MEDICAL PRACTICE

Statistics in Medicine

Current issues in the design and interpretation of clinical trials

STUART J POCOCK

Abstract

Though there have been considerable improvements in the use of statistical methods for clinical trials in recent years, there remain major practical difficulties in the design and interpretation of many trials. This paper concentrates on problems relating to randomisation, the overemphasis on significance testing, and the inadequate size of many trials. Each topic is illustrated by examples from recent trials.

Introduction

This article concentrates on three major statistical problems that remain a common cause of difficulty in the design and interpretation of clinical trials.

The first problem is randomisation. Many treatments are still being developed without properly randomised controlled trials. Unequal randomisation might be used more often in early (phase II) trials for which there exist substantial historical data on the standard treatment. The required degree of stratification in the design of randomised trials is still not clear.

The second topic is an overemphasis on significance testing. Most trial reports in medical journals rely heavily on significance tests and pay inadequate attention to estimating the potential magnitude of treatment differences (for example, confidence limits are underused). The abundant and selective use of significance tests in clinical trials may greatly increase the risk of false positive claims. Particular problems concern the use of multiple endpoints, interim analyses, and subgroup analyses.

The third problem concerns the size of trials. Many trials remain far too small to provide adequate power to detect relevant treatment differences. When power calculations have been used there is a danger of defining unduly large treatment differences under the alternative hypothesis to achieve the convenient requirement of small sample sizes. Small trials require huge observed differences to be statistically significant, and non-significant small trials are less likely to be published; these two facts lead to major "publication bias" in reports of clinical trials.

The above problems are practical. Whereas further developments in statistical methods will continue to occur in clinical trials, such theoretical advances may be of secondary importance compared with the need to convey the essentials of good statistical practice to a wider audience. Thus it is important that professional biostatisticians make a greater attempt to communicate effectively with clinicians and other non-statistical collaborators rather than concentrate on mathematically oriented topics of only peripheral relevance to medical and biological research.

The remainder of my article includes examples from actual trials in a broader discussion of the three topics.

Department of Clinical Epidemiology and General Practice, Royal Free Hospital School of Medicine, London NW3 2PF

STUART J POCOCK, MSC, PHD, reader in medical statistics

Randomisation

CHAOS CAUSED BY NON-RANDOMISED TRIALS

In Britain there has recently been considerable controversy over clinical trials to evaluate periconceptional multivitamin treatment for preventing births of children with neural tube defects to high risk mothers. There have been two non-randomised trials in which pregnant women with a previous neural tube defect birth who had had periconceptional multivitamin supplements were compared with an unsupplemented control group.[1 2] The combined results for the two trials were:

	Multivitamin group	Control group	
Total No of births	397	493	$p<0\cdot0003$ one sided test
No of neural tube defect births	3 (0·8%)	23 (4·7%)	

The authors argued that such a highly significant apparent benefit justified the use of multivitamins for all subsequent such pregnancies, as they considered that bias in the methodology of the trial could not account for the observed treatment difference. In these multicentre trials, however, the control group included some women who had elected not to take supplements and also included more women from high risk areas—for example, Northern Ireland —so that there is ample reason to claim that the treatment comparison may have been severely biased.

The problem is that there is no way of determining whether all or part of the treatment difference is attributable to bias, so that there remains uncertainty whether a subsequent randomised trial is ethically justified. Hence the Medical Research Council Vitamin Study (begun in 1983) has provoked major arguments over whether it was ethically acceptable to randomise some patients to a non-supplemented control group. There has been considerable public opposition to this randomised trial. For instance, the *Daily Telegraph* (30 March 1984) carried the headline "Dummy-pill risk of handicapped babies 'immoral' " in an article claiming that the trial was unnecessary. Such views must be regarded with some sympathy: to the layman the results of the earlier trials look impressive. Nevertheless, it is a dangerous precedent to argue that future therapeutic practice should be determined by such inadequate trials.

