STATISTICS IN MEDICINE, VOL. 3, 199–214 (1984)

APPROACHES TO SAMPLE SIZE ESTIMATION IN THE DESIGN OF CLINICAL TRIALS—A REVIEW

ALLAN DONNER

Department of Epidemiology and Biostatistics, University of Western Ontario, London, Canada N6A 5B7

SUMMARY

Over the last decade, considerable interest has focused on sample size estimation in the design of clinical trials. The resulting literature is scattered over many textbooks and journals. This paper presents these methods in a single review and comments on their application in practice.

KEY WORDS Sample size determination Comparison of rates Survival analysis

INTRODUCTION

Increased attention has focused recently on the importance of sample size considerations in the design of randomized controlled trials (RCTs). Freiman *et al.*[1] reviewed the power of 71 published RCTs which had failed to detect a significant difference between groups and found that 67 of these trials could have reasonably missed a 25 per cent therapeutic improvement, whereas 50 could have missed a 50 per cent improvement. The authors concluded that many of these studies were not only negative, but, because of insufficient numbers, might also mislead. However, it does not follow from this that an investigator should enroll as many patients as possible in a clinical trial. If the number of patients exceeds that required, the trial will be unnecessarily expensive and prolonged. An investigator must strike a balance between enrolling sufficient patients to detect important differences, but not so many patients such that he would unnecessarily waste important resources.

In this paper, we review the recent literature concerning sample size estimation in the design of RCTs. The review restricts attention to designs with the primary purpose of comparing two groups of patients with respect to the occurrence of some specified event, such as death or the recurrence of disease; we also discuss trials where interest centres on time to the terminal event, rather than the occurrence of the event itself. Schwartz, Flamant and Lellouch[2] give an excellent account of sample size estimation for continuous outcome variables, and Lachin[3] discusses the case of multi-group comparisons with respect to categorical outcome variables.

Formal sample size planning in the design of a clinical trial usually depends on relatively simple and well-known formulae presented in introductory statistics texts. Fleiss[4] points out that these formulae, although useful for short-term studies, may not prove adequate in the design of RCTs, where, over long periods of time, issues peculiar to human experimentation inevitably arise. Although published methods to deal with these issues exist, their literature is scattered over many textbooks and journals. We aim, therefore, to present these methods in a single review paper, and to comment on their practical utility, advantages and disadvantages.

Lachin[5] also presents a review of sample size evaluation for clinical trials. However, he deals

0277–6715/84/030199–16$01.60

Received 3 May 1983
Revised 23 November 1983

broadly with a variety of statistical procedures and response variables, whereas this review deals in depth with studies comparing two treatments with respect to the occurrence of (or time to) some specified event.

NOTATION

We assume comparison of an experimental treatment, E, with a control treatment, C. Standard formulae for sample size depend on the chosen probabilities α and β associated with a type I error (falsely declaring a treatment difference) and a type II error (falsely declaring no treatment difference), respectively. We express this dependence mathematically by the quantities Z_α and Z_β, defined as the values of standardized normal deviates corresponding to α and β, where $1-\beta$ is the trial power. Table I provides values of Z_α and Z_β corresponding to commonly used levels of significance and power, separately for one- and two-sided tests. Most authors (e.g. Friedman, Furberg and DeMets[6]) recommend the exclusive use of two-sided tests unless one has strong justification for expecting a difference in only one direction.

The number of patients required for a trial comparing two T-year event rates is a function of Z_α, Z_β, a measure of the anticipated effect of intervention, and estimates of various other parameters whose impact is under consideration. One usually measures the anticipated effect of intervention with either the difference or the ratio of the expected event rates. Other parameters whose impact one may wish to consider include factors such as pre-stratification in the study design, the effect of randomizing groups of patients rather than individuals, and the effect of patient dropout or non-adherence.

Table I. Values of Z_α and Z_β corresponding to specified values of significance level and power

		Two-sided tests	One-sided Tests
Level	0·01	2·576	2·326
	0·05	1·960	1·645
	0·10	1·645	1·282
Power	0·80	0·840	
	0·90	1·282	
	0·95	1·645	
	0·99	2·326	

BASIC APPROACHES TO SAMPLE SIZE ESTIMATION

All formulae for sample size estimation correspond to a specified null hypothesis (H_0) and one or more test statistics. For each approach, therefore, we will present H_0, a reference to the appropriate analytic methods, one or more remarks concerning practical application, and, finally, an example. Unless otherwise specified, the alternative hypothesis H_1 may be either one- or two-sided. For each approach discussed, we present the sample size requirements in terms of the number of patients n to be randomized to each of the two groups.

1. Sample size requirements in terms of risk difference

This approach, by far the most frequently used in the design of clinical trials, is a straightforward application of traditional sample size formulae for comparing two proportions. I et

P_C = anticipated T-year event rate among control group patients
P_E = anticipated T-year event rate among experimental group patients
$\delta = P_E - P_C$ = the difference in event rates regarded as scientifically or clinically important to detect.

Formula 1. (H_0: $P_E = P_C$)

$$n = \{Z_\alpha\sqrt{[2\overline{P}(1-\overline{P})]} + Z_\beta\sqrt{[P_E(1-P_E)+P_C(1-P_C)]}\}^2/\delta^2$$

where $\overline{P} = (P_E + P_C)/2$ (Reference 4 pp. 38–42)

Test statistic. Chi-square contingency test (Reference 4, pp. 21–27)

Remarks

(1.1) In practice, one usually has an estimate of P_C available from past experience. One may then estimate P_E by $P_C + \delta$.

(1.2) Several other sample size formulae in terms of the risk difference address the problem of comparing two proportions (Reference 7, p. 318; Reference 8, p. 180; Reference 9, p. 129). All of these formulae, including formula 1, give sample size estimates that are close to the exact values required to produce the desired power with use of the chi-square test without correction for continuity.[10] Fleiss, Tytun and Ury[11] have shown that the incorporation of the continuity correction implies, to a high degree of accuracy, that the value of n in formula 1 should be increased by an amount $2/|P_E - P_C|$.

The basic similarity among the various sample size formulae expressed in terms of the risk difference results from the fact that each is a variation of the basic formula, $n = (\sigma_0 Z_\alpha + \sigma_1 Z_\beta)^2/\delta^2$, where σ_0 and σ_1 are the standard deviations of an observation under H_0 and H_1, respectively. Different methods of estimating σ_0 and σ_1 lead to the different formulae.

