SAMPLE SIZE FOR SHORT-TERM TRIALS OF ANTI-HYPERTENSIVE DRUGS

S. FREESTONE, J.H. SILAS & L.E. RAMSAY

University Department of Therapeutics, Royal Hallamshire Hospital, Sheffield S10 2JF

1 Controlled trials of antihypertensive drugs published in the *British Journal of Clinical Pharmacology* during 1979 and 1980 were examined. Studies comparing two or more active drugs or dosage regimens nearly always failed to separate the treatments significantly. The sample size (mean 19 patients) and power of these studies were too low.

2 When planning such studies the aims should be a power of at least 0.8; significance 0.05 or less; and to detect a difference between treatments of 10/5 mmHg, or 6.7 mmHg mean arterial pressure (MAP). The sample size needed can be derived readily from a nomogram if the standard deviation of differences (SDD) between BP measurements under trial conditions is known.

3 In five studies the SDDs were fairly constant despite different observers, patient groups and measuring devices, at approximately 14 mmHg systolic, 9 mmHg diastolic, and 9 mmHg MAP. Use of three BP measurements at each visit reduced the SDD by about 1 mmHg, and would reduce the sample size required by about 20%. Replicate BP measurements at separate visits would be expected to have a larger effect on the power of the study.

4 Published studies with negative results should give an estimate of the power of the study.

Introduction

The properties of antihypertensive drugs are often examined in short-term controlled studies, either to establish efficacy against placebo treatment, to compare the efficacy of two active drugs, or to study the dose-response by comparing different doses of the same drug. A brief scan of the literature reveals that such studies are very common, yet when planning such studies ourselves we found great difficulty in predicting the appropriate sample size. We obtained little specific help from the literature, although the general principles of determining sample size are clear (Altman, 1980). In particular the sample size required depends on the variability of blood pressure measurements and it is not clear what variability may be anticipated. It is also unclear whether replicate measurements of blood pressure, which are used commonly, are effective in reducing variability. Should we use a single measurement at a single visit? Could the sample size be reduced to an important extent by using replicate measurements at a single visit; single measurements at replicate visits; or replicate measurements at replicate visits?

This article is in two parts. In the first we examine controlled trials of antihypertensive drugs published during two years in the *British Journal of Clinical Pharmacology*, to examine the sample sizes used and the pattern of replicate BP measurements employed.

In the second part we consider the requirements for estimating sample size, examine the variability of BP in our own experimental data, consider the value of replicate measurements and use this information to draw some general conclusions about the sample size needed for different kinds of study.

Survey of published studies

Methods

We examined all controlled comparative studies of antihypertensive drugs published in the *British Journal of Clinical Pharmacology* in 1979 and 1980.

Results

There were 14 studies, 12 crossover and 2 of parallel group design (Table 1). There is clearly no consensus on the value of replicate BP measurements. In some studies a single measurement at a single visit was used; in one four replicate measurements at a single visit were used; in one a single measurement at three visits was used; and there were many other permutations.

In six studies an active drug was compared with

Table 1 Survey of published studies

Study	Sample size (completing trial)	Supine BP (Readings/visit)	Number of visits on each treatment	Outcome of study	
				✓ Placebo	Active drugs
1	11	1	3	P	N
2	15	2	1	P	—
3	17	2	1	—	N
4	10	1	2	—	N
5	42 (3 groups)	1	1	—	N
6	10	1	1	—	N
7	14	1	1	—	N
8	10	4	1	P	—
9	40	2	1	—	P
10	7	1	2	P	—
11	13	3	2	P	N
12	15	3	1	—	N
13	11	1	1	P	—
14	19	1	2	—	N

P positive, N negative

placebo, and in these the mean sample size was 11 patients (range 7–15). In each study the active drug was significantly superior to placebo, suggesting that the sample sizes, although small, were adequate. In ten studies two or more active drugs, or two or more dosage regimens of the same drug were compared, and the mean sample size was 19 patients (range 10–42). In only one study were the active treatments significantly different. In the other nine studies were the active treatments truly 'not different', or were the sample sizes too small to demonstrate a significant difference between treatments (type II error)?