The MRC trial is going ahead and had recruited 132 patients by May 1984.[3] The intended accrual is over 2000 patients, and several hospital ethical committees have refused permission for patients to be entered, so that considerable doubt remains whether this (or any other) trial will succeed in resolving this therapeutic dilemma.

There is no simple "right answer" to the question of when a randomised trial is justified in the face of such suggestive (but perhaps grossly exaggerated) prior evidence. The real problem is that non-randomised trials may considerably hinder clarifying therapeutic issues. Undoubtedly the first trial of multivitamins should have been organised properly with a randomised control group. I hope that this unfortunate example will increase awareness of the need to undertake early randomised trials before uncontrolled or poorly controlled data lead to overenthusiastic (and possibly mistaken) support for a new treatment.

UNEQUAL RANDOMISATION

In many trials of cancer chemotherapy the aim is to assess the value of a new treatment when there is already a standard treatment with substantial experience. The approach is often to conduct an uncontrolled (phase II) trial of the new agent and then to consider a randomised (phase III) trial later. Nevertheless, this may pose problems: uncontrolled trials may produce wildly overoptimistic results and hence decisions on which drugs to put in phase III trials are based on inadequate data.

A useful compromise is to undertake a randomised controlled phase II trial in which most patients (say, two thirds) are assigned to the new treatment. Compared with the conventional randomised trial, such an unbalanced design permits greater experience of the new treatment, an important consideration in the early stage of testing drugs. At the same time the element of randomisation helps to ensure that patient selection, ancillary care, and evaluation of response conform to accepted standards.

Some recent experience of this approach has been in the design of trials for advanced breast cancer. One commonly accepted combination treatment is VAC (vincristine, adriamycin, and cyclophosphamide), and apparently investigators are encouraged to undertake trials of new single agents or combinations that also randomise one third of patients to VAC. Randomisation ratios of 3:2 or 2:1 preserve adequate power if comparative analyses are to be undertaken. For instance, if a trial adopts a 2:1 rather than 1:1 randomisation ratio (while preserving the same total number of patients) then type II errors of 0·05 and 0·5 would be increased to 0·075 and 0·55 respectively.[4] Nevertheless, in such randomised phase II trials the observed response and toxicity rates with the new treatment are of interest in their own right.

A further issue is that response data for the randomised control group might potentially be supplemented by response data from a larger body of historical controls. Given the potential bias in historical controls, caution is needed before undertaking any formal analysis pooling the two sets of control data. If the historical controls were collected in a similar fashion, however—that is, from a previous randomised trial with the same eligibility and response criteria—then some formal combining of control data may be worth considering. In these circumstances it is still appropriate to give greater weight to the randomised controls.

STRATIFIED RANDOMISATION

Despite the considerable developments in the past decade in statistical methods for stratified randomisation,[4,5] there has also been controversy over whether stratification is necessary. The diversity of approaches has been illustrated in an international survey of randomisation in major cancer trial centres.[6] For example, 10 centres had a current trial in primary breast cancer; two of these trials had no stratification factors (except for institution), whereas one trial had five stratification factors, with a total of $2\times2\times2\times2\times3=48$ strata. The most common decision was to have two stratification factors, each at two levels.

All of these centres were experienced at coordinating multicentre trials, so that the different approaches were due to conscious decisions rather than to oversight. Should it be argued that the two trials without stratification were inferior? Probably not: as trials in breast cancer require a substantial number of patients the risk of serious imbalance between randomised groups with respect to prognostic factors is negligible. On the other hand, close prognostic comparability of randomised groups achievable by extensive stratification enhances the acceptability of simple unstratified analyses. To achieve such closely matched groups, however, may require so called minimisation procedures. These are straightforward conceptually (they aim at minimising some overall measure of the treatment difference in prognostic factor distributions when registering patients but they increase the administrative burden.)

