Br. J. clin Pharmac. (1985), **19**, 155–160

Is the clinical trial evidence about new drugs statistically adequate?

J. M. BLAND
Department of Clinical Epidemiology and Social Medicine, St George's Hospital Medical School, London
D. R. JONES*
Department of Community Medicine, Westminster Medical School, London
S. BENNETT
Department of Applied Statistics, University of Reading, Reading
D. G. COOK
Department of Clinical Epidemiology and General Practice, Royal Free Hospital Medical School, London
A. P. HAINES**
Department of Community Medicine, Middlesex Hospital Medical School, London
ALISON J. MACFARLANE
National Perinatal Epidemiology Unit, Radcliffe Infirmary, Oxford

1 The statistical adequacy of all papers published in the period 1976–80 describing clinical trials of five non-steroidal anti-inflammatory and two analgesic drugs introduced into the UK market in 1978 and 1979 has been assessed using a checklist of simple criteria.

2 Most trials were reported to be randomised and double-blind.

3 Trial designs were less satisfactory in other important respects; the sample size of most trials was inadequate to demonstrate superiority of the new drug compared with an active control therapy.

4 The period of treatment assessment was short in view of the likelihood of prolonged prescription of drugs in these classes.

5 It is suggested that licensing authorities should demand higher standards of clinical trial evidence offered in support of new drugs.

Keywords clinical trials size of trial statistical methods non-steroidal anti-inflammatories registration procedures

Introduction

The withdrawal of benoxyprofen (Opren) and other drugs from the UK market has highlighted disquiet about the quality of evidence used to support the introduction of new drugs. Clinical trials form a major part of this evidence. In this paper we examine a series of clinical trials relating to a particular set of new drugs, report some deficiencies of the trials and discuss some of the reasons these deficiencies arise.

* Correspondence: Social Statistics Research Unit, Department of Mathematics, The City University, London EC1, UK.
** Present address: Department of General Practice,
St Mary's Hospital Medical School, Central Middlesex
Hospital, London NW10, UK.

The study reported here addresses one important aspect of the more general question:

> 'Are new drugs introduced on the market useful 'innovations' or a net gain for anyone other than the drug companies?'

In answering this question the effectiveness of new drugs, their therapeutic advantages over existing products, their cost and their safety all need to be considered.

In this paper we examine the evidence published to support the effectiveness and therapeutic advantage of new drugs. We have assessed the quality of the clinical trial evidence supporting the introduction of drugs, in particular some statistical aspects of clinical trial design. Such is the volume of the published material that we have had to limit the investigations to specific groups of drugs and diseases, and to published trials in order to ensure access to the material.

We chose to study drugs prescribed for patients suffering from arthritis and rheumatism. Arthritis and rheumatism were chosen because together they form the most commonly self-reported cause of limiting long standing illness and the associated drug market is an important one. For example, although the validity of the self-diagnosis is open to some question, amongst General Household Survey respondents aged over 15 years, 3.7% reported arthritis and rheumatism over a 3 month period, out of 20.0% who reported any cause (Office of Population Censuses and Surveys, 1973). Arthritis and rheumatism accounted for 15.8 million prescriptions in England and Wales in 1980 (Department of Health and Social Security, 1982).

The type of study considered is the comparative (phase III) clinical trial on humans. Such trials are carried out on drugs of demonstrated activity to compare new drugs with established therapies or where appropriate with no treatment, in the conditions of normal use (see, for example, Schwartz *et al.* (1980); Johnson & Johnson (1971)).

The restriction to published material may mean that some relevant trials are omitted. Under the Medicines Act (1968) applications for a product licence must be supported by evidence from experimental and animal studies and from clinical trials. The DHSS notes on applications for product licences state: (Department of Health and Social Security, 1980):

> 'In most cases evidence of efficacy will be from controlled trials. In addition, where a product is to be administered on a long-term basis (e.g. an oral hypoglycaemic agent) evidence of its long-term safety and efficacy in a substantial number of patients is needed. This can be provided from open trials in appropriate cases.'