Estimating sample size

Theory

Hills & Armitage (1979) have suggested that the sample size in a crossover study should be large enough for the first period to be analysed as a parallel group study, to allow for possible interaction between treatments and periods. Estimates of sample size for crossover and parallel group studies should therefore be identical. The requirements for estimating sample size are:

- the power of the study. This is the probability of obtaining a statistically significant result if there is a true difference between treatments, and is traditionally set at 0.8 (Hills & Armitage, 1979) or 0.9 (Bulpitt, 1975).
- the level of significance acceptable, generally set at < 0.05 .
- the minimum difference between treatments to be detected, which is set by the investigator.

- the standard deviation of the differences (SDD) between measurements under trial conditions. The SDD is a measure of the variability of BP measurements—the greater the SDD, the larger the sample size needed. Replicate measurements would be of value if they reduced the SDD and, therefore, the sample size required.

Given these four variables the sample size required can be estimated readily from the nomogram described by Altman (1980). In practice the SDD is the important unknown factor when estimating sample size, and we have been unable to obtain even approximate estimates from the literature. It can be determined by mounting a pilot study but this is time consuming and open to sampling error unless a large number of patients is studied.

Method

We examined the SDD between BP measurements from the run-in phase of five studies. The purpose of these measurements was to establish the pre-treatment baseline BP before randomisation. All patients were familiar with the clinic routine before entering so that contribution to variability from adaptation to the clinic or regression to the mean might be expected to be minimal. The BP had been measured by different trained observers in each study under actual clinical trial conditions. Random zero sphygmomanometers were used in four studies, and a semi-automated recorder, the accuracy of which has been previously proven (Silas *et al.*, 1980), was used in the fifth. Phase 5 diastolic BP (disappearance of sounds) was measured and mean arterial pressure (MAP) was calculated as diastolic BP plus $1/3$ pulse pressure.

Table 2 The standard deviation of differences between blood pressure measurements from the run-in phase of five studies

Study	Observer**	Number of patients	Initial BP (mmHg)	Interval (weeks)	Standard deviation of differences					
					Systolic		Diastolic		MAP	
					1 × 1	3 × 3*	1 × 1	3 × 3*	1 × 1	3 × 3*
1	TN.1	14	156/92	4	12.2	—	10.7	—	9.7	—
2	TN.2	39	167/103	2	15.4	—	10.8	—	9.7	—
3	D.1	22	186/100	3	14.6	12.9	5.8	4.8	7.1	6.5
4	D.2	13	141/81	2	12.9	14.5	7.9	7.2	9.2	9.0
5	SA	14	156/95	3	12.4	10.5	10.3	7.6	9.9	8.1

**TN = trained nurse; D = doctor; SA = Dinamap semi-automated recorder

* = mean of 3 readings on each occasion

Results

Using a single BP measurement at two visits the SDD was fairly consistent in the five studies (Table 2) despite the differences in observers, patient groups and measuring devices. The SDDs for systolic BP (~ 14 mmHg; range 12.2–15.4 mmHg) and MAP (~ 9 mmHg; range 7.1–9.9) were particularly consistent, although those for diastolic BP were more variable (5.8–10.8 mmHg). Use of a semiautomated recorder did not appear to reduce the SDD.

In three of the studies the effect on the SDD of using the mean of three readings at each visit rather than a single reading could be examined (Table 2). In general the use of three readings per visit reduced the SDD, but only by a small amount (~ 1 mmHg). This would have a relatively small effect on the sample size required as discussed later.

Calculating sample size

Example 1. A new diuretic is expected to have an antihypertensive effect similar to that of the thiazides, and the investigator wishes to establish this in a placebo-controlled study;

- let power = 0.9
- the significance level acceptable = 0.05
- difference to be detected (equivalent to thiazide) = 21/10 mmHg, i.e. Δ MAP = 13.7 mmHg (McMahon, 1978).
- SDD for MAP, using single measurements at single visits = 9 mmHg.

From the nomogram (Altman, 1980) the sample size needed is 20 patients, 10 randomised to the new drug and 10 to placebo.

Example 2. An investigator wishes to determine whether two drugs of proven efficacy differ in their antihypertensive effect. He must first decide what difference would be considered important in ordinary

practice and settles on a difference of 10/5 mmHg (Δ MAP 6.7 mmHg), then:

- power = 0.9 (for example)
- significance = 0.05
- difference to be detected = 6.7 mmHg
- SDD for MAP = 9 mmHg.

He will require 74 patients, i.e. two groups of 37 patients.