It may be sensible to avoid stratifying if there is uncertainty over the relevance or reliability of prognostic factors or if the trial has a simple organisation that might not cope well with complex randomisation. Nevertheless, for trials with a sophisticated and experienced organisation that have well defined prognostic factors the slight gains in statistical efficiency and the appeal of closely matching groups make it sensible to use stratified randomisation. The newer more complex methods based on a minimisation approach may enable effective balancing for more prognostic factors than would the conventional random permuted blocks within strata.

To take an overall view of the statistical design of clinical trials, it appears that stratification is a minor issue, particularly in large trials. For most trials it may be more directly profitable to use some form of stratified analysis—that is, adjustment for prognostic factors when analysing for treatment differences.

Overemphasis on significance testing

SIGNIFICANCE TESTS AND CONFIDENCE LIMITS

Some fundamental issues in significance testing are illustrated by the reported findings in the Lipid Research Clinics Coronary Prevention Trial.[8] A total of 3806 men with high serum cholesterol concentrations were randomised to treatment with cholestyramine (a drug that lowers cholesterol concentrations) or placebo. The following table presents the results for the prespecified primary endpoint, definite coronary heart disease, after an average follow up of 7·4 years:

	Placebo	Cholestyramine	
Total No of men	1900	1906	$p<0\cdot05$ one sided test
No with definite coronary heart disease	187 (9·8%)	155 (8·1%)	

There were fewer coronary events with cholestyramine, but the comparison is statistically significant at the 5% level only if a one sided test is adopted. This borderline significance should convey the message that uncertainty remains whether cholestyramine reduces the risk of coronary heart disease. Unfortunately, in an aspect of the epidemiology of heart disease that provokes some strongly committed views the study findings have been interpreted in a more dogmatic fashion, particularly by the press. For instance, the *Daily Telegraph* (January 1984) carried an article headed "Heart disease study points finger at cholesterol," which went on to state, "the study . . . proved that lowering cholesterol in the blood reduced both the signs of heart disease and fatal heart attacks."

This example illustrates that the accept/reject philosophy of significance testing based on the "magical" $p=0{\cdot}05$ barrier remains dominant in the minds of many non-statisticians. Indeed, the statistical profession itself has not always been effective enough in overcoming this misconception. We need continually to assert that p-values are only a guideline to the strength of evidence contradicting the null hypothesis of no treatment difference and that they should not be regarded as indicating proof of treatment efficacy. It may be more productive to shift the emphasis towards estimation methods such as confidence limits.

In this trial the observed percentage reduction in risk in the group given cholestyramine was $(9{\cdot}8-8{\cdot}1)\times 100/9{\cdot}8=17\%$, which is adjusted to 19% after a stratified analysis. The authors reported 90% confidence limits for this percentage reduction of +3% and +32%. One might argue in favour of the more conventional 95% confidence interval, which would have included zero reduction and helped to emphasise the lack of conclusive proof for treatment benefit.

It has become fashionable to use such percentage reductions in risk, but for potential patients the difference in risk may be more meaningful. In this trial the observed difference in favour of cholestyramine was $9{\cdot}8\%-8{\cdot}1\%=1{\cdot}7\%$, with 95% confidence limits of $-0{\cdot}1\%$ and $+3{\cdot}5\%$. Either way, the use of confidence limits readily conveys the considerable uncertainty about the effect of cholestyramine on coronary risk.

One issue highlighted by this trial is the use of one sided testing. Ideally the distinction between one and two sided tests should be unimportant if p-values are interpreted as informal guidelines where there is no radical distinction drawn between $p=0{\cdot}06$ and $p=0{\cdot}04$. Even so it would make sense always to use two sided tests, as one sided testing rests on a subjective judgment that an observed difference in the opposite direction (for example, against cholestyramine) would be of no interest whatsoever. In particular, the use of a two sided test in this trial would have resulted in $0{\cdot}05<p<0{\cdot}1$, which might have helped to tone down some of the more exaggerated claims derived from this well conducted and valuable but inconclusive trial.

