STATISTICS IN MEDICINE, VOL. 8, 467–475 (1989)

COVARIATE IMBALANCE AND RANDOM ALLOCATION IN CLINICAL TRIALS

S. J. SENN

Ciba-Geigy AG, Medical Department, 4002 Basle, Switzerland

SUMMARY

A model is developed to estimate the effect of covariate imbalance on the size of a test of treatment efficacy in randomized clinical trials comparing two treatments when dispersion parameters are known. It is concluded that tests of homogeneity on the covariates should not be performed, that covariate imbalance is just as much a problem for large studies as for small ones in terms of effect on size, and that the effect of correlation between covariates and measures of efficacy is more complex than has previously been sugested. The best way to adjust for covariate imbalance is by an analysis of covariance.

KEY WORDS Size Analysis of covariance Tests of homogeneity Covariate imbalance Clinical trials Randomization

INTRODUCTION

'Block what you can and randomize what you cannot' is eminently sensible advice given by the authors of that excellent textbook, *Statistics for Experimenters*.[1] Nevertheless, a frequent source of anxiety for clinical researchers is the process of randomization, and a commonly expressed worry, despite the care taken in randomization, is that the treatment groups will differ with respect to some important prognostic covariate whose influence it has proved impossible to control by design alone. Under such circumstances the statistician may choose to explain that type I errors are unavoidable but that the probability of their occurrence, given the null hypothesis is correct, may be fixed at any desired level and, provided that the experimenter has chosen the most powerful test available, there is nothing that may be done to improve the position. Of course, this most powerful test may take account of any uncontrolled prognostic factors through an analysis of covariance.

However, it is also common experience that arguments of power do not satisfy clinicians who are unimpressed by justification in terms of averages over all experiments. Perhaps for this reason 'tests of homogeneity' for covariates in placebo and treatment groups of randomised clinical trials are frequently carried out before proceeding to tests of treatment efficacy. This practice has been critically examined and condemned by Altman[2] who points out that covariate imbalance is not sensibly determined by significance tests. This paper will cover essentially the same ground as Altman but using an approach which facilitates definite results. In particular it will enable two interesting conclusions to be drawn: first, that covariate imbalance is of as much concern in large studies as in small ones, and secondly, that for a given observed degree of imbalance the effect on the size of the test of efficacy does not necessarily increase with the correlation between the covariate and the measure of efficacy.

0277–6715/89/040467–09$05.00

Received January 1988
Revised October 1988

MODEL

Altman writes that it is easier 'to study the possible magnitude of the effect of imbalance in the case where both outcome and baseline variable are dichotomous'. I will develop a simple model of placebo controlled trials below to study the effect of covariate imbalance when both are continuous.

Consider a randomized parallel-group study of placebo versus active treatment in which treatment response, Y, and a prognostic factor, X, (which may be either a single covariate or a suitable function of two or more covariates) are recorded. Assume further that X and Y are normally distributed with means μ_x and μ_y, standard deviations σ_x^2 and σ_y^2, and correlation coefficient ρ, in the placebo group and corresponding parameters μ_x, $\mu_y+\Delta$, $(\Delta \geqslant 0)$ σ_x^2, σ_y^2, ρ in the active treatment group.

To investigate sample size I will assume that the nuisance parameters $(\sigma_x^2, \sigma_y^2, \rho)$ are known. If d_y denotes the difference between the observed response means in the two groups (treatment minus placebo) then the distribution of d_y is normal with mean Δ and variance σ_y^2 $(1/n_1+1/n_2)$:

$$d_y \sim N\{\Delta, \sigma_y^2(1/n_1+1/n_2)\},$$

where n_1 and n_2 are the numbers of patients in the placebo and treatment groups, respectively. If d_x is the corresponding difference for the prognostic factor, X, then:

$$d_x \sim N\{0, \sigma_x^2(1/n_1+1/n_2)\}.$$

If we define a statistical test based on d_y with null hypothesis, H_0, that the mean responses are equal in the two treatment groups $(\Delta=0)$, as reject H_0 if

$$d_y/\surd\{\sigma_y^2(1/n_1+1/n_2)\} > Z_\alpha, \tag{1}$$

where Z_α is the standardized Normal deviate corresponding to a one-sided test of size α, then we may define a conditional size for this test, $\alpha(d_x)$, as the probability of a type I error given a particular value of d_x.