However, there is no requirement that this evidence should be published, and undoubtedly some is not.

Methods

The sampling frame chosen was intended to comprise all comparative clinical trials on new non-steroidal anti-inflammatories and analgesic drugs introduced into the UK market in 1978 and 1979 for treatment of (amongst other diseases) arthritis and rheumatism.

There were problems in obtaining an agreed list of these new drugs. Three hospital drug information centres and two clinicians produced four related but far from identical lists. These problems arose from difficulties in deciding what was a 'new' drug and from different interpretations of the disease area. The final list was a compromise between these lists and comprised seven drugs: buprenorphine, diclofenac, diflunisal, fenclofenac, nefopam, salsalate and tolmetin. Of these, buprenorphine and nefopam were evaluated primarily in trials of treatment of post-operative pain rather than rheumatic pain, but all trials have been included in the present study.

The sample consisted of all comparative trials on these drugs listed in Index Medicus, 1976–80 inclusive, together with the proceedings of a conference on fenclofenac (Royal Society of Medicine, 1977). The conference proceedings contained nearly all the published material on this drug. The trials examined are thus likely to include most of the published evidence from comparative trials most relevant to the introduction of the drug into the UK market. Some of the trials reviewed may not, however, have been cited in a licence application. The publications appeared in a very wide range of journals (including many of the leading journals).

For the seven drugs, 76 publications were found describing trials which met the sampling criteria. Some of these contained reports of two trials, typically one of the drugs *vs* placebo and another *vs* an active treatment. There were 80 trials altogether. The number of papers and trials for each drug are shown in Table 1.

The evaluation criteria concentrate on the statistical aspects of the trial rather than clinical or pharmacological issues such as the appropriateness of the alternative treatment. The criteria are concerned with trial design rather

Table 1 Size of the sample of trials evaluated

Drug (Approved names)	*Papers*	*Trials*
Diclofenac	18	18
Salsalate	2	2
Buprenorphine	13	13
Diflunisal	3	3
Tolmetin	9	9
Nefopam	14	15
Fenclofenac	17	20
Total	76	80

than analysis, so the emphasis of this study was rather different from that of several other surveys of statistics in medical literature such as those of Schor & Karten (1966), Gore *et al.* (1977) and White (1979), but broadly in keeping with that of Gardner *et al.* (1983). This emphasis reflects the critical nature of design in clinical trial methodology; a report of a well designed trial incorrectly analysed can often be re-analysed but no re-analysis can correct basic design faults.

The criteria were based on those of Gifford & Feinstein (1969) in their classic study of anticoagulants in acute myocardial infarction, with some additions and modifications. Mahon & Daniel (1964) and Lionel & Herxheimer (1970) have provided comparable checklists. The criteria used to judge the published account of the trial were:

1. clear statement of diagnostic criteria;
2. in multi-centre trials, co-ordination of ancillary treatments at different hospitals;
3. experimental trial as opposed to an observation study;
4. concurrence of controls;
5. random allocation to treatment;
6. stratification by prognostic factors;
7. clear statement of criteria for outcome measurement;
8. double-blindness;
9. statement of reason for choice of sample size;
10. clear statement of source of patients.

The additional items recorded were:

11. sample size;
12. length of treatment;
13. length of post-treatment follow-up;
14. nature of alternative treatment used;
15. two-sample cross-over or parallel group design;
16. source of funding;
17. involvement of a statistician;
18. outcome of the trial.

Item 17 was included because, as several of the authors are statisticians often involved in clinical trials, we were interested to know whether the involvement of a statistician improved the trial by our criteria.

