IEA Edinburgh, August 1981.

EVALUATING A SERIES OF CLINICAL TRIALS OF THE SAME TREATMENT

Douglas G Altman

INTRODUCTION

Several trials of the same treatment will usually produce widely differing results. 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 very many factors that may contribute to the heterogeneity of study findings, some of which may directly affect the validity of the results. Even when there are several studies which are all reasonably reliable, then there is the further problem of trying to combine the results statistically to get an overall picture. Unfortunately there is, however, a strong possibility that in many cases the studies which are published give a biased representation of the studies actually carried out. Each of these aspects will be considered in turn in this paper.

BETWEEN-STUDY VARIABILITY

RANDOM VARIATION

No medical treatment will produce identical responses from all individuals. When the outcome of interests is dichotomous, such as death or survival, then we are interested in the proportion of individuals surviving, which is equivalent to the probability of survival for an individual. (This type of outcome is most suitable for considering between-study variability, as it is possible to generalise without having to specify between-individual variance.)

If we consider that the true proportions surviving in the control and treatment groups are p1 and p2 respectively, and that each trial involves n subjects in each group, then one study in ten will produce an observed difference in survival rates OUTSIDE the range

$$p1-p2 \pm 1.645\sqrt{[p1(1-p1)+p2(1-p2)]/n}$$

Table I shows some examples of this range for plausible values of p1, p2 and n. Clearly the likely variability of observed results around the true difference is very wide for small sample sizes. It is also worth considering how likely it is to obtain a result that is 'opposite' to the truth; that is, how often will the treatment appear worse when it is in fact x% better (in terms of survival)? These probabilities can be derived from the preceding formula in a similar manner to sample size calculations. Table II shows such probabilities for the same values of p1, p2 and n as were used in Table I. Clearly for the sort of size studies often reported, especially pharmaceutical trials, the risk of observing p2 >p1, when p2 is truly less than p1, is quite considerable. (This concept is slightly less strong than the "error of the third kind" (gamma) described by Schwartz et al (1980), which is the probability of p2 being found to be significantly greater than p1.)

Clearly, these results suggest that a series of small studies, with perhaps 40 per group, would be expected to yield a fairly wide range of results. This, in itself, should not be taken as an indication of incompatibility of results. Similar observations apply to studies with continuous outcome measures (e.g. blood pressure reduction), although they can not be generalised for the reasons outlined above. Here the problems of small samples still exist, but the magnitude of the problem depends on between-individual variability.

OTHER SOURCES OF BETWEEN-STUDY VARIATION IN RESULTS

1. Entry criteria
2. Study populations
3. Variations in protocols
4. Selection of control group/Randomisation
5. Degree of blindness
6. Deviations from protocol

SUMMARY: BETWEEN-STUDY VARIABILITY

In terms of the between-study variance of results, the effects of different independent sources of variability are additive. Thus the cumulative effect of all the above factors will exceed the variability due to random variation alone, perhaps considerably.

We should not, therefore, be particularly surprised when clinical trials give differing results unless the sample sizes are all large, and the statistical methodology is not only sound, but also consistent in all studies.

COMBINING RESULTS FROM SEVERAL TRIALS

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

There are several different approaches to combining individual trial results. Unfortunately these can produce widely differing overall findings. Broadly speaking there are three main possibilities, involving the combination of the data, the test statistics, or the probabilities. The first and last possibilities will be considered here.

COMBINING THE DATA

It seems obvious that the best approach would be to combine the raw data from the individual studies. If the outcome measure being investigated is dichotomous (e.g. improving on treatment, survival a given length of time), then this ought to be straightforward.

Perhaps the best-known approach is that due to Mantel and Haenszel (1959). This method involves computing an "observed - 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 (1955), is to calculate the relative risk 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 in the logits of the proportions in the two groups, the overall relative risk is easily calculated using linear modelling using GLIM (Baker and Nelder, 1978). This method is particularly suitable for certain extensions to be suggested later. The two methods give very similar results in examples given by Armitage

(1971) and Miller (1980); both of these authors discuss the methods in more detail.