The problem with significance testing would not be so bad if there was only one test per trial. Many trials, however, generate a multiplicity of data, which may provoke a plethora of significance tests. It is worth focusing on three issues: interim analyses, subgroup analyses, and multiple end points.

INTERIM ANALYSES AND STOPPING RULES

Many trials continue without formalised stopping rules, with the consequent risk of exaggerating both the significance and the magnitude of treatment effects. Our experience at the Royal Free Hospital with a trial comparing D-penicillamine and placebo in treating primary biliary cirrhosis illustrates some of the problems associated with interim analyses.[9] The trial began in 1975 with survival as the main endpoint. In line with the arguments above, the randomisation was unbalanced with three fifths of patients assigned to D-penicillamine.

The changing pattern of survival results was as follows:

	No of deaths/No of patients			
	Placebo	D-penicillamine	χ^2	p
First analysis, summer 1980	8/32	2/55	9·1	0·003
Publication, 1981	10/32	5/55	7·2 (logrank)	0·01
Most recent analysis, 1984	16/37	18/61	3·0 (logrank)	0·08

In the summer of 1980 the results looked interesting and so the investigators were encouraged to seek a first analysis of the data despite the few deaths overall. This showed a highly significant result which suggested that, even though no formal stopping rule had been planned, patient entry should be stopped and the results published. The published findings in 1981 indicated that the logrank χ^2 had been reduced in the few extra months of follow up but that the treatment difference was still significant at the 1% level.[9]

Three years further on the latest analysis of patient survival in January 1984 still shows a lower death rate with D-penicillamine, but the difference is only marginally significant ($p=0{\cdot}08$). These updated findings leave greater uncertainty regarding the value of D-penicillamine, a toxic drug that needs to show clear survival benefit before its general use could be recommended. One possibility worth considering is that there is a genuine decline in the treatment difference over time—that is, does D-penicillamine delay some deaths that would otherwise occur in the first year or two of follow up? Inspection of the current life table plots, however, gives no indication of such a "treatment-time interaction," and indeed trials in this disease need more patients to establish the true answer.

This trial illustrates a couple of general messages that need wider recognition:

(*a*) Interim trial publications claiming significant treatment differences will tend to exaggerate the true magnitude of the treatment effect.

(*b*) Subsequent analyses (if possible) are likely to show a reduction in both the significance and magnitude of treatment differences.

Both these phenomena may be explained by the fact that interim publications are often timed (either deliberately or unwittingly) to reflect a "random high" in the treatment comparison. Unfortunately, this potential bias in the timing of publication is widespread as most trials have no formal policy on when to publish.

SUBGROUP ANALYSES

Clinical trials are sometimes accused of providing only global treatment comparisons, which may not be suited to the needs of individual patients. Hence there is always pressure to try to identify particular subgroups of patients who responded especially well (or badly) to a new treatment. The problems here are:

(*a*) trials can rarely provide sufficient power to detect such subgroup effects;

(*b*) medical publications tend erroneously to use separate significance tests for each subgroup rather than the appropriate (but more complex) tests of interaction; and

(*c*) there are often many possible prognostic factors from which to form subgroups, so that one has to guard against "data dredging."

To illustrate the problems of subgroup analysis I will refer to the Multiple Risk Factor Intervention Trial.[10] This randomised trial of 12 866 men at high risk of coronary heart disease compared special intervention aimed at affecting major risk factors (for example, hypertension, smoking, diet) and usual care in the community. The overall rates of coronary mortality after an average seven year follow up (1·79% with special intervention and 1·93% with usual care) are not significantly different. The trial report contains several subgroup analyses, the most striking of which is the following:

		No of coronary deaths/No of men (%)	
Hypertension	Electrocardiographic abnormality	Special intervention	Usual care
No	No	24/1817 (1·3)	30/1882 (1·6)
No	Yes	11/592 (1·9)	15/583 (2·6)
Yes	No	44/2785 (1·6)	58/2808 (2·1)
Yes	Yes	36/1233 (2·9)	21/1185 (1·8)