Five of the six authors assessed trial reports. For each drug, all of the relevant papers were evaluated by one of the five assessors, each assessor evaluating papers about one or two drugs from the list. As a check on consistency of the evaluations made by the assessors, a random sample of ten papers was drawn and each of these papers assessed by all five assessors. This is particularly important in view of the clearly subjective nature of some of the evaluation criteria. It was thought unlikely that serious disagreements would be missed in this subsidiary test of 50 assessments. Agreement was perfect or good for most criteria, but only moderate for assessments of outcome criteria (item 7: Cohen's $\kappa = 0.38$, see Fleiss (1981)) and source of patients (item 10: $\kappa = 0.47$). Two items, assessments of diagnostic criteria (item 1: $\kappa = 0.07$) and stratification by prognosis (item 6: $\kappa = -0.19$), gave very poor agreement and they have been discarded from further consideration. One of the original criteria (3) of Gifford and Feinstein was met by all trials due to our method of selection. There were too few multicentre trials in the study for analysis of (2) to be worthwhile.

Results

On some criteria the trials were generally satisfactory. All trials had concurrent controls. Random allocation was usual in these trials, being stated to have taken place in 69 (86%) of them. It may be that those which did not state that random allocation was used did in fact use it. Double-blind techniques were frequently used in these trials, being reported for 65 (81%) trials. Most of those trials not double-blind were single-blind, using blind assessment but distinguishable treatments; most occurred in trials of one drug, fenclofenac. Of the 80 trials, 31 were of crossover design.

On the whole, however, the exhortations of Bradford Hill (1953) and many others for clinical trials to be randomised and double-blind seem to have had some effect. Criteria for the outcome measure were also clearly stated in 67 (84%) trials. None of these criteria (randomisation, double-blindness, clear outcome measure) was related to the involvement of a statistician in the trial.

Other criteria were less well met. The source of patients was indicated in only 28 of the 80 trials (35%). For 30 (38%) trials the source of funding was stated to be a drug company. The other trial reports usually contained no statement whatever about funding.

One major defect in trial design and reporting concerned sample size. In 78 (98%) trials there was no statement about the criterion for the choice of sample size. The only statements concerning the power of the trial were in the two sequential trials included, where this is essential for the design. The samples used were mostly small. The frequency distribution of total sample size used in the trials is shown in Table 2. The majority of trials used fewer than 30 patients altogether. The few which are apparently large do not in fact have high power. The largest, comprising 240 patients in total, involved eight sub-groups receiving five different drugs, so most comparisons are between two groups of 30 each. Other things being equal, a small sample size leads to a trial with low power. This becomes a particular problem when the invalid conclusion that no treatment difference exists is drawn from a non-significant result.

Table 2 Total sample sizes

1- 10	6
11- 20	16
21- 30	22
31- 50	12
51-100	17
101-150	6
151-200	1
201-250	1
Total	80

Table 3 shows the outcome of the trials consisting of comparison of the new drug with placebo. We have recorded as showing a significant difference all trials where *any* outcome variable indicating the effectiveness of the treatment was reported as showing a significant difference (at the 5% level), whether or not this was in favour of the new drug.

In most of these trials there were many comparisons between treatments; some differences are likely to be significant by chance, even if there are no true differences between the treatments being compared. Most of the trials in which a new drug was compared with a placebo showed the drug to perform significantly better than the placebo (Table 3). This is not surprising as the effectiveness of these drugs should already have been demonstrated in a Phase II trial. As a result the differences between drug and placebo are, in general, relatively large. However, in view of the relatively small sample sizes, and hence limited power of most of these trials, it is possible that some of these significant results are false positive (Type I) errors.

When the new drug was compared with another drug (Table 4) the majority of trials did not produce a significant difference (and some of those that did were in favour of the standard drug). There may be insufficient power to detect a small advantage over the alternative drug. Many reports appear to interpret a non-significant difference as evidence for the satisfactory nature of the new drug.

Table 5 shows the distribution of lengths of treatment. Some trials for post-operative pain inevitably have very short treatment times, usually 24 h or less. However, although non-steroidal anti-inflammatory drugs are intended to provide relief from symptoms of a chronic disease, in only seven of the 80 trials did length of follow-up after treatment exceed 1 day.