It is of course a point for debate whether tests of outcome in clinical trials are one-sided or two-sided. Usually a two-sided five per cent significance level is quoted although in many trials it is clear that the only alternative hypothesis worth consideration is one for which the parameter of interest lies in one particular direction (for example, placebo controlled trials). Such tests may be described as one-sided tests at the $2\frac{1}{2}$ per cent level. This debate will not be taken further. Here, it is sufficient to note that for 'large' degrees of imbalance in an important prognostic factor (the cases of most concern) the conditional size of one of the tails in a two-tailed test is negligible and may be ignored. Therefore, the argument of this paper will be developed in terms of a one-tailed test. This will be sufficient to show that tests of homogeneity for covariates are pointless.

CONDITIONAL SIZE

If $\Delta=0$, then conditional upon X

$$Y \sim N\{\mu_y+\rho(X-\mu_x)\sigma_y/\sigma_x, (1-\rho^2)\sigma_y^2\},$$

and conditional upon d_x,

$$d_y \sim N\{\rho d_x\sigma_y/\sigma_x, (1-\rho^2)\sigma_y^2(1/n_1+1/n_2)\}.$$

Hence, it may be shown that the conditional size of the test defined by (1) above is given by:

$$\alpha(d_x)=P[Z>Z_\alpha/\sqrt{(1-\rho^2)}-d_x\rho/\sqrt{\{(1-\rho^2)\sigma_x^2(1/n_1+1/n_2)\}}], \tag{2}$$

where Z is a standard Normal variate and Z_α is the critical value of the test.

Since $\sigma_x\sqrt{(1/n_1+1/n_2)}$ is the standard error of d_x we may consider (2) in terms of a standardized prognostic difference, d_x^*, between the placebo and active treatment groups and write:

$$\alpha(d_x)=P[Z>Z_\alpha/\sqrt{(1-\rho^2)}-d_x^*\rho/\sqrt{(1-\rho^2)}], \tag{3}$$

from which the following points may be noted:

1. If $\rho=0$ then $\alpha(d_x)=\alpha$ as expected since if prognostic factor and response are uncorrelated the conditional size of the test should not vary.
2. The expression does not depend on sample size. It follows that covariate imbalance, contrary to what has been claimed by Altman, is just as much of a problem for large studies as for small ones. This point has not in general been well understood. For example, Pocock[3] claims that 'In randomized clinical trials one can generally expect treatment groups to be fairly well matched but occasionally one will be unlucky and discover some factor which differs substantially between treatments. This is more likely to occur if the trial is small.' Rothman[4] makes a similar point when he states that 'the larger the size of the treatment groups, the smaller is the amount of confounding which can result.'. However, it is not absolute imbalance which is important but standardized imbalance and this is independent of sample size. With benefit of hindsight this is obvious. Since the classical approach controls the probability of a type I error, given that the null hypothesis is true, and makes this constant irrespective of the number of observations, it is not surprising that conditional size is also independent of the number of observations.
3. The formula confirms Altman's results[2] that, if conditional size is considered important, a significance test for homogeneity is no safeguard against imbalance, for reasons which will become clear below.

THE EFFECT OF CORRELATION

Figure 1 is a plot of conditional size against standardized covariate difference for several values of ρ and an unconditional size, α, of 0·05. It can be seen that the conditional size exceeds the nominal value of the test well before the point at which a test of homogeneity for the covariate would call the balance of the experiment into question. To make this explicit consider the value of the covariate difference which, for a given test size, α, and correlation coefficient, ρ, will cause the conditional size, $\alpha(d_x)$, to exceed the nominal value of the test. If we denote this value of the standardized covariate difference by d', then we have $\alpha(d')=\alpha$, and it follows from (3) that:

$$d'=Z_\alpha\{1-\sqrt{(1-\rho^2)}\}/\rho. \tag{4}$$

Figure 2 shows a plot of d' against ρ for a value of $Z_\alpha=1{\cdot}645$ corresponding to $\alpha=0{\cdot}05$.

