	12	65	147	975	660	1.477	12.857	5.535	6.633	4.490	14.26	5.49	0.100
3	129	586	115	597	1427	77013	67390	1.143	122.256	50.605	53.968	47.225	135.09	50.08	0.741
4	83	308	111	634	1136	52622	34188	1.539	66.773	36.344	46.322	30.095	76.24	37.91	1.205
5	2	24	2	25	53	50	48	1.042	1.962	0.942	0.943	0.906	2.20	0.92	0.044
6	56	443	48	452	999	25312	21264	1.190	51.948	23.317	25.337	21.285	57.86	23.03	0.150
Total	303								271.633	121.936	139.464	108.099			2.385

D = died, S = survived

Note that
and

Table V. Raw data and Woolf (logit) calculations for the combination of results of six trials of anticoagulant therapy after myocardial infarction as discussed by Chalmers *et al* [12].

Study i	Control Group D	Control Group S	Treated Group D	Treated Group S									
1	18	29	13	32	92	576	377	1.528	0.424	0.198	5.045	2.139	0.907
2	15	55	12	65	147	975	660	1.477	0.390	0.184	5.448	2.125	0.829
3	129	586	115	597	1427	77013	67390	1.143	0.133	0.020	50.431	6.707	0.892
4	83	308	111	634	1136	52622	34188	1.539	0.431	0.026	36.638	16.653	7.177
5	2	24	2	25	53	50	48	1.042	0.041	1.082	0.924	0.038	0.002
6	56	443	48	452	999	25312	21264	1.190	0.174	0.043	23.169	4.031	0.701
Total	303											123.655	31.693	10.508

D = died, S = survived

Note that

and

Table VI. x2 and P values for six clinical trials of anticoagulant therapy after myocardial infarction.

Study	x2	PI	PI*
1	0.911	0.340	0.170
2	0.835	0.361	0.180
3	0.899	0.343	0.172
4	7.252	0.007	0.0035
5	0.002	0.964	0.482
6	0.705	0.401	0.201

APPENDIX

Some of the methods described in the text will be illustrated using the results of six randomised trials of the use of anticoagulants after myocardial infarction as analysed by Chalmers et al. (1977). The data for these six studies, together with various calculated quantities, are shown in table A1.

MANTEL-HAENSZEL METHOD

A summary χ^2 statistic may be calculated for the overall relationship between treatment groups and survival. If the ith 2 x 2 table is of the following structure

	Dead	Survived	Total
Treated	a_i	b_i	a_i+b_i
Control	c_i	d_i	c_i+d_i
Total	a_i+c_i	b_i+d_i	n_i

then the expected number of cases in the first cell under the hypothesis of no association is

$$E(a_i) = (a_i+b_i)(a_i+c_i)/n_i$$

and its variance is

$$V(a_i) = (a_i+b_i)(a_i+c_i)(b_i+d_i)(c_i+d_i)/[n_i^2(n_i-1)].$$

The summary χ^2 statistic is given by

$$\chi^2 = \frac{(|\sum a_i - \sum E(a_i)| - 0.5)^2}{\sum V(a_i)}$$

so that, in this case,

$$\chi^2 = (|303 - 271.633| - 0.5)^2 / 122.738$$

$$= 7.763 \quad (P = 0.005).$$

The uncorrected version of χ^2, used for a later calculation, is given by

$$\chi^2 = (303 - 271.633)^2/122.738$$

$$= 8.016.$$

The pooled estimate of the odds ratio (or approximate relative risk) ψ is given by

$$\psi_{MH} = \frac{\sum(a_i d_i/n_i)}{\sum(b_i c_i/n_i)}$$

$$= 139.464/108.099$$

$$= 1.290.$$

The simplest method of calculating a confidence interval for this estimate is due to Miettinen (1976). The lower and upper limits of a 95% confidence interval are given by

$$\psi_L, \psi_U = \psi_{MH}^{(1 \pm 1.96/\chi)}$$

where χ is the square root of the summary χ^2 statistic without the continuity correction. We have

$$\chi = \sqrt{8.016}$$

$$= 2.831$$

and $1.96/\chi = 0.692$

giving $\psi_L, \psi_U = 1.290^{0.308}$ and $1.290^{1.692}$

so $\psi_L = 1.082$ and $\psi_U = 1.539.$

The Mantel-Haenszel test for homogeneity is complicated and will not be illustrated here. A worked example is given by Breslow and Day (1981).