COMBINING PROBABILITIES

The most familiar of several methods of combining probabilities is due to Fisher (1944). The probabilities (P_i) associated with m tests of significance, usually χ^2 tests, are combined by calculating $q = -2 \sum_{i=1}^{m} \log_e P_i$ where q is a chi-squared variate on 2m degrees of freedom.

Two problems arise with the use of this method. Firstly, it does not incorporate any weighting to take account of reliability. A result with P=0.10 would have the same weight whether based on a sample of 20 or 2000. Secondly, there can be interpretational difficulties ivolving one- or two-sided tests. This method is generally felt to be less sensitive than using the raw data (Peto et al, 1977).

AN EXTENSION

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

Firstly, it is possible to weight each study not only to take account of its size, but also to consider its statistical quality. Chalmers et al (1981) have recently proposed such a scoring system which could be the basis for such a method. Secondly, account can be taken of various differences between the studies in the types of patient, treatment, follow-up period etc.

Time trends in treatment effects are probably common as a result of temporal changes in other aspects of medical care, and have already been noted for the studies of anticoagulant therapy. These can be studied in the same way.

SUMMARY: COMBINING RESULTS

It is possible to use the results from several trials, carried out in varying circumstances, to get an overall measure of treatment efficacy, and to study other factors related to this. Nevertheless there are many unknown

differences between studies that may influence results, so that the conclusions drawn from such an analysis will usually need to be somewhat guarded. Much more important, perhaps, is the possibility of bias in what studies get published. This topic will be discussed in the final section of this paper.

PUBLICATION BIAS

The first point to note is that 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 suggestion that smoking causes lung cancer.

Two possible sources of bias are that 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), and 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 sufficient to deter the authors from resubmitting the paper elsewhere if their results were 'negative'.

These suggestions are a mixture of speculation and anecdote. What evidence is there to support the idea of publication bias?

If we consider trials with a categorical 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. Two series of published trials show a relationship between study size and treatment effect, suggesting a publication bias of the kind postulated.

Peto (1978), discussing the results of various studies of rapid 5-fluorouracil injection for advanced colorectal cancer (Moertel and Reitemeier, 1969), observed that the treatment effect was half as large again in the smaller studies than in the bigger studies. Also, the 30 trials of

imipramine reported by Rogers and Clay (1975) indicate a relationship between treatment effect and study size. These two series suggest 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. Maxwell (1981), 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 unworkable.

Few authors have discussed the possible biases in what papers get published. Chalmers et al (1965) have written about the "understandable tendency of clinicians to report unusual rather than expected phenomena". They point out 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. Chalmers et al (1977) have also discussed bias in relation to the studies of anticoagulant therapy for myocardial infarction.

Lastly, Zelen (1980) has suggested that 5% of reported clinical trials will be false positive results. This would only be true 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 provide, it is hard to see how a proper estimate of this sort can be obtained.

CONCLUSIONS

The main weakness of trying to combine the results of several studies is the problem of publication bias of unknown magnitude. Nevertheless, analyses of this sort can yield useful information, and, if it is possible to incorporate information about patient characteristics, differences in therapy etc., may provide additional information and suggest hypotheses for further investigation. Pressure should be brought to bear on journals, however, to realise the danger of discriminating against 'negative studies'.

REFERENCES

ARMITAGE P. STATISTICAL METHODS IN MEDICAL RESEARCH. OXFORD: BLACKWELL, 1971;427-33.

BAKER RJ AND NELDER JA. THE GLIM SYSTEM. RELEASE 3. OXFORD: NUMERICAL ALGORITHMS GROUP, 1978.

CHALMERS TC, KOPP RS, GRADY GF. A NOTE OF FATALITY IN SERUM HEPATITIS. GASTROENTEROLOGY 1965;49:22-6.