From Figure 1 there appears a boundary beyond which conditional size cannot increase, whatever the value of ρ, given a particular value of d_x^*. In fact for each value of d_x^* there is a worst possible value of ρ. Maximizing (3) with respect to ρ this value is given by:

$$\rho=\begin{cases} 1 & \text{if } d_x^*\geqslant Z_\alpha \\ -1 & \text{if } d_x^*\leqslant -Z_\alpha \\ d_x^*/Z_\alpha & \text{if } -Z_\alpha<d_x^*<Z_\alpha \end{cases} \tag{5}$$

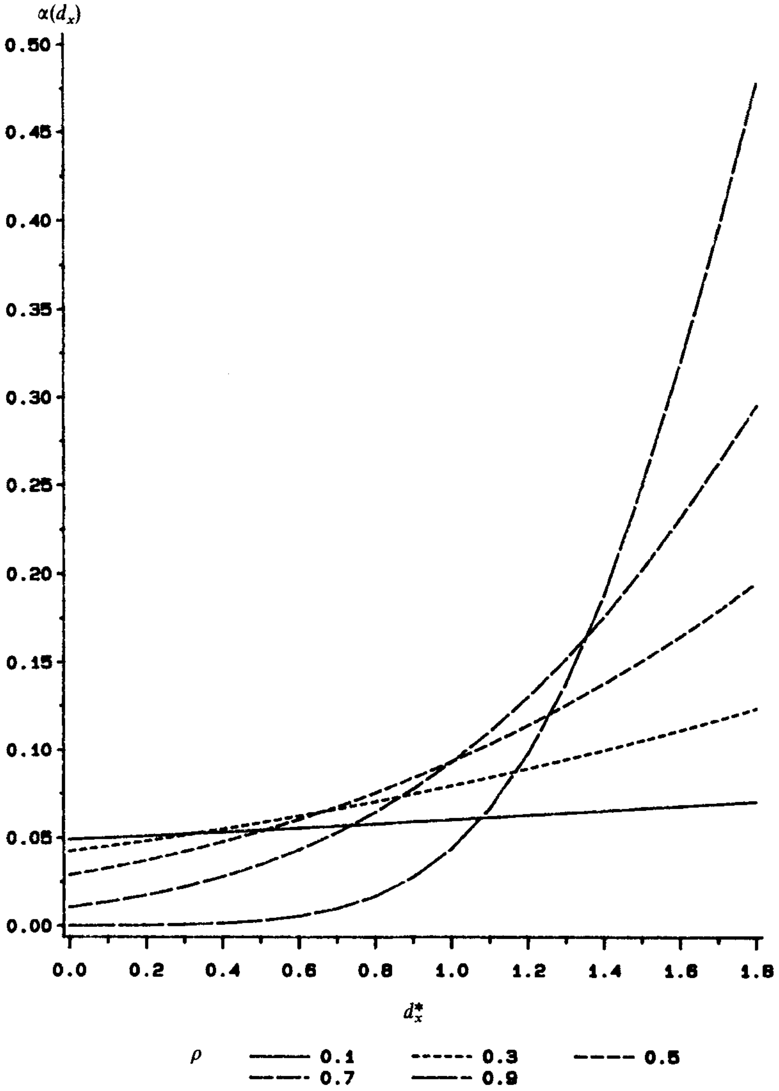

Figure 1. Conditional size $\alpha(d_x)$ as a function of standardized covariate difference (d_x^*) and correlation (ρ) for a test of nominal size $\alpha = 0{\cdot}05$

Figure 3 shows a plot of maximum conditional size and corresponding value of ρ against the standardized difference, d_x^* for $\alpha = 0{\cdot}05$. From this figure it is possible to confirm the general point made by Altman, namely that significance tests on covariates are not guaranteed to control size. Indeed, if a two-sided test of covariate homogeneity, of the same size as a one-sided test on the measure of efficacy is used, the maximum possible size is 100 per cent.