(1.3) Formula 1 is well-approximated[5] by the even simpler formula $n = (Z_\alpha + Z_\beta)^2[2\overline{P}(1-\overline{P})]/\delta^2$. For example at $\alpha = 0.05$, (one sided), $\beta = 0.10$, this formula will yield total sample sizes within six units of the total sample size yielded by formula 1.

(1.4) Fleiss, Tytun and Ury[11] have extended formula 1 to the case of randomization of unequal numbers of patients to the two groups. Suppose we wish to randomize n patients to the experimental group and sn to the control group, where $0 < s < \infty$. Then the required sample size is

$$n = \{Z_\alpha\sqrt{[(s+1)\overline{P}_s(1-\overline{P}_s)]} + Z_\beta\sqrt{[sP_E(1-P_E)+P_C(1-P_C)]}\}^2/(s\delta^2),$$

where $\overline{P}_s = (P_E + sP_C)/(s+1)$.

(1.5) Clinical trials in which patients are individually matched occur only infrequently. However, use of formula 1 for matched designs will tend to give conservative (somewhat too large) sample size requirements. We note that many authors (e.g. Schlesselman[12]) recommend that one ignore matching in the determination of sample size, since, prior to the study, the assessment of its impact will be difficult to estimate. For the same reason, the reduction in n which may be obtained by taking into account factors which produce variability, such as age and sex, is also usually ignored in sample size planning.

Example 1

Investigators anticipate a 3-year death rate among control patients of about 60 per cent. They view a reduction of this mortality rate to 40 per cent among experimentally treated patients as clinically important. They would like to detect such a reduction with 80 per cent power and with a two-tailed test at the 5 per cent significance level. Thus $P_C = 0{\cdot}60$, $P_E = 0{\cdot}40$, $\overline{P} = (0{\cdot}60 + 0{\cdot}40)/2 = 0{\cdot}50$, $Z_\alpha = 1{\cdot}96$, $Z_\beta = 0{\cdot}84$. Substitution into formula 1 yields $n = 96{\cdot}8$ or 97 as the number of patients required in each group. With employment of a correction for continuity, remark 1.2 implies an increase in patient intake to $n + 2/|P_E - P_C| = 97 + 2/0{\cdot}2 = 107$ per group.

Suppose that the investigators prefer to study only half as many experimental as control group subjects. Then, (ignoring the continuity correction) the required number of patients in the experimental group is

$$n = \frac{\{1{\cdot}96\sqrt{[(2+1)(0{\cdot}53)(0{\cdot}47)]} + 0{\cdot}84\sqrt{[2(0{\cdot}40)(1-0{\cdot}40)+(0{\cdot}60)(1-0.60)]}\}^2}{2(0{\cdot}40-0{\cdot}60)^2} = 72{\cdot}4$$

Thus the number of patients required in the control group is $2(73) = 146$, and the total size of the trial should be $73 + 146 = 219$ patients. Note that this is greater than the total number of patients (194) needed to achieve the same precision with assignment of equal number of patients to the two groups. This illustrates the general result that, for a given total number of patients, the most information is obtained under equal patient allocation.

2. Sample size requirements in terms of relative risk

Let $R = P_E/P_C$ = relative risk regarded as clinically or scientifically important to detect.

Formula 2. (H_0: $R = 1$)

$$n = \{Z_\alpha\sqrt{[2\overline{P}_R(1-\overline{P})]} + Z_\beta\sqrt{[P_C\{1+R-P_C(1+R^2)\}]}\}^2/[P_C(1-R)]^2$$

where $\overline{P}_R = \frac{1}{2}P_C(1+R)$ (Reference 12)

Test statistic. Chi-square contingency test (Reference 4, pp. 21–27)

Remarks

(2.1) Formula 2 is algebraically equivalent to formula 1. It pertains to the situation where one can more easily specify the anticipated effect of intervention in terms of R and P_C rather than in terms of P_E and P_C.

(2.2) One may adjust formula 2 for use of a continuity correction by adding $2/[P_C|1-R|]$ to the calculated value of n.

Example 2

Suppose in example 1 the investigators had specified detection of a relative risk of 2/3 among experimental group subjects. Then, substitution of $P_C = 0{\cdot}60$, $R = 2/3$, $Z_\alpha = 1{\cdot}96$ and $Z_\beta = 0{\cdot}84$ into formula 2 yields $n = 97$ patients required in each group. Note that this is identical to the solution obtained with formula 1.

3. Sample size requirements for clinical trials designed to show equivalence

An increasing number of clinical trials seek to show that an experimental therapy is equivalent in efficacy to a control therapy, rather than (necessarily) superior. This often occurs when the experimental therapy is 'conservative' and the standard control therapy is invasive or toxic. In this case the null hypothesis may specify that the success rate P_C on the control therapy is higher than the success rate P_E on the experimental therapy by at least some amount θ. The alternative hypothesis specifies that $P_C - P_E < \theta$, which implies that the two therapies are equivalent. Formula 3 provides the required number of patients for such a trial.

Formula 3. (H_0: $P_C \geqslant P_E + \theta$ vs. H_1: $P_C < P_E + \theta$)

$$n = \{Z_\alpha\sqrt{[2\overline{P}(1-\overline{P})]} + Z_\beta\sqrt{[P_E(1-P_E)+P_C(1-P_C)]}\}^2/(P_E - P_C - \theta)^2$$

where $P_C < P_E + \theta$ and $\theta > 0$.

Test statistic. Critical ratio test for difference between two proportions (Reference 13, pp. 296–298).

Remarks

(3.1) Makuch and Simon[14] and Blackwelder[15] argue that in the planning of equivalence trials formula 3 is more appropriate than formula 1, which reflects the standard significance-testing approach. This is because, while equivalence trials seek, under the standard approach, to accept the null hypothesis, the reporting of the test does not explicitly account for the associated type II error. Thus a 'non-significant result' may invite false confidence in the equivalence of the two treatments. With the above approach, on the other hand, the equivalence trial seeks to reject H_0 with an accompanying error rate of at most α.

(3.2) The most convenient method of using formula 3 in practice is to set $P_E = P_C$, regarding θ as the difference in treatment efficacy that the investigator wishes to rule out with probability $(1-\beta)$. In this case we may also regard formula 3 as yielding the number of patients required to ensure with probability $(1-\beta)$ that the upper $100(1-\alpha)$ per cent confidence limit for the true difference does not exceed θ when $P_E = P_C$. Makuch and Simon[14] recommend $\alpha = 0{\cdot}10$, $\beta = 0{\cdot}20$ and $\theta = 0{\cdot}10$ as useful values in practice for this approach.