If a power of 0.8 was acceptable (perhaps more realistic) then the sample size needed would be 56 patients. If he was willing to take three replicate measurements at each visit and use the mean, this would reduce the SDD for MAP to approximately 8 mmHg, and the sample size needed would be 60 (power 0.9) or 46 (power 0.8). Use of three measurements reduces the required sample size by about 20%.

Power of published studies

The consistency of the SDDs for BP in our studies suggests that they may be generally applicable and we have used them to examine the approximate power of the published studies tabulated above. For those studies which compared two or more active antihypertensive drugs and which showed no significant difference between treatments the estimated power (probability of detecting a 10/5 mmHg difference at $P < 0.05$) ranged from 0.22–0.52 which is well below the desirable 0.8–0.9. In fact, use of the nomogram for retrospective calculations tends to overestimate power for samples of this size (Altman, 1980).

Discussion

In the studies surveyed the average sample size for trials comparing active drugs with placebo was small (11 patients), but this proved sufficient to separate the treatments significantly in each case. Our calculations confirm that few patients are needed for such a

study, even when the drug is of only moderate potency and a parallel group design is used. The fact that all of the placebo-controlled trials surveyed gave positive results despite using 15 or fewer patients may be explained by the drugs being more potent than thiazides and the tendency for small negative studies against placebo to be suppressed or rejected. On the other hand nearly all studies comparing two or more active drugs or dosage regimens failed to separate the treatments. Our data suggest that the power of these studies was well below the acceptable value of 0.8, and that the average sample size (19 patients) was much too small. The 'no difference' outcome could simply indicate type II error rather than similar efficacy of the treatments in each case. It is disturbing that these were consecutive studies in a journal generally considered to have a policy of strict refereeing. It should also be noted that the results of such studies are commonly used for commercial promotion of drugs as equivalent in efficacy, when the drugs may in fact differ to a clinically important degree.

The other point to emerge from the survey was the absence of a consensus on the optimum number of BP measurements. Our data suggest that use of the mean of three replicate measurements at single visits will reduce the sample size needed by about 20%. Measurements with a random-zero sphygmomanometer are time-consuming, and it may be preferable to settle for a larger sample size and use single

measurements unless patients for a particular study were difficult to find. Single measurements at *replicate visits* would be expected to have a larger effect on the power of a study (Armitage & Rose, 1966; Rosner & Polk, 1979), although we did not have data to examine this. According to Armitage & Rose (1966) a single BP reading on two separate occasions gives a better estimate of mean BP than infinite readings on a single occasion. This has interesting implications for continuous intra-arterial monitoring which has been increasingly used for clinical trials recently. The continuous record represents, in effect, infinite readings on a single occasion, and one would not expect a large effect on the power of comparative studies, although the measurements may be more accurate in the absolute sense.

In our experience the SDDs between BP measurements varied little between studies with different observers, patient groups and measuring devices. We suggest that SDD estimates of 14 mmHg systolic, 9 mmHg diastolic and 9 mmHg MAP could be used to make approximate estimates of sample sizes when those planning studies have no personal data and cannot conduct a pilot study. Those asked to referee articles describing comparative studies of antihypertensive drugs for scientific journals should seek information on the power of the study. When this information is not given our figures can be used to make an approximate estimate of the power.

References

- ALTMAN, D.G. (1980). Statistics and ethics in medical research. III. How large a sample? *Br. med. J.*, **281**, 1336-1338.
- ARMITAGE, P. & ROSE, G.A. (1966). The variability of measurements of casual blood pressure. I. A laboratory study. *Clin. Sci.*, **30**, 325-335.
- BULPITT, C.J. (1975). The design of clinical trials *Br. J. hosp. Med.*, **13**, 611-620.
- HILLS, M. & ARMITAGE, P. (1979). The two-period cross-over clinical trial. *Br. J. clin. Pharmac.*, **8**, 7-20.
- McMAHON, F.G. (1978). *Management of essential hypertension*, p. 23. New York: Futura.
- ROSNER, B. & POLK, B.F. (1979). The implications of blood pressure variability for clinical and screening purposes. *J. chronic Dis.*, **32**, 451-461.
- SILAS, J.H., BARKER, A.T. & RAMSAY, L.E. (1980). Clinical evaluation of Dinamap 845 automated blood pressure recorder. *Br. Heart J.*, **43**, 202-205.

(Received January 11, 1982,
accepted April 5, 1982)