Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials.

STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:2589–2603
Published online 8 November 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2740

Relative efficiency of unequal *versus* equal cluster sizes in cluster randomized and multicentre trials

Gerard J. P. van Breukelen*,†, Math J. J. M. Candel and Martijn P. F. Berger

Department of Methodology and Statistics, Maastricht University, The Netherlands

SUMMARY

Cluster randomized and multicentre trials evaluate the effect of a treatment on persons nested within clusters, for instance, patients within clinics or pupils within schools. Optimal sample sizes at the cluster (centre) and person level have been derived under the restrictive assumption of equal sample sizes per cluster. This paper addresses the relative efficiency of unequal *versus* equal cluster sizes in case of cluster randomization and person randomization within clusters. Starting from maximum likelihood parameter estimation, the relative efficiency is investigated numerically for a range of cluster size distributions. An approximate formula is presented for computing the relative efficiency as a function of the mean and variance of cluster size and the intraclass correlation, which can be used for adjusting the sample size. The accuracy of this formula is checked against the numerical results and found to be quite good. It is concluded that the loss of efficiency due to variation of cluster sizes rarely exceeds 10 per cent and can be compensated by sampling 11 per cent more clusters. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: cluster randomized trials; mixed regression; multicentre trials; optimal design; relative efficiency; sample size; power

1. INTRODUCTION

The effect of medical and behavioural treatments is often evaluated by a randomized trial involving many clusters (or centres). Examples are a study of a new antibiotic drug for a urinary tract infection among patients of general practices, and of a smoking prevention programme for pupils in high school. Data analysis has to take this nesting of persons within clusters into account. Clusters may be treated as fixed or random, depending on several considerations [1]. This paper treats clusters as random, which is a prerequisite for generalization of treatment effects, not only to other patients but also to other clusters than those included in the study. Our focus is therefore on studies with

*Correspondence to: Gerard J. P. van Breukelen, Department of Methodology and Statistics, Maastricht University, P. O. Box 616, Maastricht, 6200 MD, The Netherlands.
†E-mail: gerard.vbreukelen@stat.unimaas.nl

Received 23 December 2005
Accepted 26 September 2006

a large number of clusters. Such studies are often called cluster randomized trials if clusters are randomized to treatment [2, 3], and multicentre trials if persons within each cluster are randomized to the treatment [1, 4]. In the sequel we will adopt this terminology. The data of both types of trials are best analysed with mixed regression, although ANOVA of cluster means is also used [5].

The planning of cluster randomized and multicentre studies involves several design choices, among others the choice between randomization of clusters or of persons, and the sample size, that is, the number of clusters and the number of persons per cluster. The optimal sample size can be defined as the sample size that gives the most precise treatment effect estimate and the largest test power under the precondition of a fixed budget for sampling. Alternatively, it can be defined as the cheapest sample size under the precondition of a certain test power and precision of effect estimation. Formulae for computing the optimal sample size are available for cluster randomized and multicentre trials [2, 4, 6–10]. These formulae assume that the number of persons sampled is the same for each included centre or cluster. Although equal cluster sizes are optimal for estimating treatment effects [11], they are rarely encountered in practice. Among others, variation of the actual size of the organizational units studied (schools, hospitals, general practices) and non-response and dropout of persons generate unequal cluster sizes in the study. Apart from the issue of bias due to selective dropout, an obvious question then is how much power is lost by cluster size variation and how to compensate for this. Some results on the relative efficiency (RE) of unequal *versus* equal cluster sizes have already been obtained in literature. A lower bound for this RE has been derived for treatment effect estimation in cluster randomized trials [12] and multicentre trials [13]. It has also been shown that the design effect, which is the ratio of the sample size needed for cluster randomization and the sample size needed for person randomization for both designs to have the same power [14], is hardly affected by cluster size variation of the type to be expected in trials with general practices in the U.K. [15].

The present paper expands the results from literature in two ways. Starting from maximum likelihood (ML) estimation in cluster randomized trials, it will be shown that although equal cluster sizes are optimal, the RE of unequal *versus* equal cluster sizes usually remains high. In addition, a Taylor approximation of the RE will be presented as a simple function of the intraclass correlation and the coefficient of variation of cluster size. The practical implications of both results for study design will also be shown. The structure of this paper is as follows. Section 2 presents the model and estimation procedure for cluster randomized trials, and Section 3 introduces the design criteria that we use and defines RE in the context of nested designs. Formulae for the RE of unequal cluster sizes are presented in Section 4. Section 5 applies these formulae to several cluster size distributions, and evaluates the accuracy of the Taylor approximation. Person randomization is covered in Section 6, showing that the RE results for cluster randomized trials by and large also hold for multicentre trials with treatment by centre interaction. Section 7 shows how our results can be used for planning a trial. The paper ends with some conclusions on the practical use of our results, and discusses some simple extensions.

2. CLUSTER RANDOMIZED TRIALS: MODEL AND ESTIMATION PROCEDURE

2.1. Model

Suppose that the effectiveness of a new treatment is evaluated by a cluster randomized trial, where K clusters (e.g. general practices or schools) are randomly allocated to treatment or control, where

DOI: 10.1002/sim

control refers to no treatment or to some standard treatment. Within each cluster j ($j = 1, \ldots, K$) a total of n_j patients is included and all patients are given the treatment to which their cluster was allocated. Let the effectiveness of the treatment be expressed as its effect on the average of a quantitative outcome Y. The model for data analysis is

$$Y_{ij} = \beta_{0j} + \beta_1 X_j + e_{ij} \tag{1}$$

with

$$\beta_{0j} = \beta_0 + u_{0j}$$

where Y_{ij} is the outcome for person i in cluster j, and X_j is the treatment to which cluster j is allocated and will be coded $+1$ for treated and -1 for control clusters, although other codings are possible. Given this coding scheme, β_0 is the grand mean of the outcome, and β_1 is half the treatment effect on this outcome. Finally, u_{0j} and e_{ij} are a random cluster effect and a residual error reflecting person and measurement error effects and are assumed to be independently and normally distributed with variances σ_0^2 and σ_e^2, respectively. The outcomes Y_{ij} and $Y_{i'j}$ of any two persons i and i' in an arbitrary cluster j are correlated, and this intraclass correlation is equal to

$$\rho = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_e^2} \tag{2}$$

where the numerator is the covariance of Y_{ij} and $Y_{i'j}$ and the denominator is their variance.