(3.3) Suppose that an investigator wishes to randomize n patients to the experimental group and sn to the control group, where $s > 0$. Formula 3 generalizes to give

$$n = \frac{\{Z_\alpha\sqrt{[(s+1)\overline{P}_s(1-\overline{P}_s)]} + Z_\beta\sqrt{[sP_E(1-P_E)+P_C(1-P_C)]}\}^2}{s(P_E-P_C-\theta)^2},$$

where $\overline{P}_s = (P_C + sP_E)/(s+1)$.

Example 3

The control regimen for a planned trial of treatment for hypertension consists of standard drug treatment, whereas the experimental intervention consists of a 'lifestyle modification package', including relaxation therapy and diet. The investigators anticipate that the proportion of patients whose blood pressure is under control is about 80 per cent in each group. However, because of the potential for harmful pharmacological side-effects, they regard the treatments as equivalent if the proportion of experimental patients under control is no more than ten percentage points less than the corresponding percentage for control patients. Thus the required number of patients in the trial at $\alpha = 0{\cdot}10$ (one-sided) and $\beta = 0{\cdot}20$ is, from formula 3,

$$\begin{aligned} n &= \{1{\cdot}282\sqrt{[2(0{\cdot}80)(0{\cdot}20)]} + 0{\cdot}84\sqrt{[(0{\cdot}80)(0{\cdot}20) + (0{\cdot}80)(0{\cdot}20)]}\}^2/(0{\cdot}10)^2 \\ &= 145 \end{aligned}$$

If the anticipated proportion of patients under control in the experimental group is only 75 per cent, then the required number of patients is

$$\begin{aligned} n &= \{1{\cdot}282\sqrt{[2(0{\cdot}78)(0{\cdot}22)]} + 0{\cdot}84\sqrt{[(0{\cdot}80)(0{\cdot}20) + (0{\cdot}75)(0{\cdot}25)]}\}^2/(0{\cdot}05)^2 \\ &= 624 \end{aligned}$$

The required number of patients is larger in this second case because one would like to rule out a difference of 0·10 when the true difference is 0·05. It is intuitively clear that one requires fewer patients to rule out a given value of θ when the true difference is zero.

Note also that rejection of H_0 under this approach (and the consequent conclusion of equivalence) will be associated with a 10 per cent error rate, explicitly stated in terms of α.

4. Sample size requirements that account for stratification of subjects

This approach assumes stratification of subjects into K risk categories (e.g. based on age) with n_j subjects randomly assigned to each of an experimental and control group within the jth stratum, $j = 1, 2, \ldots, K$. One wishes to compare event rates within each of the resulting 2 × 2 tables, and to obtain an overall comparison to test whether the (assumed) common relative odds equals unity. Let P_{C_j} and P_{E_j}, $j = 1, 2, \ldots, K$, denote the event rate among control and experimental group patients in the jth stratum. Then the common relative odds is $\text{OR} = P_{E_j}(1-P_{C_j})/P_{C_j}(1-P_{E_j})$, for all j. Define

$$\Delta = \log_e(\text{OR})$$

$$g_j = \frac{\Delta^2}{\dfrac{1}{P_{C_j}(1-P_{C_j})} + \dfrac{1}{P_{E_j}(1-P_{E_j})}}, \quad j = 1, 2, \ldots, K$$

f_j = fraction of observations contained in the jth table, $j = 1, 2, \ldots, K$.

Formula 4. (H_0: OR = 1).

$$n = \frac{(Z_\alpha + Z_\beta)^2}{\sum^K g_j f_j}, \text{ where } n = \sum^K n_j = \sum^K f_j n \text{ (Reference 16)}$$

Test statistic. Unconditional large sample test (Reference 17, pp. 76–87). Mantel–Haenszel chi-square test (Reference 4, pp. 173–178).

Remarks

(4.1) We can write the term P_{E_j} as $P_{E_j} = P_{C_j}/[P_{C_j} + (1-P_{C_j})\exp(-\Delta)]$. Thus we can regard the term g_j as a function of Δ and P_{C_j} alone. Gail[16] provides a table of g_j in terms of OR and P_{C_j}.

(4.2) Formula 4 assumes that the relative odds OR (rather than the relative risk) is constant over strata, since this is an empirically reasonable and frequently adopted model for the analysis of independent 2 × 2 tables. Although in a single 2 × 2 table a test of H_0: OR = 1 is equivalent to a test of H_0: $R = 1$ the two parameters will in general have quite different values unless the underlying event rate is very small (Reference 4, p. 81).

(4.3) Formula 4 depends on the assumption that the marginal totals of the 2 × 2 tables are random variables. Thus it most strictly corresponds to the unconditional analysis proposed by Cox (Reference 17, pp. 76–87) rather than the conditional analysis implied by the Mantel–Haenszel test (which assumes the marginal totals are fixed). However, Gail[16] suggests that formula 4 yields sample sizes appropriate to conditional testing provided all n_j exceed 15 and all P_{C_j} and P_{E_j} lie in the interval (0·1, 0·9).

(4.4) The multiple-table approach to sample size planning is feasible only with information available from previous studies to provide estimates of the P_{C_j} and f_j. This might be the case, for example, in planning drug trials with previous incidence data available by stratum. The use of this information, when accurate, leads to more precise estimates of sample size, especially if the P_{C_j} vary widely. However, reliable prior information on stratum-specific rates and relative stratum sizes often remains difficult to obtain. In this case one should use formulae 1 or 2 with P_C or P_E representing 'average' incidence rates.

Example 4

Suppose there are two strata of increasing risk such that the anticipated control group rates are $P_{C1} = 0{\cdot}30$ and $P_{C2} = 0{\cdot}60$, with previous studies suggesting $f_1 = 2/3$ and $f_2 = 1 - 2/3 = 1/3$. It is of interest to detect OR $= 2{\cdot}5$ as statistically significant, with $\alpha = 0{\cdot}05$ (1-tailed) and $\beta = 0{\cdot}10$. Thus $\Delta = 0{\cdot}92$, $g_1 = 0{\cdot}0958$, $g_2 = 0{\cdot}0824$, and

$$n = \frac{(1{\cdot}645 + 1{\cdot}282)^2}{2/3(0{\cdot}0958) + 1/3(0{\cdot}0824)} = 93{\cdot}79.$$

Thus one should enter 94 patients into the trial (per group), 63 in stratum 1 and 31 in stratum 2.