Table 3 Statistical significance of results: drug *vs* placebo comparisons

Significant drugs vs *placebo difference*	*Not significant*	*No test performed*	*Total*
21	2	1	24

Table 4 Statistical significance of results: drug *vs* drug comparisons

Significant drug vs *drug difference*	*Not significance*	*No test performed*	*Total*
23	44	1	68

Table 5 Distribution of length of test treatment (days)

	Number of trials
< 1	25
2- 7	9
8- 14	25
15- 28	6
29- 56	4
57- 91	7
92-365	4
Total	80

Discussion

It was reassuring to see that a large majority of the trials studied were reported as being double-blind, randomised, controlled trials. These methodological principles seem to be well established.

There were, however, two main areas of concern regarding the design of these trials: sample size and duration of treatment. It would be interesting to know the criteria by which the sample sizes were chosen. That the reason for choosing a sample size was seldom given suggests that the choice was not considered important by experimenters. It may be that the sample size was determined entirely by the number of patients available for the trial, and no effort was made to consider this further. The sample sizes themselves were large enough to produce significant differences compared with placebo in most of the placebo control trials, but in the trials where the control treatment was active most treatment comparisons did not result in significant differences.

In some trial reports it was incorrectly concluded that because the new drug was not significantly different in outcome from a standard treatment, in a trial including 20 or 30 patients, the new drug was effective and safe. A finding of 'not significant' simply means that a trial has failed to demonstrate that a difference exists. There may in fact be quite large differences, in either direction, and estimates of the size of such possible differences could easily be presented instead of the results of significant tests.

The second main area of concern in these trials relates to length of treatment. Most of these trials are of drug treatments which may be taken for many years by a patient with a chronic disease. Drugs should be tested in conditions of normal use, and normal use for most of these drugs is continual administration for a long period of time. It is therefore very disturbing to see so few trials in which assessment of the effectiveness of treatment continues for more than 1 month. Only clinical trials under 'normal' conditions can establish whether a new treatment is really superior to existing ones. However, only four of the trials studied had a treatment time longer than three months. Guidelines for the appropriate length of treatment in trials of drugs intended for treatment of chronic diseases would appear not to exist. We recognise that detection of *side effects* will, however, in general require a different approach, such as a phase IV, event monitoring study (see Inman, 1981).

The weakness of the clinical trial data available, on the evidence of this study, in support of the introduction of new drugs should be of concern to all involved in the prescribing of drugs. As we have noted here, there were problems in the sampling of drugs for this project. Some non-steroidal anti-inflammatory drugs were not included, and two drugs which were evaluated for the control of other types of pain were. However, the sample of drugs was chosen before any of the reports of trials relating to the drugs were inspected and it seems unlikely that the sample we have is highly atypical. This view is supported by the concordance between our conclusion and those of Hemminki (1981) who studied trials of psychotropic drugs, and Der Simonian *et al.* (1982) whose emphasis was on the quality of reporting of clinical trials. Many of the conclusions thus may well be true of clinical trials for new drugs in general. It may be argued that the demand for larger and longer clinical trials would further delay the introduction of beneficial new drugs, to the detriment of patient welfare, but introduction of inadequately tested drugs clearly may also be detrimental. It has been argued by Altman (1980) that trials which are so designed that they cannot provide clear answers to the questions asked are unethical, in that therapy based on their misleading results may harm patients and patients in the trial are exposed to risk for no good reason.

As we have already noted, not all clinical trial evidence cited in support of the introduction of a new drug will be published. However, it seems to us unlikely that the trials which remain unpublished are those which are best designed, with the largest samples sizes, the most rigorous data collection procedures, and the most clear-cut demonstrations of the superiority of the new product. It seems more likely that the opposite will be the case and that our survey of published literature is biased towards a favourable report on trial standards.