Estimation

Ignoring the individual random effects u_{0j} and e_{ij}, model (1) has four parameters of interest: two fixed regression parameters and two variance components, summarized by the vector $\Theta = (\beta_0, \beta_1, \sigma_e^2, \sigma_0^2)'$ which can be estimated with the ML method. If variance components $(\sigma_e^2, \sigma_0^2)'$ are known, the ML estimator of $(\beta_0, \beta_1)'$ is equal to the generalized least squares (GLS) estimator [16]. For model (1), GLS reduces to weighted least squares (WLS) estimation. Assume that $K/2$ clusters are randomly assigned to the treatment group, and let $j = 1, \ldots, K/2$ denote those treated clusters, and $j = (K/2) + 1, \ldots, K$ the control clusters. The WLS estimator of $(\beta_0, \beta_1)'$ can then be shown to be

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \hat{\mu}_\mathrm{t} + \hat{\mu}_\mathrm{c} \\ \hat{\mu}_\mathrm{t} - \hat{\mu}_\mathrm{c} \end{pmatrix} \tag{3}$$

where μ_t and μ_c are the expected outcome under treatment and control, with estimator:

$$\hat{\mu}_\mathrm{t} = \frac{\sum_{j=1}^{K/2} w_j \overline{Y}_j}{\sum_{j=1}^{K/2} w_j}, \quad w_j = \frac{1}{\mathrm{Var}(\overline{Y}_j)} = \left(\sigma_0^2 + \frac{\sigma_e^2}{n_j}\right)^{-1} \tag{4}$$

where $\overline{Y}_j$ is the observed mean outcome in cluster j. The estimator $\hat{\mu}_\mathrm{c}$ is defined analogously. Since $\hat{\mu}_\mathrm{t}$ and $\hat{\mu}_\mathrm{c}$ are based on independent samples of $K/2$ clusters each, it follows that $\mathrm{Cov}(\hat{\mu}_\mathrm{t}, \hat{\mu}_\mathrm{c}) = 0$,

DOI: 10.1002/sim

and the WLS estimator (3) of $(\beta_0, \beta_1)'$ has variance-covariance matrix

$$\text{VarCov}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = \frac{1}{4} \times \begin{bmatrix}\text{Var}(\hat{\mu}_t) + \text{Var}(\hat{\mu}_c) & \text{Var}(\hat{\mu}_t) - \text{Var}(\hat{\mu}_c)\\ \text{Var}(\hat{\mu}_t) - \text{Var}(\hat{\mu}_c) & \text{Var}(\hat{\mu}_t) + \text{Var}(\hat{\mu}_c)\end{bmatrix} \tag{5}$$

where

$$\text{Var}(\hat{\mu}_t) = \left(\sum_{j=1}^{K/2} w_j\right)^{-1} \tag{6}$$

and likewise for $\hat{\mu}_c$. Due to the randomization, the cluster size distribution is the same in both treatment arms apart from sampling error. As a result, $\text{Var}(\hat{\mu}_t) = \text{Var}(\hat{\mu}_c)$ and so $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0$, assuming homogeneity of variance components across treatments. In practice, the variance components $(\sigma_e^2, \sigma_0^2)'$ are unknown and must be estimated. Asymptotically, however, the equations (4)–(6) then still hold and the fixed parameter estimators in (3) are orthogonal to the variance component estimators [11, 16].

3. DESIGN CRITERIA

If model (1) is correct, the ML estimators are asymptotically unbiased and their covariance matrix can be used as a criterion for evaluating a design ξ. A well-known criterion from optimal design theory is the D-criterion [17, 18]. A design is D-optimal if it minimizes the determinant $\text{Det}(\text{VarCov}(\hat{\Theta}))$ of the asymptotic covariance matrix of estimators. Since $\hat{\Theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_e^2, \hat{\sigma}_0^2)'$ and the fixed estimators are asymptotically orthogonal to the variance component estimators, $\text{Det}(\text{VarCov}(\hat{\Theta}))$ is the product of the determinants of two submatrices: the covariance matrix of fixed estimators given by equation (5), and the covariance matrix of the variance component estimators [11, 16]. This paper uses the determinant of the 2×2 submatrix of fixed estimators in equation (5) as optimality criterion. Minimization of the determinant of such a $p \times p$ submatrix of p estimators is known as D_S-optimization [17]. Additionally, we use $\text{Var}(\hat{\beta}_1)$ as an optimality criterion. This is again a D_S-criterion, now with a 1×1 submatrix.

Two observations motivate our choice for the D-criterion. First, the D-optimal design does not depend on the scale or coding of the predictors, i.e. in our case the coding of treatment indicator variable X. Secondly, the square root of the determinant of the covariance matrix in (5) is proportional to the area of the confidence ellipsoid for $(\beta_0, \beta_1)'$.

Using the D-criterion, and denoting the optimal design as ξ^*, the RE of a design ξ is defined as

$$\text{RE} = \left(\frac{\text{Det}(\text{VarCov}(\hat{\Theta}_s)|\xi^*)}{\text{Det}(\text{VarCov}(\hat{\Theta}_s)|\xi)}\right)^{1/p} \tag{7}$$

where Θ_s denotes the subvector of parameters of interest, and p is its dimensionality. For instance, for the D-criterion of the fixed part, $\Theta_s = (\beta_0, \beta_1)'$ and $p = 2$, and for a single parameter, such as β_1, $p = 1$. The inverse of the RE indicates the number of times design ξ must be replicated to have the same statistical information on Θ_s, and thereby also the same power, as the optimal design ξ^* has.

DOI: 10.1002/sim

In the next section it will be seen that, given the number of clusters K, mean cluster size $\overline{n}$, and 50:50 allocation of clusters to treatment arms, equal cluster sizes are D-optimal. Therefore, ξ and ξ^* will be the design with unequal and equal cluster sizes, respectively.