Disregarding stratification in this example, one would apply formula 1 with $P_C = 0{\cdot}30(2/3) + 0{\cdot}60(1/3) = 0{\cdot}40$, for which $g = 0{\cdot}0996$ and $n = (1{\cdot}645 + 1{\cdot}282)^2/0{\cdot}0996 = 87$, an underestimate of 7 patients per group.

5. Sample size requirements in terms of time to some critical event

This approach, originally due to Pasternack and Gilbert,[18] assumes that greater interest attaches to the time to some critical event, such as death or the recurrence of disease, rather than the occurrence or non-occurrence of the event. Thus the approach pertains particularly to studies that aim to compare length of survival on different treatments.

We assume, as in Reference 19, that the time-to-event (survival time) has an exponential distribution with means μ_C and μ_E in the control and experimental groups, respectively. This is equivalent to the assumption that the 'hazard function' or instantaneous probability of death (recurrence of disease, etc.) is constant within each group. We also assume that patients enter the trial according to a Poisson process.

Let $\theta = \mu_E/\mu_C$ = ratio of mean survival times regarded as important to detect. If all patients are followed-up in each group until the occurrence of the critical event, i.e. there are no censored observations, then formula 5 gives the required number of patients.

Formula 5. (H_0: $\theta = 1$)

$$n = 2(Z_\alpha + Z_\beta)^2/[\log_e(\theta)]^2 \text{ (Reference 19)}$$

Test statistic. Cox's F-test (Reference 20, pp. 133–135)

Remarks

(5.1) Formula 5 derives from an approximation to the exact sample size solution. However George and Desu[19] showed that the formula is accurate to within two sample units. A generalization of formula 5 to comparative trials involving more than two treatment groups is given by Makuch and Simon.[21]

(5.2) The exponential distribution is the simplest and most widely used distribution for describing survival data: Gross and Clark[22] discuss it in detail. Under the assumption of exponential time-to-event, the median survival time is a constant multiple of the mean survival time. Thus we may assess the parameter θ in formula 5 as either a ratio of mean times-to-event of interest or of median times-to-event. Freedman[23] has generalized formula 5 to the situation with no distributional assumptions made concerning survival times.

(5.3) The F-test developed by Cox[20] provides the most powerful test of H_0: $\theta = 1$ when the times-to-event are exponentially distributed, and is the procedure which analytically corresponds to formula 5. However one may also use non-parametric procedures for testing H_0, such as the Wilcoxon rank-sum test.[24] That is, we do not require the assumption of exponential survival for analysis.

(5.4) There is a close relationship between this approach and the T-year event rate approach reflected by formula 1. Under the assumption of exponential survival time, the T-year event rates among control and experimental group patients become, respectively, $P_C = 1 - \exp$

$(-T/\mu_C)$ and $P_E = 1-\exp(-T/\mu_E)$. Thus if the event in question is death, for example, the proportion of survivors at any time T is $\exp(-T/\mu_i)$, $i =$ C, E. It follows that $\theta = \mu_E/\mu_C = \log_e(1-P_C)/\log_e(1-P_E)$.

(5.5) The approach described above assumes follow-up of all patients in each group until failure, i.e. until the occurrence of the event in question. This will not frequently occur in practice, both because of loss to follow-up and because of an excessive required duration of the trial. If we assume instead that patients enter the trial at a uniform rate over a T-year period, Gross and Clark[22] developed a further approximation, which depends on the separate median survival times μ_E and μ_C. If the trial terminates at a time T, the required number of patients in each group is

$$n = (Z_\alpha + Z_\beta)^2[\phi(\mu_E) + \phi(\mu_C)]/(\mu_E^{-1} - \mu_C^{-1})^2$$

where

$$\phi(\mu_i) = \frac{T}{\mu_i^3}\Bigg/\left[\frac{T}{\mu_i} - 1 + \exp(-T/\mu_i)\right], \; i = \text{C, E}$$

Lachin[5] presented a further generalization of this formula. He assumed recruitment of patients over the interval $(0, T_0)$, but with a follow-up until time T, where the interval (T_0, T) may be called a 'continuation period'. We obtain the desired sample size under these conditions from the formula given in the previous paragraph, setting

$$\phi(\mu_i) = \frac{1}{\mu_i^2}\left[1 - \frac{\{\exp[-(t-T_0)/\mu_i] - \exp(-T/\mu_i)\}\,\mu_i}{T_0}\right]^{-1}$$

Test procedures for comparing survival distributions that can handle censored data include Gehan's generalization of the Wilcoxson test[25] and the Mantel–Haenszel test[26] for comparing survival distributions.

(5.6) An alternative approach to estimating the number of patients required for a clinical trial comparing survival distributions is to determine the required duration of the trial as a function of θ and the yearly entry rate. Pasternack and Gilbert,[18] George and Desu[19] and Rubenstein, Gail and Santner[27] all pursue this approach. Pasternak and Gilbert assume that all patients are followed until the time of the critical event, have exponentially distributed survival times, and a uniform accrual into the trial. George and Desu generalize this approach, allowing for Poisson rather than uniform accrual. Rubenstein, Gail and Santner generalize both approaches by allowing for a continuation period, during which patient follow-up persists but accrual has terminated. These approaches also take into account loss to follow-up.

Example 5

Consider a trial to evaluate a new chemotherapeutic agent for the treatment of childhood leukaemia. The investigators anticipate that mean survival time for patients on this drug might increase by a factor of 1·5, with follow-up of all patients until death. The required number of patients for the trial (per group) is, from formula 5,

$$n = 2(1{\cdot}645 + 1{\cdot}282)^2/\log_e^2(1{\cdot}5) = 105$$

Suppose now that patients will enter the study over a five-year time period, and that average survival in the control group is 3 years. This implies interest in detection of an increase in this figure to 4·5 years. Then the required sample size per group is obtained from the formula in remark 5.5, where

$$\phi(\mu_C) = (5/3^3)[(5/3) - 1 + e^{-5/3}]^{-1} = 0{\cdot}217$$

$$\phi(\mu_E) = (5/4{\cdot}5^3)[(5/4{\cdot}5) - 1 + e^{-5/4{\cdot}5}]^{-1} = 0{\cdot}125$$

$$n = (1{\cdot}645 + 1{\cdot}282)^2(0{\cdot}217 + 0{\cdot}125)/(4{\cdot}5^{-1} - 3^{-1})^2 = 238$$

This calculation assumes that patients enter the study throughout a five-year period. Suppose instead that they are recruited over a four-year period, but followed-up for an additional twelve months, so that the total study duration remains five years. Then we may use the formula given in the last line of remark 5·5 to obtain $\phi(\mu_C) = 0{\cdot}184$, $\phi(\mu_E) = 0{\cdot}105$, yielding $n = 207$. Note therefore that fewer patients need enter a 5-year trial in which recruitment terminates within 4 years than a trial of similar duration in which patient accrual continues throughout.