WOOLF (LOGIT) METHOD

The approximate relative risk is estimated for each study as

$$\psi_i = \frac{a_i d_i}{b_i c_i}.$$

The weighted mean of the logarithms of these relative risks is

$$\Sigma w_i \log_e \psi_i / \Sigma w_i$$
$$= 31.693/123.655$$
$$= 0.256$$

so that the pooled estimate of ψ is

$$e^{0.256} = 1.292$$

which is almost exactly the same as the Mantel-Haenszel estimate. A test of homogeneity (Armitage, 1971a) is given by the χ^2 statistic on 5 degrees of freedom

$$\Sigma w_i (\log \psi_i)^2 - (\Sigma w_i \log \psi_i)^2 / \Sigma w_i$$
$$= 10.508 - 31.693^2/123.655$$
$$= 2.385$$

for which P=0.79. There is thus no *statistical* reason for not combining the results of these six trials.

Approximate 95% confidence limits for $\log\psi$ (under the assumption of a constant relative risk across all the trials) are given by

$$\log\psi \pm 1.96\sqrt{(1/\Sigma w_i)}$$
$$= 0.256 \pm 1.96/\sqrt{123.655}$$
$$= 0.080 \text{ and } 0.432.$$

The 95% confidence limits for ψ are thus 1.083 and 1.540, which again agree almost exactly with those derived by the Mantel-Haenszel method.

The Woolf method may easily be implemented using a program such as GLIM (Baker and Nelder, 1978) to perform logistic regression, although the connection may not be immediately obvious. The logit of a proportion p is given by log[p/(1-p)], so that, for example, the logit of the proportion dying in the treated group is

$$\log \frac{a/(a+b)}{b/(a+b)}$$
$$= \log(a/b).$$

The log odds ratio, $\log\frac{ad}{bc}$, can be rewritten as log(a/b) - log(c/d) which is the

difference between the logits of the proportions dying in the two treated and control groups. Analysis by GLIM using this approach yields identical answers to those shown above.

More generally one can use log-linear models (also possible with GLIM) to analyse the numbers in the cells of each 2 x 2 table. This method gives the same answers as the logistic regression, but has the advantage that it can be extended to the case where there are more than two outcome categories. Adena and Wilson (1982) have illustrated this use of GLIM using data from Breslow and Day (1980). With either of these methods it is simple to incorporate other variables, as suggested in the main paper.

COMBINATION OF χ^2 VALUES AND P VALUES

The techniques for combining χ^2 or P values are more widely applicable but, in this situation, less powerful than the methods just illustrated. Also they only lead to an overall test of significance, without any estimate of the pooled relative risk. These techniques are shown for purposes of comparison - they would not normally be used when the more powerful methods could be used.

The individual (uncorrected) χ^2 values and corresponding P values are shown in table A2. All the trials favoured the treatment, but only one result was statistically significant. One of the trials was very small (with only four deaths in all), and might be felt to be receiving undue weight in these analyses.

The χ^2 values can be combined by

$$z = \Sigma \sqrt{\chi_1^2} / \sqrt{6}$$

since all the differences were in the same direction, giving

$$z = 2.338$$

and a two-tailed P value of 0.019.

The weighted version of this combination yields $z = 2.80$ with $P = 0.01$.

The $\chi 2$ test is a two-sided test, so that the P values need to be converted into one-sided values before Fisher's formula can be applied. This can be done by letting $P_i^* = P_i/2$ if $s_i=1$ and $1-P_i/2$ if $s_i=-1$. We calculate

$$\chi^2 \text{ (12 df)} = -2\sum \log_e P_i^*$$

$$= 26.45 \text{ (P = 0.009).}$$

Omitting the fifth study has little effect, the combination giving

$$\chi^2 \text{ (10 df)} = 25.01 \text{ (P=0.005).}$$

Alternatively, instead using a weighting for s. since $\chi^2(12df)$ =

The column of weights in Table A1 shows that the fifth study had virtually no influence on the Woolf method, and its effect on the Mantel-Haenszel χ^2 was also negligible.

COMMENT

All the methods suggest that anticoagulant therapy is beneficial. The 95% confidence interval for the relative risk was 1.08 to 1.54, indicating between 8% and 54% more deaths in the control group, or, alternatively, that the treatment leads to a reduction of between 7% and 35% in the death rate after myocardial infarction.