4. RELATIVE EFFICIENCY OF UNEQUAL *VERSUS* EQUAL CLUSTER SIZES

4.1. The RE: formal expression and some properties

The RE of unequal *versus* equal cluster sizes for estimating $\Theta_s = (\beta_0, \beta_1)'$ follows from equations (5) and (6) and is equal to

$$\mathrm{RE} = \frac{\left(\sum_{j=1}^{K/2} w_j \times \sum_{j=1+K/2}^{K} w_j\right)^{1/2}}{(K/2) \times w_e} \tag{8}$$

where w_e is w_j, the cluster size weight in equation (4), with $n_j = \overline{n}$, the mean cluster size. Note that $w_e \neq \overline{w}$ (where $\overline{w}$ is the mean of the K weights) except if $\rho = 0$. Assuming that both treatment arms have the same cluster size distribution due to randomization, and plugging in (4) gives:

$$\mathrm{RE} = \left(\frac{\overline{w}}{w_e}\right) = \left(\frac{\overline{n} + \alpha}{\overline{n}}\right) \times \frac{1}{K} \sum_{j=1}^{K} \left(\frac{n_j}{n_j + \alpha}\right) \tag{9}$$

where $\alpha = (1 - \rho)/\rho$. Equation (9) shows that

1. The RE does not depend on the number of clusters K, only on the cluster size distribution and the intraclass correlation ρ.
2. There is a trade-off between mean cluster size $\overline{n}$ and ρ. Multiplying all n_j's and α by a factor >0 does not change the RE in (9). So the RE is independent of $\overline{n}$ in the sense that a larger (smaller) $\overline{n}$ gives the same RE at a suitable smaller (larger) ρ.
3. As $\rho \to 0$ or as $\rho \to 1$, the RE $\to 1$, for any cluster size distribution. For $0 < \rho < 1$, the RE is smaller than 1 and so the equal cluster size design is optimal.

The RE in equation (9), RE$\leqslant$1, can be seen by treating cluster size as a random variable U and letting K go to infinity. We then get: $\overline{w} \to E(w(U))$ and $w_e \to w(E(U))$, where $w(U) = \sigma_0^{-2} U/(U + \alpha)$ is a concave function of U. It then follows from the Jensen inequality [19] that RE$\leqslant$1. It may further be observed that, since $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0$ in equation (5) due to randomization, equations (8) and (9) also hold for the RE with respect to β_0 or β_1 apart.

To design a cluster randomized trial, it is important to know how the RE in equation (9) behaves. The RE will therefore be studied in two ways. First, a Taylor approximation of (9) will be given, expressing the RE as a function of the intraclass correlation ρ and the mean and standard deviation of cluster sizes only. Second, equation (9) will be investigated numerically for various cluster size distributions in order to evaluate its behaviour and the accuracy of the Taylor approximation.

4.2. Taylor approximation of the RE

Let the n_j's in the RE of equation (9) be independent realizations of a random variable cluster size with expectation μ_n and standard deviation σ_n. Then define $\mathrm{CV} = \sigma_n/\mu_n$, the coefficient of variation of cluster size. Let $\lambda = (\mu_n/(\mu_n + \alpha))$, which increases from $\lambda \approx 0$ for $\rho \approx 0$ to $\lambda = 1$ for

DOI: 10.1002/sim

$\rho = 1$ (remember that $\alpha = (1 - \rho)/\rho$). The following 2nd order Taylor approximation of (9) can then be obtained (see Appendix A1)

$$\mathrm{RE_t} \approx 1 - \mathrm{CV}^2 \times \lambda(1 - \lambda) \tag{10}$$

which is a quadratic function of λ and CV, with a maximum $\mathrm{RE_t} = 1$ as $\rho \to 0$ (so that $\lambda \to 0$) and as $\rho \to 1$ (so that $\lambda \to 1$) and minimum $\mathrm{RE_t} = 1 - \mathrm{CV}^2/4$ at $\rho = 1/(\mu_n + 1)$ (so that $\lambda = 0.5$). So the Taylor approximation of the RE in (10) has a minimum that depends on the CV of the cluster size only, which is attained at an intraclass correlation ρ that depends on the expected cluster size μ_n, or the mean cluster size $\overline{n}$ for that matter.

4.3. Approximations of the RE from literature

The RE approximation in (10) may be compared with two approximations from the literature. The first approximation [12] estimates $(\beta_0, \beta_1)'$ by unweighted averaging of all individual observations, which is equivalent to weighting cluster means by cluster size in equation (4), that is, $w_j = n_j$. This gives the following approximation [12, equation (3)]

$$\mathrm{RE_{wn}} \approx \frac{1}{1 + \lambda\,\mathrm{CV}^2} \tag{11}$$

with λ and CV as defined in (10). This approximation is a decreasing function of the intraclass correlation ρ, running from a maximum of 1 for $\rho \to 0$ to a minimum $1/(1 + \mathrm{CV}^2)$ for $\rho \to 1$. Since equation (11) is based on a cluster weighting method that is less efficient than the WLS method [20], it underestimates the RE.

Another approximation is based on equal weights, that is, $w_j = 1$ for $j = 1, 2, \ldots, K$, in equation (4), giving the RE in [15, equation (3)]. By Taylor expansion we get (see Appendix B):

$$\mathrm{RE_{w1}} \approx \frac{1}{1 + (1 - \lambda)\,\mathrm{CV}^2} \tag{12}$$

which resembles equation (11) but behaves differently, as it is an increasing function of ρ, running from a minimum $1/(1+\mathrm{CV}^2)$ for $\rho \to 0$ to a maximum of 1 for $\rho \to 1$. Just like (11), equation (12) is based on a weighting method that is less efficient than WLS, yielding an underestimation of the RE to the extent that the Taylor approximation (12) is accurate. The accuracy of all approximations of the RE, (10)–(12), is evaluated in the next section.

5. NUMERICAL EVALUATION OF THE RELATIVE EFFICIENCY AND ITS TAYLOR APPROXIMATIONS

5.1. Study conditions

The RE of equation (9) was numerically investigated, and the accuracy of its approximations $\mathrm{RE_t}$ in (10), $\mathrm{RE_{wn}}$ in (11) and $\mathrm{RE_{w1}}$ in (12) was evaluated, under the following conditions:

1. Five different cluster size distributions: uniform, unimodal, bimodal, positively skewed, and negatively skewed distributions.