6. Sample size requirements that account for patient dropout

Commonly during the course of a clinical trial some patients assigned to the experimental regimen ‘drop out’ or fail to adhere to the prescribed protocol, although their outcomes are still recorded. Since one must count such individuals against the experimental group in the statistical analysis (Reference 6, Chapter 13), the effect of patient drop-out is to dilute the effective treatment difference. Several approaches have evolved for taking this problem into account in the calculation of sample size requirements. We summarize three such approaches here, all of which model the effect of drop-outs through the ultimate effect on the proportions under comparison.

6.1. *An approach based on characterization of drop-outs by the control group event rate*

Lachin[5] has proposed a very simple method of adjusting sample size requirements for an anticipated drop-out rate d among patients in an experimental group. This approach characterizes drop-outs by the control event rate P_C, rather than the event rate P_E corresponding to their original group assignment. It follows that the effective value P_E^* of the T-year event rate P_E is $P_E^* = P_E(1-d) + P_C d$, and the effective treatment difference δ^* by $P_E^* - P_C = (1-d)(P_E - P_C)$. Substitution of δ^* for δ in formula 1 implies division of the usual formula for sample size requirements by the factor $(1-d)^2$ to inflate appropriately the number of patients entered into the trial.

6.2. *An approach that accounts for specific patterns of drop out*

Schork and Remington[28] proposed an approach which takes into account yearly ‘shifts’ of subjects from the experimental group to the control group. (They also deal with shifts in the reverse direction, which may occur, for example, in life-style intervention trials, where control group subjects may voluntarily seek out the benefits of treatment. For simplicity of discussion, we do not consider the effect of such ‘drop-ins’ in this paper.) Under Schork and Remington's approach, subjects who drop-out effectively become characterized by the control group event rate P_C from that point onward. For convenience, we assume that a subject's shift occurs mid-year.

To apply this approach, one must anticipate, on the basis of past experience, a particular pattern of shift, i.e. the percentage of subjects anticipated to shift from the experimental to control group per unit time (say, year) of the trial. For example, we might assume for a 5-year trial that 20 per cent of experimental patients will discontinue therapy during the first year, and that an additional 10 per cent will do so in each of the remaining four years of the study. For any given pattern of shift, Schork and Remington provide a formula for the effective T-year experimental group event rate P_E^* for use in formula 1. P_E^* will in general be closer to P_C than P_E, thus inflating the number of subjects required for the trial.

Suppose that the anticipated *yearly* event rates in the experimental and control groups are P_{EY} and P_{CY}, respectively, and the yearly drop-out rate in the experimental group is $d_i, i = 1, 2, \ldots, L$, where L is the study duration in years. Then the effective T-year event rate in the experimental group is

$$P_{\mathrm{E}}^* = \sum_{i=1}^{L} d_i[1-(1-P_{\mathrm{EY}})^{i-1/2}(1-P_{\mathrm{CY}})^{L-i+1/2}]+C[1-(1-P_{\mathrm{EY}})^2]$$

where $C = 1 - \Sigma d_i$ is the proportion of experimental group subjects anticipated to complete the study. One may easily evaluate this formula for any given values of P_{EY}, P_{CY} and $d_i, i = 1, 2, \ldots, L$ and then substitute P_{E}^* for P_{E} in formula 1.

6.3. *An approach based on modelling the distribution of time to drop-out*

Halperin *et al.*[29] developed a theoretical model to account for patient drop-outs in a clinical trial. They assumed that time-to-event among control group subjects and time-to-drop-out among experimental group subjects each follow an exponential distribution, equivalent to the assumption that the corresponding instantaneous risks (hazards) of event and drop-out in these two groups, respectively, are constant. In their simplest model, they also assume that the instantaneous event rate among drop-outs returns immediately to the control group rate. With these assumptions, they derived the effective T-year experimental group event rate P_{E}^* in terms of P_{C}, the anticipated maximum reduction in P_{C} (specified as a proportion k of P_{C}), and the anticipated T-year drop-out rate d. Formula 1 then applies as before, with P_{E}^* replacing P_{E}.

The calculation of the effective P_{E} with the above assumptions requires numerical integration; thus no simple formula for this quantity exists. However, the authors provide extensive tables of P_{E}^* in terms of the assessed values P_{C}, d and k.

Remarks

(6.1) We should not confuse the issue of patient drop-out with the related issue of patient loss to follow-up. The approaches described above assume that drop-outs remain under surveillance for the duration of the study, and are thus not lost to follow-up. The term loss to follow-up refers instead to those patients whose end-point status does not become available, in spite of (possibly extensive) surveillance efforts. If the reason for loss to follow-up is related to group assignment, serious problems of interpretation can arise. If, on the other hand, one expects loss to follow-up rates, l, to be identical in the two groups, and one can further assume that the character of the drop-outs is no different between groups, an appropriate adjustment consists of multiplying the estimated value of n by the factor $1/(1-l)$.

(6.2) As mentioned above, the assumption of exponential time to event is equivalent to the assumption that the instantaneous event rate (the probability of having an event within any unit time interval, given that the event has not already occurred) is constant throughout the T-year study period. Wu, Fisher and DeMets[30] have generalized this assumption by dividing the time dimension into intervals and allowing for different instantaneous event rates in the different intervals. They present a similar generalization for the drop-out model, allowing as well for a time-dependent instantaneous drop-out rate. The authors present an equation for the effective value of P_{E} under their generalized model, and provide an example of its application.

(6.3) Lachin's approach is the most conservative of the three described, since it characterizes drop-outs completely by the control group event rate. The other two approaches characterize drop-outs by the experimental group rate up until the time they drop out. Schork and Remington's approach most suits trials in which evidence from previous studies provides a characterization of the drop-out pattern that is likely to emerge from year to year.