It is worth justifying the very strange plot in Figure 3 heuristically. The effect on conditional size of correlation between covariate and response is first to shift the conditional mean of the response and secondly to affect its conditional variance. For low values of d_x^* the slight shift in the conditional mean of d_y^* away from zero induced by high correlation is negligible compared with the associated reduction in variance. However, as d_x^* approaches the critical value of the test of

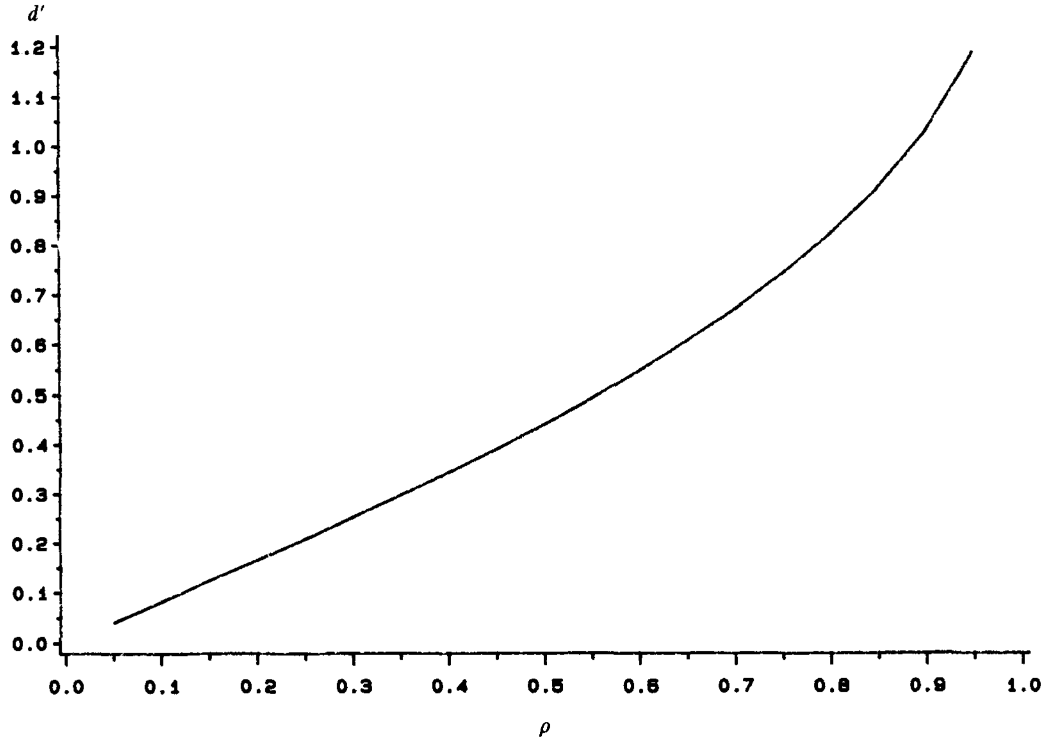

Figure 2. Standardized covariate difference at which conditional size $>0{\cdot}05$ (d') against correlation (ρ)

efficacy, the size is affected more and more adversely by large values of ρ. With d_x^* just less than Z_α a value of ρ just less than one will produce a size of 0·5, this being the worst possible case. However, as soon as d_x^* exceeds Z_α, clearly, a correlation of 1 produces a conditional size of 1.

From Figure 3 it follows that Altman's assertion, that it is the strength of association between prognostic and response variables which is important, is incorrect if this is taken to mean that the stronger the association the worse the position with respect to size. Unconditionally all (correct) tests have the same size irrespective of correlation between covariate and measure of response. Conditional upon a given covariate imbalance there is a worst possible correlation from the point of view of size and this is not in general equal to 1.