For those with hypertension and electrocardiographic abnormalities at initial screening it appears that the coronary death rate is higher in the special intervention group, whereas the three other subgroups show a difference in the opposite direction. At face value such a subgroup effect looks impressive and worthy of clinical interpretation. Nevertheless, a formal test for interaction[11] shows no significant departure from the null hypothesis that the logit difference in coronary death rates for the special intervention and usual care groups is the same in all four subgroups (p=0·1). Given that this was not the only subgroup analysis performed, we should assert that there are inadequate grounds for supposing that the special intervention harmed those with hypertension and electrocardiographic abnormalities.

The message is that we should be wary of overinterpreting subgroup analyses. It would be too extreme to suggest that they should be avoided altogether—rather that they should be used cautiously in a spirit of exploratory data analysis, provoking ideas to be confirmed (or refuted) in future studies.

MULTIPLE END POINTS

In many trials it is appropriate to record and analyse several different aspects of each patient's response to treatment. Descriptive statistics on such multiple end points may provide valuable insight into the pattern of progress of the disease with each treatment. Nevertheless, the corresponding use of multiple significance tests carries an increased risk of a type I error—that is, false claims of treatment benefit. Accordingly, at the planning stage it has become standard practice to designate one primary end point whose significance test will be the main criterion for assessing treatment differences.

As an illustration, consider the report of a trial of 1232 men in Oslo, Norway, at high risk of coronary heart disease who were randomly assigned to the intervention or control group.[12] The intervention was recommendations to change diet and stop smoking. After five years the mortality and cardiovascular events in each group were as follows:

	Intervention group (604 men)	Control group (628 men)	p
Sudden death	3	12	0·02
Fatal myocardial infarction (MI)	3	2	
Fatal MI+sudden death	6	14	0·09
Non-fatal MI	13	22	0·15
Total coronary events	19	36	0·03
Fatal stroke	2	1	
Non-fatal stroke	1	2	
Total cardiovascular events	22	39	0·04
Total cardiovascular deaths	8	15	0·17
Total mortality	16	24	0·25

The above seven significance tests would pose considerable problems if all were presented on equal terms. In fact, the authors prespecified total coronary events as the primary end point, in which case they interpreted this single test at p=0·03 as evidence for intervention being beneficial.

Interestingly, one sudden death in the control group was "unexplained," in that confirmatory evidence of a coronary cause could not be obtained. If this death was not related to coronary disease it might be considered to have been unrelated to intervention. Let us suppose that this man had been randomised to the intervention group instead—then if this one sudden death had still occurred the significance tests for sudden death, total coronary events, and total cardiovascular events would all have become non-significant at the 5% level. Thus the statistical significance of this trial depends on one possibly unavoidable death being in the control group. This illustrates how fickle statistical significance may be, particularly if one were to rely too heavily on specific cut off points such as $p<0·05$.

Size of trials

All too frequently statisticians claim that most trials do not have enough patients to provide a reliable comparison of treatments. It would help to emphasise this fact if trial publicatons had to indicate the uncertainty of therapeutic differences by using interval estimation methods such as confidence intervals. Nevertheless, once the results of a trial are analysed and published it is too late to improve things.

Hence greater emphasis should be given at the planning stage, where power calculations should be used realistically. Unfortunately power calculations have the habit of producing unduly large sample sizes which are incompatible with the number of patients available. It is tempting to "improve the situation" by modifying the arbitrary levels of power and treatment difference to be detected, but this may lead to overoptimistic specifications.

For instance, the Oslo trial specified a 60% chance of detecting a 50% reduction in coronary events in the intervention group at the 5% level of significance, which required a trial size of 1230 men.[12] Such a dramatic reduction in coronary heart disease by diet and smoking intervention is highly desirable, but is it realistic to expect such a large effect? A 30% risk reduction would also be important to detect, but a trial size of 1230 men has only about 30% power of picking up such an effect as being significant at the 5% level—that is, if a true 30% risk reduction were to exist, this would be detected as significant only if the observed difference happened by chance to be somewhat greater. Thus, although recruitment of 1230 patients is a substantial undertaking, in the context of primary prevention trials for coronary heart disease such a sample size is inadequate.