Example 6

Consider a clinical trial to compare groups of myocardial infarction patients with respect to the five-year recurrence of myocardial infarction. Suppose that the experimental intervention consists of enrolling patients in a mild-activity exercise programme, whereas control group subjects receive no intervention. Past experience and the chosen eligibility criteria for the trial suggest that 80 per cent of the participants will adhere to the exercise programme. Assume further a 5-year recurrence rate among control group subjects of about 0·25, and interest in detection of a reduction in this rate to 0·15 among experimental group subjects.

Ignoring drop-outs and setting $\alpha = 0{\cdot}05$ (one-sided) and $\beta = 0{\cdot}20$, formula 1 yields 213 as the required number of subjects per group. Using Lachin's adjustment factor to account for the anticipated drop-out rate, the required number of subjects in each group increases to $213/(1 - 0{\cdot}20)^2 = 333$.

Using the approach developed by Schork and Remington, one must specify the anticipated yearly drop-out rate in the experimental group. If we set $d_1 = 0{\cdot}10$, $d_2 = 0{\cdot}05$, $d_3 = d_4 = 0{\cdot}025$, $d_5 = 0$, then Schork and Remington's solution for the effective value of P_E is

$$P_E^* = \sum_{i=1}^{5} d_i[1-(1-0{\cdot}03)^{i-1/2}(1-0{\cdot}05)^{5-i+1/2}] + 0{\cdot}80[1-(1-0{\cdot}03)^5]$$
$$= 0{\cdot}154$$

where $P_{EY} = 0{\cdot}03$ and $P_{CY} = 0{\cdot}05$ are the approximate yearly event rates anticipated in the two groups. Substitution of this value of P_E in formula 1, with $P_C = 0{\cdot}25$, shows the required number of subjects for the trial is $n = 298$.

Using the approach developed by Halperin *et al.*, one must also calculate the effective value of P_E, which, under their model, depends only on P_C, the overall anticipated drop-out rate d, and the expected reduction in P_C, expressed as a proportion k. For our example, $P_C = 0{\cdot}25$, $d = 0{\cdot}20$ and $k = (0{\cdot}25 - 0{\cdot}15)/0{\cdot}25 = 0{\cdot}40$, yielding, from the author's tables, $P_E^* = 0{\cdot}128$. Applying formula 1, the required number of subjects for the trial is $n = 326$, only slightly less than the result obtained with Lachin's much simpler approach.

7. Sample size requirements that account for the length of time required to achieve maximum benefit of treatment

In trials of long-term therapy, the realization of full benefit may not be immediate, and may, in fact, require a fairly lengthy period of treatment. To account for this possibility in sample size planning, Halperin *et al.*[29] assumed achievement of the full effect of treatment among non-drop-outs in the exposed group in a linear fashion in f years. This assumption allows the development of an expression for the effective T-year event rate in the experimental group, denoted by P_E^*, as a function of P_C (the anticipated T-year event rate in the control group), the anticipated maximum reduction in P_C (as a proportion, k, of P_C), and f (as a fraction of T). One can then apply formula 1 as usual. As a further refinement, one may also extend this model to account for an anticipated T-year drop-out rate d, as discussed in 6.3. In this case, we assume further that the time until drop-out has an exponential distribution, and that the underlying instantaneous event rate for drop-outs in the experimental group returns to the control group level in a linear fashion, and at the same rate as incidence declined before drop-out. This is a generalization of the development given in 6.3, where the assumption was that drop-outs returned immediately to the control group level ($f = 0$).

The calculation of P_E^* under the above assumptions requires numerical integration. However, Halperin *et al.* provided tables for the quantity in terms of P_C, d and k for $f = \{0, T/2 \text{ and } T\}$. Clearly P_E^* will approach P_C as d and f increase, thus inflating the eventual sample size requirements.

Example 7

Continuing example 6, suppose we now assume that the full benefit of treatment required at least eighteen months for achievement. Setting $P_C = 0{\cdot}20$, $k = 0{\cdot}40$, $d = 0{\cdot}20$, and $f = T/2$, tables provided by Halperin *et al.* give $P_E^* = 0{\cdot}146$ as the effective value of P_E. Applying formula 1, the required number of subjects for the trial is $n = 599$. Assuming total adherence ($d = 0$), the required number is $n = 581$.

8. Sample size requirements that account for patient accrual by cohorts

The model developed by Halperin *et al.*[29] assumes follow-up of at least T years for each patient entered into the trial so that fixed sample-size formulae (such as formula 1) will apply. Suppose, however, that patients enter the study at essentially a uniform rate over a T-year period, at which time patient accrual terminates, but with patient follow-up continuing for an additional r years. Thus not all patients will have experienced T years of follow-up, although the purpose of the trial is to compare T-year event rates (using life-table methods).

Extending the assumptions of Halperin *et al.*, Pasternack[31] assumed an exponential survivorship function in both the experimental and control groups. With this assumption, and using an analytic approach similar to that of Halperin *et al.*, Pasternack derived sample size requirements in terms of the expected T-year event rate in the control group and its anticipated decrease in the experimental group. He presented tables for $T = 5$ and $r = 1$, with the assumption of a one-tailed significance test. Palta[32] has pointed out that these tables are generally conservative when survival in the control group is 50 per cent or higher, and recommends they not be used for control group survival probabilities greater than 60 per cent. She also suggested caution with accrual rates less than 30 per year. In this case, the five-year survival rate may not be estimable because of small numbers, and Palta therefore recommended an increase in the trial size beyond the level indicated by Pasternack's tables.

Example 8

Suppose, as in example 7, patients enter the trial at approximately a uniform rate over a five-year period, with patient follow-up continuing for one additional year. Then, without adjusting for subject drop-out or the length of time required to achieve full treatment benefit, the size of the annual cohort per group is, from Pasternack's tables (with $P_E = 0{\cdot}25$ and $P_C = 0{\cdot}15$), 115. Thus the total number of subjects per group over the five years of entry is $n = 5(115) = 575$. This is considerably more than the 249 subjects per group as calculated for a fixed sample trial with five-year follow-up for each patient.