RECOMMENDATIONS

Just as for every degree of covariate imbalance for each nominal (unconditional) test size there is a worst (largest) possible conditional size, so for a given degree of covariate imbalance there is an unconditional test size which ensures that the conditional size cannot exceed a given desired level.

From (3) and (5), a formula may be derived for the unconditional size, α_u, of a test which will ensure that, for a given value of the standardized covariate difference, d_x^*, the conditional size, α_c, is less than or equal to some desired level, α^*. This formula is:

$$\alpha_u = P(Z > Z_u),$$

where Z is a standardized Normal variate and

$$Z_u = \sqrt{(Z_c^2 + d_x^{*2})}.$$

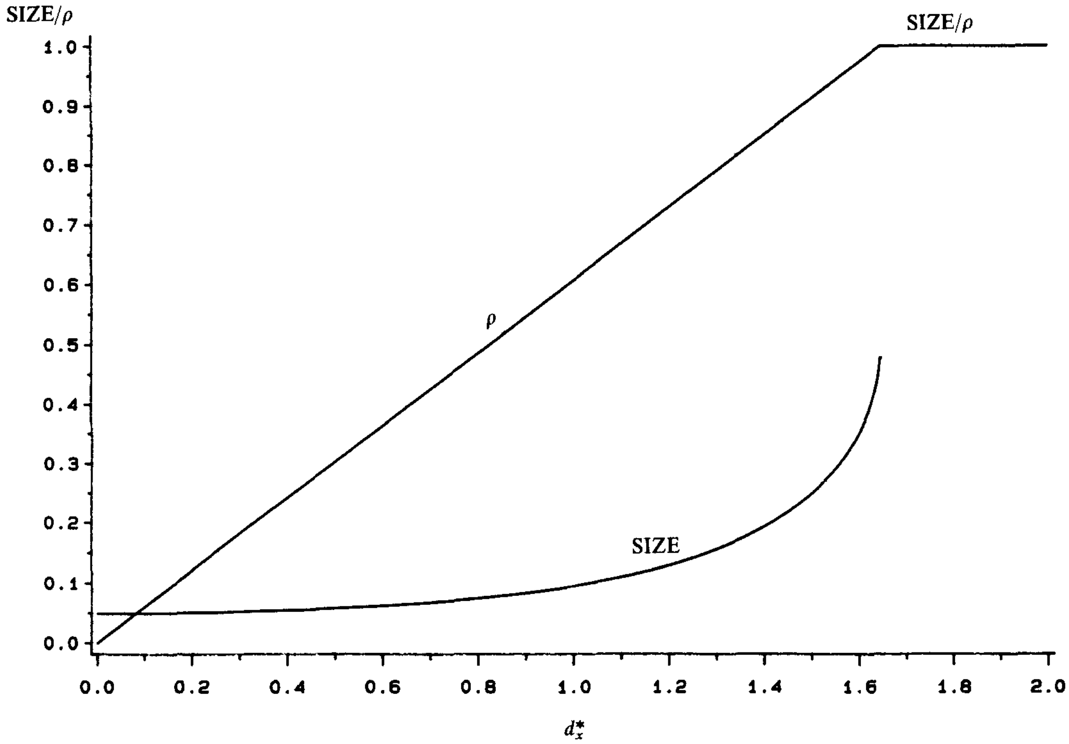

Figure 3. Worst possible values of conditional size and corresponding correlation against the standardized covariate difference (d_x^*)

Z_c is a critical value such that

$$P(Z > Z_c) = \alpha_c = \alpha^*. \tag{6}$$

Figure 4 is a plot of the unconditional size required for a given standardized difference to ensure that the conditional size should be no greater than 0·05.

Therefore, to allow for covariate imbalance, the following procedure might be useful:

1. Establish the standardised covariate difference between groups.
2. Use this in connection with formula (6) to choose a nominal value for the unconditional size of the test.
3. Carry out the statistical test of efficacy using this selected size.

However there are a number of disadvantages to this conservative approach:

1. Cost in terms of reduced power.
2. The procedure is unclear when there are many covariates though if covariate imbalance is measured in terms of standardized distance of the linear discriminant function this will be conservatively adequate.
3. There is a superior alternative in the analysis of covariance, which may be used to correct for covariate imbalance.