DOI: 10.1002/sim

Table I. Cluster size distributions used for computing the relative efficiency.

Distribution	Frequencies	Smallest size	Range R	CV
Uniform	$f_a = f_b = f_c = 40$	$g_a = 20 - R/2$	$R = 4, \ldots, 36$	$0.08 - 0.73$
Unimodal	$f_a = f_c = 30$, $f_b = 60$	$g_a = 20 - R/2$	$R = 4, \ldots, 36$	$0.07 - 0.64$
Bimodal	$f_a = f_c = 48$, $f_b = 24$	$g_a = 20 - R/2$	$R = 4, \ldots, 36$	$0.09 - 0.80$
+Skewed	$f_a = 60$, $f_b = 40$, $f_c = 20$	$g_a = 20 - R/3$	$R = 12, \ldots, 36$	$0.22 - 0.67$
−Skewed	$f_a = 20$, $f_b = 40$, $f_c = 60$	$g_a = 20 - 2R/3$	$R = 12, \ldots, 24$	$0.22 - 0.45$

Note: f_a = number of clusters of size g_a (small), f_b = number of clusters of size g_b (medium), f_c = number of clusters of size g_c (large), $K = 120$ clusters, $\overline{n} = 20$ = average cluster size, $\overline{n} = g_b$ for symmetric distributions, R = range of cluster sizes = $g_c - g_a$, restricted by $g_a > 0$.

2. Three different cluster sizes, $g_a < g_b < g_c$, with frequencies f_a, f_b, f_c.
3. A cluster size range R $(= g_c - g_a)$ increasing from 0 to almost $2\overline{n}$, where $\overline{n}$ is again the mean cluster size.

Details of these conditions are listed in Table I, showing that large variation of cluster size was covered. As a benchmark, note that if cluster sizes are normally distributed, the CV will be about 0.50 at most, and if only the extreme cluster sizes of 2 and $2(\overline{n}-1)$ occur, each with 50 per cent probability, then the CV is almost 1.0. The number of clusters was fixed at $K = 120$ to have integer values for all cluster size frequencies, and the mean cluster size $\overline{n}$ was fixed at 20. As shown in Section 4, the RE does not depend on K, and choosing a larger (smaller) $\overline{n}$ gives the same RE at a suitably chosen smaller (larger) value of the intraclass correlation ρ.

For each of the five distributions, the RE was computed with equation (9) and then plotted as a function of ρ from 0 up to 0.25, for different amounts of cluster size variation. Larger values for ρ are unrealistic for cluster randomized trials [21], and the RE approaches 1 for larger ρ. Results for the positively and negatively skewed distributions were combined into one plot, since cluster size variation is limited for the negatively skewed distribution due to the fact that all cluster sizes must be larger than zero. Additionally, the RE approximations (10) based on ML (11) based on cluster size weighting, and (12) based on equal weighting of cluster means, were plotted to evaluate their accuracy.

5.2. Results

Figure 1 shows the RE plots of the uniform, unimodal, bimodal and skewed cluster size distributions, for the following ranges of cluster size R : $1.0\overline{n}$, $1.2\overline{n}$, $1.5\overline{n}$ and $1.8\overline{n}$. For all distributions and all ranges and CVs, the RE drops from 1 at $\rho = 0$ to a minimum at ρ somewhere between $\rho = 0.05$ and 0.10, and then increases, returning to 1 for $\rho = 1$. Comparing curves within each plot shows that this minimum RE decreases as cluster size variation increases. More specifically, the minimum RE is about 0.90 for a cluster size range $R = 1.5\overline{n}$ (CV ≈ 0.60, depending on the distribution), and about 0.80 for the extreme range $R = 1.8\overline{n}$ (CV ≈ 0.70).

Comparisons between the four plots show that the positively skewed distribution gives the highest RE, followed by the unimodal, uniform and bimodal distribution. This ordering is in line with a theoretical result in Appendix A (equation (A5)) on the effects of skewness and kurtosis of the cluster size distribution on the RE. The negatively skewed distribution is limited to $R = 1.2\overline{n}$ to prevent cluster sizes <0, giving an RE similar to that for the bimodal distribution. It follows

DOI: 10.1002/sim

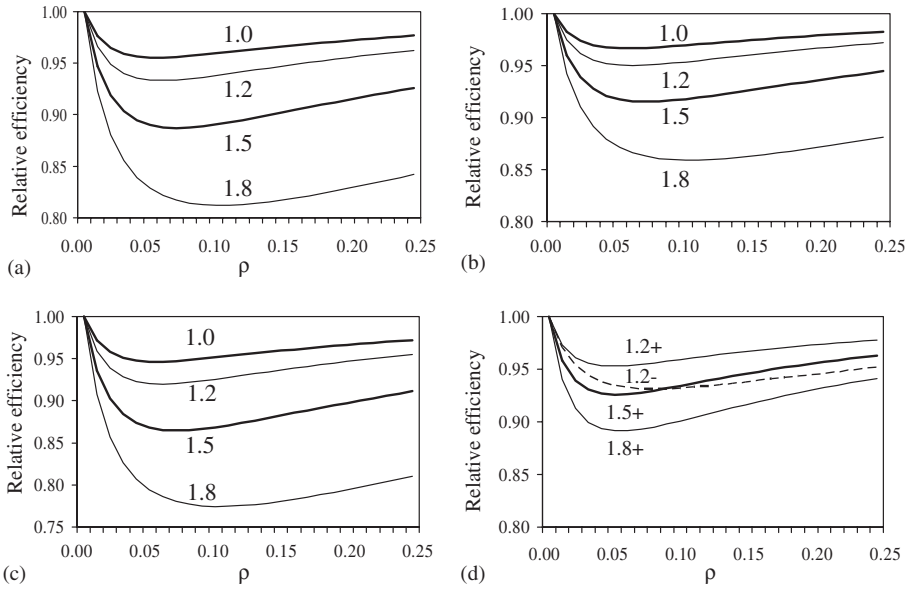

Figure 1. Relative efficiency of unequal *versus* equal cluster sizes plotted against the intraclass correlation ρ for different distributions of cluster size: (a) uniform; (b) unimodal; (c) bimodal; and (d) skewed. Curve numbers 1.0, 1.2, 1.5, 1.8 indicate the ratio $R/\overline{n}$ ($=$range/mean) of the cluster size. In plot (d) a $+$sign indicates positive skew and a $-$sign indicates negative skew. For details per distribution, see Table I.

from equations (5) and (6) that the results in Figure 1 also hold for the RE when based on either $\hat{\beta}_0$ or $\hat{\beta}_1$ rather than on both simultaneously.