Other trials on this question, such as the multiple risk factor intervention trial, have not shown such large benefits of intervention, so that it seems that this Oslo trial, though well executed in all other respects, may have achieved an inflated point estimate of risk reduction (47%) because of the random error inherent in having only 55 coronary events overall.

Evidently prevention trials in heart disease require particularly large numbers of patients compared with therapeutic trials for other diseases. Nevertheless, the deficiency in patient numbers in clinical trials is a general phenomenon whose full implications for restricting therapeutic progress are not widely appreciated. The fact is that trials with truly modest treatment effects will achieve statistical significance only if random variation conveniently exaggerates these effects. The chances of publication and reader interest are much greater if the results of the trial are statistically significant. Hence the current obsession with significance testing combined with the inadequate size of many trials means that publications on clinical trials for many treatments are likely to be biased towards an exaggeration of therapeutic effect, even if trials are unbiased in all other respects. Such "publication bias" and its liability to produce an excess of false positive findings has been reported elsewhere.[4 13 14]

In a short paper it is not possible to explore fully these fundamental statistical problems affecting clinical trials. Nevertheless, I hope that airing such issues will prove thoughtprovoking and may encourage colleagues, both clinical and statistical, to an increased awareness of the subtle biases that may arise in published reports of clinical trials.

I am grateful to Dr Owen Epstein for permission to include recent unpublished results in his trial for primary biliary cirrhosis. I thank Derek Cook, Stephen Evans, Nancy Geller, Sheila Gore, and Gerry Shaper for their valuable comments on a first draft. I am also grateful to the referee for helpful comments.

References

1 Smithells RW, Shepperd S, Schorah CJ, *et al*. Possible prevention of neural tube defects by periconceptional vitamin supplementation. *Lancet* 1980;i:339-40.

2 Smithells RW, Shepperd S, Schorah CJ, *et al*. Vitamin supplementation and neural tube defects. *Lancet* 1981;ii:1425.

3 Medical Research Council. Vitamin study. *Lancet* 1984;i:1308.

4 Pocock SJ. *Clinical trials: a practical approach*. New York: Wiley, 1983.

5 Simon R. Restricted randomization designs in clinical trials. *Biometrics* 1979;35:503-12.

6 Pocock SJ, Lagakos SW. Practical experience of randomization in cancer trials: an international survey. *Br J Cancer* 1982;**46**:368-75.

7 White SJ, Freedman LS. Allocation of patients to treatment groups in a controlled clinical study. *Br J Cancer* 1978;**37**:849-57.

8 Lipid Research Clinics Program. The lipid research clinics coronary prevention trial results. *JAMA* 1984;**251**:351-74.

9 Epstein O, Lee RG, Bass AM, *et al*. D-penicillamine treatment improves survival in primary biliary cirrhosis. *Lancet* 1981;i:1275-7.

10 Multiple Risk Factor Intervention Trial Research Group. Multiple risk factor intervention trial: risk factor changes and mortality results. *JAMA* 1982;**248**:1465-77.

11 Halperin M, Ware JH, Byar DP, *et al*. Testing for interacton in an I×J×K contingency table. *Biometrika* 1977;**64**:271-5.

12 Hjermann I, Holme I, Velve Byre K, Laren P. Effects of diet and smoking intervention on the incidence of coronary heart disease. *Lancet* 1981;ii:1303-10.

13 Peto R, Pike MC, Armitage P, *et al*. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. 1. Introduction and design. *Br J Cancer* 1976;**34**: 585-612.

14 Zelen M. Guidelines for publishing papers on cancer clinical trials: responsibilities of editors and authors. *Journal of Clinical Oncology* 1983;**1**:164-9.

(Accepted 2 October 1984)