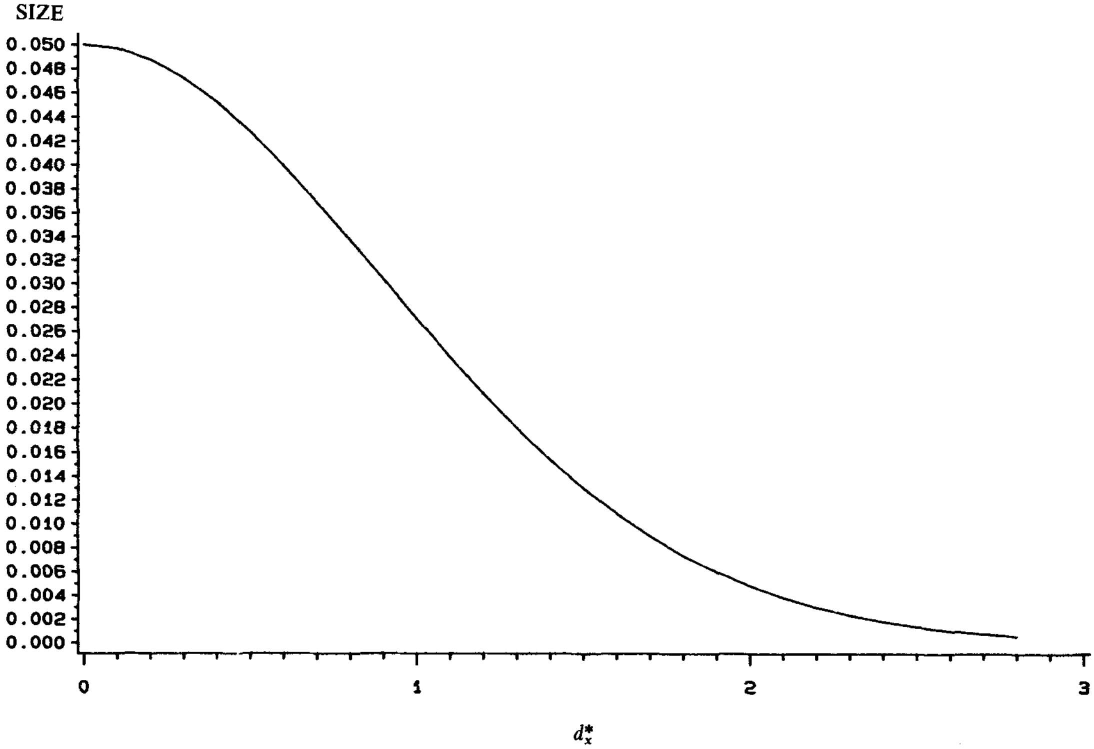

Figure 4. Unconditional size required to produce conditional size not greater than 0·05 against standardized covariate difference (d_x^*)

In the context of our model, the statistic

$$d_y - \rho d_x \sigma_y / \sigma_x$$

may be used to examine the difference between groups with an increase in the precision of the estimate of Δ, and a consequent increase in power, as well as the advantage of a test which has constant conditional size. The first of these advantages is well known and has often been remarked upon; the second is less often referred to. For example, Cox and McCullagh[5] in their important review of analysis of covariance give four broad areas of application. They include adjustment for precision in designed experiments and adjustment for bias in observational experiments but they do not mention adjustment for imbalance in designed experiments. Armitage,[6] in a review of the analysis of clinical trials, gives as one of three possible advantages that taking prognostic variables into account, 'may correct for the effect of disparities in these variables which may have occurred in spite of randomization'; he does not include any specific discussion of conditional size.