5.3. Accuracy of the Taylor approximations of the RE

How accurate are the RE approximations for the cluster size distributions in Table I? Figure 2 shows the RE for the uniform distribution and its approximations RE_t in (10) based on ML, RE_{wn} in (11) based on cluster size weighting, and RE_{w1} in (12) based on equal weighting of cluster means. This is done for cluster sizes 10–20–30 ($R=\overline{n}$, CV $=0.41$) in Figure 2(a), and for cluster sizes 5–20–35 ($R=1.5\overline{n}$, CV $=0.61$) in Figure 2(b). In both the figures, the Taylor approximation RE_t is closer to the RE than approximations RE_{wn} and RE_{w1}. More specifically, RE_t gives an approximation error that is less than 0.01 in Figure 2(a), and less than 0.05 in Figure 2(b), and less than 0.03 at the true RE minimum in Figure 2(b). Plots for the other cluster size distributions in Table I gave approximation errors within these same bounds, and confirmed the superiority of RE_t as compared to RE_{wn} and RE_{w1}. Plots for other cluster size ranges than $R=\overline{n}$ or $1.5\overline{n}$ gave similar results, noting that the error by all approximations increased with the CV.

In practice, extreme cluster sizes such as 2 and $2\overline{n}$ are a minority. We therefore also plotted the RE and its approximations for a uniform distribution with 10 different cluster sizes ranging from 2 to $2\overline{n}-2$, and a CV of 0.57. Its minimum RE was 0.90 and the maximum approximation error by RE_t was 0.04. Plotting the RE for the distribution of general practice size in the U.K. [15], which is a positively skewed distribution with CV $=0.63$ gave a minimum RE of 0.91 and a maximum error by RE_t of 0.01.

DOI: 10.1002/sim

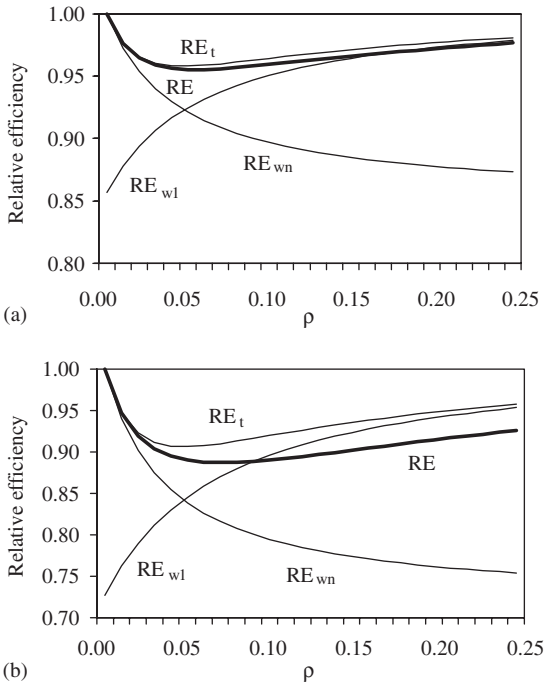

Figure 2. Relative efficiency (RE) of uniformly distributed cluster sizes 10–20–30 (a) and 5–20–35 (b), and Taylor series approximation (RE_t), and approximations based on cluster size weighting (RE_{wn}) and equal weighting (RE_{w1}).

Additional plots of a 4th order Taylor approximation of the RE, which takes into account skewness and kurtosis of the cluster size distribution (see Appendix A equation (A5)), gave an error smaller than 0.02 for all distributions with a range up to $1.5\overline{n}$, and for the uniform distribution with 10 cluster sizes ranging from 2 to $2\overline{n} - 2$.

In summary, the minimum RE was at least 0.90 and the error by the Taylor approximation (12) of the RE was small, for all reasonable distributions of cluster sizes considered. Remember that the RE and its approximation do not depend on K or $\overline{n}$ (see Section 4). So the loss of efficiency due to unequal cluster sizes mainly depends on the CV of cluster size, and this loss can often be compensated by including 11 per cent more clusters. As Figure 2 further shows, the approximations RE_{wn} in (11) based on cluster size weights and RE_{w1} in (12) based on equal weights can be very conservative, depending on the intraclass correlation ρ. Since cluster size weights and equal weights are equivalent to ML in the case of equal cluster sizes (see equation (4)), it also follows from Figure 2 that cluster size weighting and equal weighting of cluster means can be very inefficient compared to ML if cluster size does vary, and that their use for data analysis should be discouraged.

6. MULTICENTRE TRIALS

6.1. Model and estimation procedure

So far, our RE results were obtained for cluster randomized trials. Multicentre trials, that is, trials with randomization of persons within clusters (centres), are also often done. They require a smaller

DOI: 10.1002/sim

sample size [6], but are not always possible and more prone to treatment contamination [22]. This section extends our RE results to the case of multicentre trials with treatment by cluster interaction, assuming 50:50 randomization within each cluster. It will be shown that the RE of unequal *versus* equal cluster sizes in multicentre trials is at least as high as the RE in cluster randomized trials. We assume the following model for the effect of treatment X on a quantitative outcome Y for person i in cluster j:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \tag{13}$$

where

$$\beta_{0j} = \beta_0 + u_{0j} \quad \text{and} \quad \beta_{1j} = \beta_1 + u_{1j}$$

and X is again coded $+1$ for treated and -1 for control, so that β_0 is the grand mean of the outcome, and β_1 is half the treatment effect. As before, u_{0j} and e_{ij} are a random cluster effect and residual error, independently and normally distributed with variances σ_0^2 and σ_e^2. What is new is the random slope effect u_{1j}, reflecting treatment by cluster interaction and assumed to be normally distributed with variance σ_1^2 and covariance σ_{01} with u_{0j}.