One could also use Pasternack's tables to incorporate an anticipated drop-out rate d and the time f required to achieve maximum benefit of treatment, with assumptions similar to those adopted by Halperin *et al.*

9. Sample size requirements for group randomization

The clinical therapeutic trial involves randomization of individual subjects to experimental and control groups. Controlled trials of preventive measures and innovations in health care, however, necessitate randomization of groups of people rather than individuals.[33] Since one cannot regard the individuals within such groups as statistically independent, standard sample size formulae underestimate the total number of subjects required for the trial. Cornfield[34] addressed this problem in the context of community studies with randomization of equal-sized groups to each of two interventions. Donner[35] extended his results to the case of unequal-sized groups.

Consider a sample of m groups with the size of the ith group denoted by g_i, $i = 1, 2, \ldots, m$. One may view each of these groups as having its own proportion of successes P_i, $i = 1, 2, \ldots, m$. Suppose we randomly allocate $m/2$ of the groups to each of two interventions. Then the relative efficiency of group randomization to individual randomization is $R = \bar{P}(1-\bar{P})/\bar{g}\sigma^2$, where $\bar{g} = \Sigma g_i/m$, $\bar{P} = \Sigma g_i P_i/\Sigma g_i$ and $\sigma^2 = \Sigma g_i(P_i - \bar{P})^2/(m\bar{g}^2)$. This expression implies that one should multiply the usual estimate of the required number of individuals in the trial by an inflation factor $\text{IF} = 1/R$ to provide the same statistical power under group randomization as would be obtained under individual randomization. Donner[35] has shown that IF may be written as $\text{IF} = 1 + (g-1)\hat{\kappa}$, where $\hat{\kappa}$ measures the degree of within-group dependence and corresponds to the version of the kappa statistic presented by Fleiss.[4] For small relatively homogeneous groups (clusters) of fixed size g, Donner, Birkett and Buck[36] have shown that one can interpret κ directly in terms of the underlying concordancy rate, where a concordant cluster is defined as one in which the responses of all members are identical. Specifically in this case

$$\hat{\kappa} = \frac{P_{\text{CT}} - [P_{\text{C}}^g + (1 - P_{\text{C}})^g]}{1 - [P_{\text{C}}^g + (1 - P_{\text{C}})^g]}$$

where P_{CT} is an advance estimate of the proportion of concordant clusters in the control population, and P_{C} is the anticipated success rate among control patients. One may also interpret κ as an intraclass correlation coefficient.

Group randomization requires modification in standard analytic methods since these assume statistical independence among individuals. The essential feature of any valid analytic approach is the development of a variance expression for the difference between the two proportions which accounts for between-group variation.[34] The application of standard inference techniques to data arising from group randomization may result in spurious statistical significance, since they ignore this source of variation.

Example 9.1

Consider a study of cardiovascular mortality with randomization of either a control condition or a life-style intervention designed to modify coronary risk factors. Data from 15 cities show that the mean and standard deviation of cardiovascular mortality rates among these communities are 250×10^{-5} and 19×10^{-5}, respectively. Assuming randomization of equal-sized cities of approximately 50,000 individuals, the appropriate inflation factor is

$$\text{IF} = \frac{\bar{g}\sigma^2}{\bar{P}(1-\bar{P})} = \frac{(50{,}000)(0{\cdot}00019)^2}{(0{\cdot}00025)(0{\cdot}99975)} = 7{\cdot}22$$

Thus the result of standard sample calculations, as provided by formula 1, for example, must have an inflation factor of 7·22 to allow for the reduced efficiency of group randomization.

Example 9.2

Consider a study with randomization of spouse pairs to either a control group or a group receiving a reduced amount of dietary sodium, where the interest is hypertensive status ($<$ 140/90 mmHg vs. $\geqslant$ 140/90 mmHg). Previous data indicate that the proportion of couples concordant with respect to hypertensive status is 0·85, whereas the anticipated rate of hypertension in the control population is 0·15. Thus

$$\hat{\kappa} = \frac{P_{\mathrm{CT}} - [P_{\mathrm{C}}^{g} + (1-P_{\mathrm{C}})^{g}]}{1 - [P_{\mathrm{C}}^{g} + (1-P_{\mathrm{C}})^{g}]} = \frac{0{\cdot}85 - [0{\cdot}15^2 + 0{\cdot}85^2]}{1 - [0{\cdot}15^2 + 0{\cdot}85^2]} = 0{\cdot}41$$

Consequently, one must multiply the value of n as obtained from standard sample size calculations by a factor $1 + (2-1)(0{\cdot}41) = 1{\cdot}41$ to allow for the clustering effect within spouses.

10. Sample size requirements for cross-over designs

The two-period cross-over design involves a single group of patients, each of whom serves as his own control in the comparison of two treatments. One randomizes patients to one of two treatment sequences—half the patients receive the treatments in the order EC, the other half in the order CE. The advantage of this design over a parallel group or completely randomized design is that it allows the effects of the treatments to be compared within the same patients.

For quantitative responses, the total number of subjects N required in a cross-over experiment to provide the same statistical power as would be obtained in a parallel group experiment is $N = n(1-\rho)$, where n derives from formula 1 and ρ in the correlation between two responses in a single subject. This result shows that the advantage of the cross-over trial is greatest when variation of responses within subjects is small compared to variation between subjects.

As applied to a dichotomous response, the correlation ρ between responses in a single subject is not, in a strict sense, meaningfully defined. However, Schwartz, Flamant and Lellouch[2] show that one may still obtain approximate sample size requirements in this case by estimating ρ from previous data, with observations scored as 0 or 1.

Remarks

(10.1) Although use of the cross-over design always requires fewer patients than the completely randomized design (even at $\rho = 0$, where half as many patients are required), the cross-over design, in practice, has severe limitations. The most important is that there must be no 'carry-over' effect of the first treatment into the period in which the second treatment is applied.[37] A second limitation is that the cross-over design does not apply with responses such as death or total cure, since each patient must receive both treatments. Cross-overs are most suited to the comparison of treatments for chronic conditions, such as insomnia, pain and asthma.[2]

(10.2) The structure of this design implies that the eventual test of $\mathrm{H}_0{:}P_{\mathrm{E}} = P_{\mathrm{C}}$ will employ methods for comparing dichotomous outcomes in matched pairs, where each pair consists of the binary observations produced by a subject in periods 1 and 2, respectively. The most well-known such method is McNemar's test for correlated proportions (Reference 24, pp. 177–179).