The development of a new drug is a costly procedure and insistence on adequate clinical trials would add to that cost. However, these trials do not take place in isolation but in the wider context of a pharmaceutical industry, with its need to make profits. This is reflected in the large research effort that goes into the development of 'me too' drugs. The many trials in this study which were content to conclude that the new product was no different from an older one are a striking manifestation of this major research objective.

One solution would be for licensing authorities to insist that any new drug to be released should be shown by adequate clinical trials to be an improvement over existing drugs in terms of effectiveness, convenience, side effects or cost as is required in some Scandinavian countries. Responsibility for improvements in the clinical trial evidence offered in support of new drugs rests jointly with the licensing authorities and the drug companies, and *all* such evidence should surely be available in the public domain.

We are pleased to be able to thank Dr Peter Williams, Dr Carol Black and the drug information services of St George's, Westminster and Northwick Park Hospitals who helped compile lists of new drugs and several other members of the Radical Statistics Health Group for their help with and comments on this project.

References

Altman, D. G. (1980) Statistics and ethics in medical research. Misuse of statistics is unethical. *Br. med. J.*, **281**, 1182–1184.

Bradford Hill, A. (1953). Observation and experiment. *New Engl. J. Med.*, **248**, 995–1001.

Department of Health and Social Security. (1982). *Health and Personal Social Services Statistics for England 1982*, (Table 5.25A) London: HMSO.

Department of Health and Social Security Medicines Division. (1980). *Medicines Act Leaflet MAL 2. Notes on applications for product licences.* London: DHSS.

DerSimonian, R., Charette. L. J., McPeek, B. & Mosteller, F. (1982). Reporting on methods in clinical trials. *New Engl. J. Med.,* **306**, 1332–1337.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*, 2nd ed. New York: Wiley.

Gardner, M. J., Altman, D. G., Jones, D. R. & Machin, D. (1983). Is the statistical assessment of papers submitted to the 'British Medical Journal' effective? *Br. med. J.*, **2**, 1485–1488.

Gifford, R. H. & Feinstein, A. R. (1969). A critique of methodology in studies of anticoagulant therapy for acute myocardial infarction. *New Engl. J. Med.*, **280**, 351–357.

Gore, S. M. Jones, I. G. & Rytter, E. C. (1977). Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. *Br. med. J.*, **1**, 85–87.

Hemminki, E. (1981). Quality of reports of clinical trials submitted by the drug industry to the Finnish and Swedish Control Authorities. *Eur. J. clin. Pharmac.*, **19**, 157–165.

Inman, W. H. W. (1981). Postmarketing surveillance of adverse drug reactions in general practice. *Br. med. J.*, **282**, 1216–1217.

Johnson, F. N. & Johnson, S. (1977). *Clinical trials.* Oxford: Blackwell Scientific Publications.

Lionel, N. D. W. & Herxheimer, A. (1970). Checklist for assessing a therapeutic trial report. *Br. med. J.*, **3**, 637–640.

Mahon, W. A. & Daniel, E. E. (1964). A method for the assessment of reports of drug trials. *Can. med. Ass. J.*, **90**, 565–569.

Office of Population Censuses and Surveys. (1973). *The General Household Survey* SS457, (Table 8.11). London: HMSO.

Royal Society of Medicine. (1977). A seminar on fenclofenac. Papers presented at the 14th. International Congress on Rheumatology, San Francisco, 28th. June, 1977. *Proc. Roy. Soc. Med.*, **70**, Supplement 6.

Schor, S. & Karten, I. (1966). Statistical evaluation of medical journal manuscripts. *J. Am. med. Ass.*, **195**, 1123–1128.

Schwartz, D., Flamant, R. & Lellouch, J. (1980), *Clinical trials.* London: Academic Press.

White, S. J. (1979). Statistical errors in papers in the 'British Journal of Psychiatry'. *Br. J. Psychiat.*, **135**, 336–342.

(Received August 28, 1984, accepted October 21, 1984)