The outcome variance for this model is

$$\mathrm{Var}(Y_{ij}|x_{ij}) = \sigma_0^2 + \sigma_1^2 x_{ij}^2 + 2\sigma_{01} x_{ij} + \sigma_e^2 \tag{14}$$

showing that slope–intercept covariance σ_{01} yields heterogeneity of outcome variance between treatments (remember that X is coded as $-1/+1$). Ignoring individual random effects, model (13) has six parameters: two regression parameters and four covariance parameters, which can be estimated with ML. Asymptotically, ML estimation of fixed parameters is again equivalent to GLS estimation [16], which for (13) with $\sigma_{01} = 0$ reduces to WLS:

$$\hat{\beta}_0 = \frac{\sum_{j=1}^{K} w_j \hat{\beta}_{0j}}{\sum_{j=1}^{K} w_j}, \quad \hat{\beta}_1 = \frac{\sum_{j=1}^{K} v_j \hat{\beta}_{1j}}{\sum_{j=1}^{K} v_j} \tag{15}$$

where

$$\hat{\beta}_{0j} = \tfrac{1}{2}(\hat{\mu}_{\mathrm{t}j} + \hat{\mu}_{\mathrm{c}j}), \quad \hat{\beta}_{1j} = \tfrac{1}{2}(\hat{\mu}_{\mathrm{t}j} - \hat{\mu}_{\mathrm{c}j})$$

and $\hat{\mu}_{\mathrm{t}j}$ and $\hat{\mu}_{\mathrm{c}j}$ are the sample means of treated and untreated persons in centre j, and

$$w_j = \frac{1}{\mathrm{Var}(\hat{\beta}_{0j})} = \frac{1}{(\sigma_0^2 + \sigma_e^2/n_j)} \tag{16}$$

and

$$v_j = \frac{1}{\mathrm{Var}(\hat{\beta}_{1j})} = \frac{1}{(\sigma_1^2 + \sigma_e^2/n_j)}$$

are the weights given to cluster j in estimating the overall mean and the treatment effect, respectively, and are (almost) the same as for cluster randomization, increasing with the cluster size at a rate that depends on the (estimated) variance components.

DOI: 10.1002/sim

The covariance matrix of the WLS estimator of $(\beta_0, \beta_1)'$ equals

$$\text{Var Cov}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = \begin{bmatrix}\left(\sum_{j=1}^{K} w_j\right)^{-1} & \sigma_{01}g(\mathbf{w},\mathbf{v})\\ \sigma_{01}g(\mathbf{w},\mathbf{v}) & \left(\sum_{j=1}^{K} v_j\right)^{-1}\end{bmatrix} \tag{17}$$

where $g(\mathbf{w}, \mathbf{v})$ is a function of all weights w_j and v_j. So if $\sigma_{01} = 0$, implying homogeneity of variance, see equation (14), then $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0$. We will assume $\sigma_{01} = 0$.

6.2. Relative efficiency of unequal versus equal cluster sizes

If $\sigma_{01} = 0$, then the RE as based on the determinant of (17) as efficiency criterion, is the geometric mean of the RE for β_0 and the RE for β_1. Comparing (17) with (5) and (6) shows

1. For estimating β_0, the RE satisfies equation (9) and so its Taylor approximation (10) can be used, if we define the intraclass correlation as $\rho_0 = \sigma_0^2/(\sigma_0^2 + \sigma_e^2)$.
2. For estimating β_1, the RE satisfies equation (9) and so its Taylor approximation (10) can be used, if we define the intraclass correlation as $\rho_1 = \sigma_1^2/(\sigma_1^2 + \sigma_e^2)$.

The latter result differs from [13] where, starting from ML estimation and then assuming that ρ_1 goes to zero, the same lower bound (11) is derived for the RE with respect to β_1 in multicentre trials as in [12] for cluster randomized trials. As we saw, this lower bound is very conservative unless the intraclass correlation indeed goes to 0.

As said before, the RE based on the determinant of (17) is the geometric mean of the RE for β_0 and the RE for β_1, which both satisfy equation (9). Further, unless $\rho_0 = \rho_1$, the RE for β_0 and the RE for β_1 reach their minimum at different values of the mean cluster size $\overline{n}$. As a result, the RE based on the determinant of (17) has a minimum which is larger than its counterpart for cluster randomized trials in equations (9) and (10), except if $\rho_0 = \rho_1$ in which case the RE minimum for multicentre trials is the same as that for cluster randomized trials.

7. PRACTICAL USE OF THE RE RESULTS FOR PLANNING A TRIAL

Current sample size formulae for cluster randomized and multicentre trials assume equal cluster sizes, which is optimal but rarely encountered in practice. Our results suggest a simple correction of those formulae for the loss of efficiency and power that is due to variation of cluster size: divide the number of clusters as computed with current sample size formulae by the minimum RE of the cluster size distribution at hand. For instance, the distribution of general practice size in the U.K. had a minimum RE of 0.91 according to equation (9) and its Taylor approximation (10). The number of clusters K must then be divided by 0.91, implying an increase of the number of clusters with 10 per cent, to compensate for the loss of efficiency and power due to cluster size variation. Instead of using the minimum RE, one may compute the RE from equation (10) by inserting an educated guess for λ, which in turn depends on the intraclass correlation ρ and the mean cluster size $\overline{n}$. But the existing sample size formulae that are based on equal cluster sizes require a specification of ρ anyway, and they give K and $\overline{n}$ as output.

DOI: 10.1002/sim

To illustrate this adjustment for cluster size variation, suppose that a cluster randomized trial among general practices in the U.K. is planned. The following optimal number of clusters K and the number of patients per cluster n have been derived [6] for a cluster randomized trial with a sampling budget C and sampling costs c_2 per general practice and c_1 per patient within a sampled practice:

$$K = \frac{C}{\sqrt{\alpha c_1 c_2} + c_2} \quad \text{and} \quad n = \sqrt{\frac{\alpha c_2}{c_1}} \tag{18}$$

where $\alpha = (1 - \rho)/\rho$ as before. Suppose that the sampling budget C is 100.000 Euro's, and the sampling costs are $c_2 = 1000$ Euro per general practice and $c_1 = 100$ Euro per patient. Assume an intraclass correlation $\rho = 0.05$, implying $\alpha = 19$. The optimal design then includes $K = 42$ general practices and $n = 14$ patients per general practice, which requires a budget $C = 100\,800$ Euro.