CHALMERS TC, MATTA RJ, SMITH H, KUNZLER AM. EVIDENCE FAVORING THE USE OF ANTICOAGULANTS IN THE HOSPITAL PHASE OF ACUTE MYOCARDIAL INFARCTION. NEW ENGL J MED 1977;297:1091-6.

CHALMERS TC, SMITH H, BLACKBURN B, SILVERMAN B, SCHROEDER B, REITMAN D, AMBROZ A. A METHOD FOR ASSESSING THE QUALITY OF A RANDOMIZED CONTROL TRIAL. CONTR CLIN TRIALS 1981;2:31-49.

FISHER RA. STATISTICAL METHODS FOR RESEARCH WORKERS. 9TH EDITION. EDINBURGH: OLIVER & BOYD, 1944;99.

MANTEL N AND HAENSZEL W. STATISTICAL ASPECTS OF THE ANALYSIS OF DATA FROM RETROSPECTIVE STUDIES OF DISEASE. J NAT CANCER INST 1959;22:719-48.

MAXWELL C. CLINICAL TRIALS, REVIEWS, AND THE JOURNAL OF NEGATIVE RESULTS. BR J CLIN PHARMACOL 1981;1:15-8.

MILLER RG. COMBINING 2 x 2 CONTINGENCY TABLES. IN: MILLER RG, EFRON B, ET AL, EDS. BIOSTATISTICS CASEBOOK. NEW YORK: WILEY, 1980.

MOERTEL CG AND REITEMEIER RJ. ADVANCED GASTROINTESTINAL CANCER/CLINICAL MANAGEMENT AND CHEMOTHERAPY. NEW YORK: HOEBER MEDICAL DIVISION, HARPER AND ROW, 1969;73.

PETO R. CLINICAL TRIAL METHODOLOGY. BIOMEDICINE (SPECIAL ISSUE) 1978;28:24-36.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomised clinical trials requiring prolonged observation of each patient. II. Analysis and examples. Br J Cancer 1977;35:1-39.

Rogers SC and Clay PM. A statistical review of controlled trials of imipramine and placebo in the treatment of depressive illness. Br J Psychiat 1975; 127:589-603.

Schwartz D, Flamant R, Lellouch J. Clinical trials. London: Academic Press, 1980.

Woolf B. On estimating the relation between blood group and disease. Ann Hum Genet 1955;19:251-3.

Zelen M. Guidelines for publishing papers on cancer clinical trials: responsibilities of editors and authors. Draft paper for UICC project. 1981.

Table I. 90% range of observed % improvement on treatment $(p_1 - p_2)$ for different values of p_1 and p_2, with n subjects in both treatment and control groups.

	p_1	p_2	n (per group) 40	75	100	250	500	1000	2500	True % Differ
$p_2 =$	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
	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
	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
$p_2 = 0.75\,p_1$	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
	0.25	0.1	−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
$p_2 = 0.9\,p_1$	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
	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
	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

Table II. Probability of observing $p_2 > p_1$ when $p_2 = 0.5p_1, 0.75p_1, 0.9p_1$ for different sample siz

	p_1	p_2	n (per group) 40	75	100	250	500	1000	2500
$p_2 = 0.5\,p_1$	0.5	0.25	1%	–	–	–	–	–	–
	0.25	0.125	7%	2%	1%	–	–	–	–
	0.15	0.075	14%	7%	5%	–	–	–	–
$p_2 = 0.75\,p_1$	0.5	0.375	13%	6%	4%	–	–	–	–
	0.25	0.19	25%	18%	14%	4%	1%	–	–
	0.15	0.11	31%	25%	21%	11%	4%	1%	–
$p_2 = 0.9\,p_1$	0.5	0.45	33%	27%	24%	13%	6%	1%	–
	0.25	0.225	40%	36%	34%	25%	18%	10%	2%
	0.15	0.135	42%	40%	38%	32%	25%	17%	6%