Of course, randomized experiments are balanced over all possible allocations, but, as was discussed irr the introduction, this 'consolation of the marathon experimenter' has few charms for the physician who, on finally decoding the double blind allocation, discovers covariate imbalance. Analysis of covariance can be recommended on two grounds: increased power and constant conditional size. The former should be sufficient to recommend the method to those who consider that the latter is irrelevant but for those who are concerned about conditional size this is an added advantage. The disadvantages are, first, that in practice dispersion parameters have to be estimated and the corresponding loss of degrees of freedom may be a problem in small studies, and, secondly, that the technique may be difficult to apply in experiments with awkward variables.

Cox and McCullough's[5] discussion of the first problem leads them to conclude that covariate adjustment is of little value if ρ is less than 0·3 and is impractical for more than one or two concomitant variables if the residual degrees of freedom ($n_1 + n_2 - 2$ in the example considered in this paper) are small. However, the position is more complex with regard to size. Again we come back to the fact that unconditionally all correct tests at a given level have the same size. Conditionally, the maximum possible size is 1 for any value of ρ not equal to 0, although, if ρ is small, large values of conditional size are less likely. Therefore, it is not possible to eliminate the effect of imbalance, the possibility of which would appear part of the process of randomization. In practice, some compromise may be necessary between the requirements of precision and the control of size.

The problems engendered by the response variable having a non-linear relation to the covariate can be overcome sometimes by suitable transformation, by use of dummy varibles (as might be appropriate for categorical data) or by using additional higher order powers (squared, cubic) of the concomitant variable. These last two options can be expensive in terms of degrees of freedom.

For the analysis of clinical trials the following procedure is recommended:

1. Before the data are collected, relevant prognostic variables should be identified using *a priori* information on correlation with treatment response and taking into account requirements regarding conditional size and precision (consideration of Figure 1 may be of help here);
2. Other covariates collected 'for the record' should be ignored in the analysis;
3. Do not carry out tests of homogeneity on the covariates;
4. Perform an analysis of covariance using the identified prognostic factors (step 1 above).

CONCLUDING REMARKS

The argument in this paper has been developed in terms of a bivariate normal response with known dispersion parameters; it is worth considering whether the conclusions regarding tests of homogeneity are still valid more generally.

It was stated, for example, that balance does not improve with sample size. This is obviously true for a standardized normal measure of balance but is it for a 't' statistic, where the degrees of freedom will change with sample size? All that is necessary to make it so is to describe the degree of imbalance in terms of the 'P value'. Under randomization this has a uniform distribution between 0 and 1, whatever the size of the sample. Of course the effect of a given degree of imbalance on conditional size may vary with degrees of freedom when the dispersion parameters are unknown. In the limit, the relationship is that provided by the normal distribution, and since, as the argument of this paper has shown, tests of homogeneity are useless under such circumstances, there are no logical grounds for using them when the sample size is small.

The argument has been stated in terms of a possible adverse effect on size (an increase) when the measure of balance (assumed positively correlated with outcome) was positive. However, it is also possible with a one-sided test of outcome, for a given degree of 'negative' imbalance to produce excessively small conditional size and adversely affect the power of the test. Again, analysis of covariance provides a solution, but the decision to use it must be taken before examination of the data. With two-sided tests of efficacy, small conditional size is still possible for low degrees of imbalance if the correlation with outcome is high. Whatever the circumstances, there can be no justification for tests of homogeneity whether formal or informal. Their use should be eliminated from the practise of clinical trials.

ACKNOWLEDGEMENT

I should like to thank the referees for helpful comments.

REFERENCES

1. Box, G. E. P., Hunter, W. G. and Hunter, J. S. *Statistics for Experimenters*, Wiley, New York, 1978.
2. Altman, D. G. 'Comparability of randomised groups', *The Statistician*, **34**, 125–136 (1985).
3. Pocock, S. J. *Clinical Trials, A Practical Approach*, Wiley, Chichester, 1983.
4. Rothman, K. J. 'Epidemiologic methods in clinical trials', *Cancer*, **39**, 1771–1775 (1977).
5. Cox, D. R. and McCullough, P. 'Some aspects of covariance', *Biometrics*, **38**, 541–561 (1982).
6. Armitage, P. 'The analysis of data from clinical trials', *The Statistician*, **28**, 171–183 (1979).