How many more clusters must be included to compensate for the loss of efficiency and power due to cluster size variation? Given that the CV of general practice size in the U.K. is about 0.63, the minimum RE according to equation (10) is 0.90 and so K must be $42/0.90 = 47$, or 48 to allow 50:50 allocation. The expected RE also follows from (10), by inserting $\overline{n} = 14$ and $\rho = 0.05$, giving RE = 0.90 and the same K as assuming the minimum RE. In this example, choosing the minimum RE or the expected RE does not make a difference.

8. DISCUSSION

Due to variation of the actual size of the organizational units under study, such as general practices or schools, and due to non-response and dropout, cluster randomized trials and multicentre trials show variation of cluster size in the sample. Equal cluster sizes are, however, optimal for estimating a treatment effect with maximum precision and testing it with maximum power. This paper shows that in most cases the loss of efficiency for unequal cluster sizes is at worst about 10 per cent. This loss can be compensated by sampling about 11 per cent more clusters than computed with current sample size formulae which assume equal cluster sizes. Only in case of extreme cluster size variation, for instance with 30 per cent of all clusters having a size close to zero and 30 per cent having a size of twice the average cluster size, is the efficiency loss about 20 per cent. Instead of increasing the number of clusters, one might consider increasing the average cluster size in the sample. But this only reduces sampling variance at the person level, not at the cluster level (see equations (4)–(6) and (15)–(17)). So increasing cluster size is an option only if increasing the number of clusters is impossible or expensive. For details on including cost functions into sample size calculations, see [4, 6, 9].

Additionally, we showed that the RE of unequal cluster sizes can be approximated by a simple function of the mean and CV of cluster size and the intraclass correlation. This is useful for adjusting sample sizes as computed under the assumption of equal cluster sizes.

The RE equations in this paper are based on a 50:50 treatment allocation and on homogeneity of variance components between treatments. Variance components may be heterogeneous and allocation ratio's other than 50:50 can be more efficient if variance components or sampling costs depend on the treatment [9]. However, neither heterogeneity of variance components nor unequal allocation affects the RE of unequal *versus* equal cluster sizes adversely. This can be seen as follows. For cluster randomized trials, it follows from equation (8) that the RE is the geometric mean of two RE's: that for estimating the treated mean μ_t and that for estimating the

DOI: 10.1002/sim

untreated mean μ_c. Each of these two REs is independent of the number of clusters and each of the two still satisfies equation (9) if the variance components are heterogeneous between treatments. In multicentre trials, unequal allocation means that the proportion treated persons per cluster is unequal to 0.50. This does not affect the results in Section 6, except that (σ_e^2/n_j) in equation (16) becomes slightly more complicated. Heterogeneity of σ_e^2 between treatments likewise affects the term (σ_e^2/n_j) only. Heterogeneity of cluster level variance implies $\sigma_{01} \neq 0$, see equation (14). A numerical study and analytical derivations which are beyond the present scope have shown that $\sigma_{01} \neq 0$ affects the RE noticeably only if the slope–intercept correlation and the CV of cluster size are both extremely large. In short, the present results can be generalized to unequal allocation and heterogeneity of variance components.

The models studied, (1) and (13), did not include covariates. Similar results are expected if the models are extended with covariates, at least if we assume fixed slopes for the covariates. Due to randomized treatment assignment, including covariates with a fixed slope into model (1) or (13) does not affect the covariance matrix of $(\hat{\beta}_0, \hat{\beta}_1)'$, apart from sampling error and a decrease of the variance components [7]. However, the RE of unequal *versus* equal cluster sizes in the presence of covariates deserves further study, and the same holds for the RE for estimating variance components, and for the RE in case of binary outcomes.

APPENDIX A: TAYLOR SERIES APPROXIMATION (10) OF THE RE IN EQUATION (9)

Taylor approximation (10) is derived from the RE in (9) by the following steps:

Step 1: Assume that the n_j's are realizations of a random variable cluster size, denoted by U, with expectation μ_n and standard deviation σ_n. Equation (9) then is a moments estimator of

$$\mathrm{RE} = \left(\frac{\mu_n + \alpha}{\mu_n}\right) \times E\left(\frac{U}{U + \alpha}\right) \tag{A1}$$

where $\alpha = (1 - \rho)/\rho \geqslant 0$.

Step 2: Defining $d = (U - \mu_n)$, the last term in (A1) can be rewritten as

$$E\left(\frac{U}{U + \alpha}\right) = E\left(\frac{\mu_n + d}{\mu_n + \alpha + d}\right) = E\left\{\left(\frac{\mu_n + d}{\mu_n + \alpha}\right)\left(\frac{1}{1 + (d/(\mu_n + \alpha))}\right)\right\}$$

The last term is a Taylor series [19, p. 533, equation (34)]

$$\left(\frac{1}{1 + (d/(\mu_n + \alpha))}\right) = \sum_{m=0}^{\infty} \left(\frac{-d}{\mu_n + \alpha}\right)^m$$

if $-(\mu_n + \alpha) < d < (\mu_n + \alpha)$ to ensure convergence.

Since $d = U - \mu_n$ and $\alpha \geqslant 0$, this convergence condition will be satisfied, except for a small probability $P(U > 2\mu_n + \alpha)$ for strongly positively skewed cluster size distributions combined with a large ρ (= small α).

So we have

$$E\left(\frac{U}{U + \alpha}\right) = E\left\{\left(\frac{\mu_n + d}{\mu_n + \alpha}\right) \sum_{m=0}^{\infty} \left(\frac{-d}{\mu_n + \alpha}\right)^m\right\} \tag{A2}$$

DOI: 10.1002/sim

Step 3: Ignoring all terms d^m with $m>2$, and rearranging terms in (A2), will give

$$E\left(\frac{U}{U+\alpha}\right) \approx \lambda \times \{1 - \mathrm{CV}^2 \times \lambda(1-\lambda)\} \tag{A3}$$

where $\lambda = (\mu_n/(\mu_n + \alpha)) \in (0, 1]$ and $\mathrm{CV} = \sigma_n/\mu_n$ is the CV of the random variable U.