Example 10

We focus on the estimation of ρ from previous data, since this is the only new feature. Consider a cross-over study to compare the preference of patients for two drugs A and B, where a score of 0 denotes a preference for A and a score of 1 a preference for B. To extract an estimate of ρ from past data, a fourfold table may be constructed whose entries provide the frequencies of 0 and 1 in each of the two periods of the previous trial. It is well-known that the square of the correlation between the responses in periods 1 and 2 is given by the value of χ^2 for this table divided by the number of observations. Thus for a prior study of 20 patients yielding $\chi^2 = 4{\cdot}5$, the estimate of ρ is given by 0·47, which could equivalently be obtained by applying the usual formula for a correlation coefficient to the (0, 1) observations. Use of this estimate in the planning of a cross-over trial implies that the total number of subjects required is slightly more than half the number of subjects required in each group of a completely randomized trial.

REFERENCES

1. Freiman, J. A., Chalmers, T. C., Smith, H. Jr. *et al.* 'The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of "negative" trials', *New England Journal of Medicine*, **299**, 690–694 (1978).
2. Schwartz, D., Flamant, R. and Lellouch, J. *Clinical Trials*, Academic Press, London, 1980.
3. Lachin, J. M. 'Sample size determinations for $r \times c$ comparative trials', *Biometrics*, **33**, 315–324 (1977).
4. Fleiss, J. L. *Statistical Methods for Rates and Proportions*, Wiley, New York, 1981.
5. Lachin, J. M. 'Introduction to sample size determination and power analysis for clinical trials', *Controlled Clinical Trials*, **2**, 93–114 (1981).
6. Friedman, L. M., Furberg, C. D. and DeMets D. L. *Fundamentals of Clinical Trials*, John Wright, Boston, 1981.
7. Feinstein, A. R. *Clinical Biostatistics*, C. V. Mosley, Saint Louis, 1977.
8. Cohen, J. R. *Statistical Power Analysis for the Behavioural Sciences*, Revised Edition, Academic Press, New York, 1977.
9. Snedecor, G. W. and Cochran, W. G. *Statistical Methods*, Seventh Edition, Iowa State University, Ames, Iowa, 1980.
10. Hornick, C. W. and Overall, J. E. 'Evaluation of three sample size formulae for 2 × 2 contingency tables', *Journal of Educational Statistics*, **5**, 351–362 (1980).
11. Fleiss, J. L., Tytun, A. and Ury, H. K. 'A simple approximation for calculating sample size for comparing independent proportions', *Biometrics*, **36**, 343–346 (1980).
12. Schlesselman, J. J. 'Sample size requirements in cohort and case-control studies of disease', *American Journal of Epidemiology*, **99**, 381–384 (1974).
13. Zar, J. H. *Biostatistical Analysis*, Prentice-Hall, Englewood Cliffs, 1974.
14. Makuch, R. and Simon, R. 'Sample size requirements for evaluating a conservative therapy', *Cancer Treatment Reports*, 1037–1040 (1978).
15. Blackwelder, W. C. 'Proving the null hypothesis in clinical trials', *Controlled Clinical Trials*, **3**, 345–354 (1982).
16. Gail, M. 'The determination of sample sizes for trials involving several independent 2 × 2 tables', *Journal of Chronic Diseases*, **26**, 669–673 (1973).
17. Cox, D. R. *The Analysis of Binary Data*, Methuen, London, 1970.
18. Pasternack, B. S. and Gilbert, H. S. 'Planning the duration of long-term survival time studies designed for accrual by cohorts', *Journal of Chronic Diseases*, **27**, 681–700 (1971).
19. George, S. L. and Desu, M. M. 'Planning the size and duration of a clinical trial studying the time to some critical event', *Journal of Chronic Diseases*, **27**, 15–24 (1977).
20. Lee, E. T. *Statistical Methods for Survival Data Analysis*, Lifetime Learning, California, 1980.
21. Makuch, R. W. and Simon, R. M. 'Sample size requirements for comparing time-to-failure among k treatment groups', *Journal of Chronic Diseases*, **35**, 861–867 (1982).
22. Gross, A. J. and Clark, V. A. *Survival Distributions: Reliability Applications in the Biomedical Sciences*, Wiley, New York, 1975.
23. Freedman, L. S. 'Tables of the number of patients required in clinical trials using the logrank test', *Statistics in Medicine*, **1**, 121–130 (1982).
24. Colton, T. *Statistics in Medicine*, Little, Brown, Boston, 1974.
25. Gehan, E. A. 'A generalized Wilcoxon test for comparing arbitrarily singly-censored samples', *Biometrika*, **52**, 203–223 (1965).
26. Mantel, N. 'Evaluation of survival data and two new rank order statistics arising in its consideration', *Cancer Chemotherapy Reports*, **50**, 163–170 (1966).
27. Rubenstein, L. V., Gail, M. H. and Santner, T. J. 'Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation', *Journal of Chronic Diseases*, **34**, 469–479 (1981).
28. Schork, M. A. and Remington, R. D. 'The determination of sample size in treatment-control comparisons for chronic disease studies in which dropout or non-adherence is a problem', *Journal of Chronic Diseases*, **20**, 233–239 (1967).

29. Halperin, M., Rogot, E., Gurian, J. and Ederer, F. 'Sample size for medical trials with special reference to long-term therapy', *Journal of Chronic Diseases*, **21**, 13–24 (1968).
30. Wu, M., Fisher, M. and DeMets, D. 'Sample sizes for long-term medical trials with time-dependent dropout and event rates', *Controlled Clinical Trials*, **1**, 111–123 (1980).
31. Pasternack, B. S. 'Sample sizes for clinical trials designed for patient accrual by cohorts', *Journal of Chronic Diseases*, **25**, 673–681 (1972).
32. Palta, M. 'Determining the required accrual rate for fixed-duration clinical trials', *Journal of Chronic Diseases*, **35**, 73–77 (1982).
33. Buck, C. and Donner, A. 'The design of controlled experiments in the evaluation of non-therapeutic interventions', *Journal of Chronic Diseases*, **35**, 531–539 (1982).
34. Cornfield, J. 'Randomization by group: a formal analysis', *American Journal of Epidemiology*, **108**, 100–102 (1978).
35. Donner, A. 'An empirical study of cluster randomization', *International Journal of Epidemiology*, **11**, 283–286 (1982).
36. Donner, A., Birkett, N. and Buck, C. 'Randomization by cluster-sample size requirements and analysis', *American Journal of Epidemiology*, **114**, 906–914 (1981).
37. Brown, B. 'The crossover experiment for clinical trials', *Biometrics*, **36**, 69–79 (1980).