Step 4: Plugging (A3) into (A1) gives:

$$\mathrm{RE_t} \approx \{1 - \mathrm{CV}^2 \times \lambda(1-\lambda)\} \tag{A4}$$

Remark

Ignoring in (A2) only those terms d^m with $m>4$ instead of 2, will give

$$\mathrm{RE_t} \approx 1 - \{(1-\lambda)[\lambda \mathrm{CV}^2 - \lambda^2 \mathrm{CV}^3 \mathrm{skew} + \lambda^3 \mathrm{CV}^4(\mathrm{kurt}+3)]\} \tag{A5}$$

where skew and kurt are the skewness and kurtosis of the cluster size distribution, that is, skew = the 3rd central moment of U divided by σ_n^3, and kurt = the 4th central moment of U divided by σ_n^4, minus 3 (see e.g. [19, p. 76]).

APPENDIX B: APPROXIMATION (12) OF THE RE

Step 1: Let $w_j = 1$, so that $\hat{\mu}_\mathrm{t} = (2/K) \times \sum_{j=1}^{K/2} \overline{Y}_j$ and $\mathrm{Var}(\hat{\mu}_\mathrm{t}) = (4/K^2) \times \sum_{j=1}^{K/2} \mathrm{Var}(\overline{Y}_j)$.

Step 2: Insert $\mathrm{Var}(\overline{Y}_j) = (\sigma_0^2 + \sigma_e^2/n_j)$ into $\mathrm{Var}(\hat{\mu}_\mathrm{t})$ for equal and for unequal cluster sizes, and likewise for $\mathrm{Var}(\hat{\mu}_\mathrm{c})$.

Step 3: Apply equations (4) and (5) in Section 2, and let $K \to \infty$ to obtain after some rewriting

$$\mathrm{RE} = \frac{(1-\rho) + \rho\mu_n}{\mu_n E(1/U)(1-\rho) + \rho\mu_n} \tag{B1}$$

which is the asymptotic version of [15, equation (3)].

Step 4: Using the Taylor series expansion

$$E\left(\frac{1}{U}\right) = \left(\frac{1}{\mu_n}\right) E\left(\frac{1}{1+(d/\mu_n)}\right) = \left(\frac{1}{\mu_n}\right) E\left(\sum_{m=0}^{\infty} \left(\frac{-d}{\mu_n}\right)^m\right)$$

and ignoring all terms d^m where $m>2$, gives upon rewriting the following Taylor approximation of equation (B1):

$$\mathrm{RE_{w1}} \approx \frac{1}{1 + (1-\lambda)\mathrm{CV}^2}$$

REFERENCES

1. Senn S. Some controversies in planning and analyzing multi-centre trials. *Statistics in Medicine* 1998; **17**: 1753–1765.
2. Raudenbush SW. Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods* 1997; **2**:173–185.
3. Donner A, Klar N. Cluster randomization trials in epidemiology: theory and application. *Journal of Statistical Planning and Inference* 1994; **42**:37–56.

DOI: 10.1002/sim

4. Raudenbush SW, Liu X. Statistical power and optimal design for multisite trials. *Psychological Methods* 2000; **5**:199–213. DOI 10.1037//1082-989X.5.2.199
5. Moerbeek M, Van Breukelen GJP, Berger MPF. A comparison between traditional methods and multilevel regression for the analysis of multicentre intervention studies. *Journal of Clinical Epidemiology* 2003; **56**: 341–350. DOI 10.1016/SO895-4356(03)00007-6
6. Moerbeek M, Van Breukelen GJP, Berger MPF. Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics* 2000; **25**:271–284.
7. Moerbeek M, Van Breukelen GJP, Berger MPF. Optimal experimental designs for multilevel models with covariates. *Communications in Statistics A* 2001; **30**:2683–2697. DOI 10.1081/STA-100108453
8. Moerbeek M, Van Breukelen GJP, Berger MPF. Optimal experimental design for multilevel logistic models. *The Statistician* 2001; **50**:17–30.
9. Liu X. Statistical power and optimum sample allocation ratio for treatment and control having unequal costs per unit of randomization. *Journal of Educational and Behavioral Statistics* 2003; **28**:231–248.
10. Headrick TC, Zumbo BD. On optimizing multi-level designs: power under budget constraints. *Australian New Zealand Journal of Statistics* 2005; **47**:219–229.
11. Ankenman BE, Aviles AI, Pinheiro JC. Optimal designs for mixed-effects models with two random nested factors. *Statistica Sinica* 2003; **13**:385–401.
12. Manatunga AK, Hudgens MG, Chen S. Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal* 2001; **43**:75–86.
13. Fedorov V, Jones B. The design of multicentre trials. *Statistical Methods in Medical Research* 2005; **14**:205–248. DOI 10.1191/0962280205sm399oa
14. Cochran WG. *Sampling Techniques* (3rd edn). Wiley: New York, 1977.
15. Kerry SM, Bland JM. Unequal cluster sizes for trials in English and Welsh general practices: implications for sample size calculations. *Statistics in Medicine* 2001; **20**:377–390.
16. Searle S, Casella G, McCulloch C. *Variance Components*. Wiley: New York, 1992.
17. Atkinson AC, Donev AN. *Optimum Experimental Design*. Clarendon: Oxford, 1996.
18. Silvey S. *Optimal Design*: *An Introduction to the Theory of Parameter Estimation*. Chapman & Hall: London, 1980.
19. Mood AM, Graybill FA, Boes DC. *Introduction to the Theory of Statistics* (3rd edn). McGraw-Hill: Tokyo, 1974.
20. Searle S, Pukelsheim F. Effect of intraclass correlation on weighted averages. *The American Statistician* 1986; **40**:103–105.
21. Smeeth L, Siu-Woon E. Intraclass correlation coefficients for cluster randomized trials in primary care: data from the MRC trial of the assessment and management of older people in the community. *Controlled Clinical Trials* 2002; **23**:409–421.
22. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? *British Medical Journal* 2001; **322**:355–357. DOI 10.1136/bmj.322.7282.355