Research Article

Statistics in Medicine

Received 1 June 2010, Accepted 23 January 2012 Published online 11 April 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5352

A simple sample size formula for analysis of covariance in cluster randomized trials

Steven Teerenstra,[a*†] Sandra Eldridge,[b] Maud Graff,[c] Esther de Hoop[a] and George F. Borm[a]

For cluster randomized trials with a continuous outcome, the sample size is often calculated as if an analysis of the outcomes at the end of the treatment period (follow-up scores) would be performed. However, often a baseline measurement of the outcome is available or feasible to obtain. An analysis of covariance (ANCOVA) using both the baseline and follow-up score of the outcome will then have more power. We calculate the efficiency of an ANCOVA analysis using the baseline scores compared with an analysis on follow-up scores only. The sample size for such an ANCOVA analysis is a factor r^2 smaller, where r is the correlation of the cluster means between baseline and follow-up. This correlation can be expressed in clinically interpretable parameters: the correlation between baseline and follow-up of subjects (subject autocorrelation) and that of clusters (cluster autocorrelation). Because of this, subject matter knowledge can be used to provide (range of) plausible values for these correlations, when estimates from previous studies are lacking. Depending on how large the subject and cluster autocorrelations are, analysis of covariance can substantially reduce the number of clusters needed. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: ANCOVA; pretest–posttest; cluster randomization; power; sample size

1. Introduction

Cluster randomized trials are trials in which groups or clusters of individuals, rather than individuals themselves, are randomized to intervention groups. One reason to randomize complete clusters is that the intervention can only or is naturally implemented at cluster level (e.g. implementation of protocols in hospital wards). Another reason may be to reduce the risk of contamination (e.g. when a general practitioner has to coach half of its patients, he or she may easily bring elements of coaching to his or her control patients). Logistical, financial or ethical reasons may also dictate the choice for a cluster randomized design [1]. Cluster randomized trial are often applied to evaluate nontherapeutic interventions, including lifestyle modification, educational programmes and innovations in the provision of health care.

Sample size calculations for a continuous outcome in a cluster randomized trial usually proceeds as follows: first, calculating the sample size as if a t-test on the follow-up scores would be carried out, then multiply this number of subjects by the design effect (variance inflation factor) $[1 + (n - 1)\rho]$, where n is the number of subjects in a cluster and ρ the intracluster correlation [1].

Actually, this procedure is applied to an analysis that compares the treatments on the basis of the follow-up scores at the end of treatment period. More precisely, it is the cluster means at follow-up that are compared using a t-test or analysis of variance (ANOVA), or mixed model.

[a]*Departments of Epidemiology, Biostatistics and Health Technology Assessment, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands*

[b]*Centre for Health Sciences, Barts and the London School of Medicine and Dentistry, Queen Mary, University of London, U.K.*

[c]*Scientific Institute for Quality of Healthcare, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands*

**Correspondence to: Steven Teerenstra, Epidemiology, Biostatistics, and HTA , Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands.*

[†]*E-mail: s.teerenstra@ebh.umcn.nl*

However, often also a measurement of the outcome is taken at baseline (baseline scores). Then another analysis is possible: the change from baseline scores can be compared between treatment arms. Again, it is actually the differences of the cluster means at baseline and follow-up that enter the analysis. In this case, power/precision can be gained because (part of) the variation as a result of clusters and possibly subjects can be removed [2].

Actually, even more power could be gained when the two analyses methods, that is, comparison of the follow-up scores and comparison of the change from baseline scores, are combined in a statistically optimal manner. This comes down to an analysis of covariance (ANCOVA), that is, the outcome is analyzed in a linear regression model that includes treatment arm and baseline outcome [3, 4].

In fact, empirical evidence that the power can be gained by analyzing cluster randomized trials using ANCOVA is available. Intracluster correlations (ICCs) that account for covariates such as the outcome at baseline are generally smaller than unadjusted ICC [5, 6], which means that including covariates in the analysis can explain variance at cluster level.

Although this potential for gain in power is commonly known among statistical experts, it is underused in practice when planning cluster randomized trials. The reasons for that include the lack of ICCs that account for covariates and unfamiliarity with this potential for gain among those that plan cluster randomized trials.

Our aim is to provide a simple to use and simple to interpret sample size formula for planning a cluster randomized trial that is analyzed using ANCOVA with the outcome at baseline as a covariate.

2. Methods and results

2.1. Model

The outcome y_{gtik} of subject $k (k = 1, .., n)$ in cluster $i (i = 1, \ldots, I)$ at time t (baseline $t = 0$, follow-up $t = 1$) in treatment arm g(control $g = 0$, intervention $g = 1$) is modelled as [2]

$$y_{gtik} = \mu + \gamma_g + \tau_t + (\gamma\tau)_{gt} + c_i + (c\tau)_{i,t} + s_{ik} + (s\tau)_{ik,t} \tag{1}$$

with c_i, $(c\tau)_{i,t}$, s_{ik}, $(s\tau)_{ik,t}$, normally distributed with mean 0 and variance σ_c^2, $\sigma_{c\tau}^2$, σ_s^2 , $\sigma_{s\tau}^2$, respectively. The first four terms are fixed effects, where μ is the mean outcome in the control clusters at baseline, γ_1 is the difference at baseline between the mean of the intervention and control clusters ($\gamma_0 = 0$), τ_1 is the change from baseline to follow-up of the control clusters means ($\tau_0 = 0$), and $\delta = (\gamma\tau)_{11}$ is the difference in change from baseline between intervention and control cluster means (for $g \neq 1$ or $t \neq 1 : (\gamma\tau)_{gt} = 0$). In a randomized cluster trial, no differences between treatment groups are expected at baseline and $\gamma_1 = 0$, so that $\delta = (\gamma\tau)_{11}$ is also the expected difference of the follow-up scores. The random effects c_i, $(c\tau)_{i,t}$ describe the variation of the clusters, where the first models the variation between clusters at a fixed time point, while the second the variation of each cluster at different time points. Similarly, the random effects s_{ik}, $(s\tau)_{ik,t}$ decompose the variation of subjects in a time-invariant and time-varying part. Following [2], we define

$$\rho_c = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_{c\tau}^2} \quad \text{and} \quad \rho_s = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_{s\tau}^2}, \tag{2}$$

which are called the cluster autocorrelation and subject autocorrelation. They describe the (auto) correlation between baseline and follow-up of the cluster specific scores $c_i + (c\tau)_{i,t}$ and of the subject specific scores $s_{ik} + (s\tau)_{ik,t}$. Apart from the fixed effects for treatment group and time, these cluster specific scores are the precise cluster means, that is, without the sampling error because of averaging over a finite number of subjects in the cluster. (In fact, the observed cluster means (calculated from a finite number of subjects) approximate these cluster specific values if the cluster size is very large).

Note that ρ_s is the correlation between baseline and follow-up of the subject score y_{gtik} , in a fixed cluster (i.e. conditional on the cluster).

The above model describes cohort designs, cross-sectional designs, and even mixtures of these [2]. If $\rho_s = 0$, then all $s_{ik} = 0$, that is, for each time there are different subjects, which is the case of the cross-sectional design. If $\rho_s = 1$, then all $(s\tau)_{ik,t} = 0$, that is, there is no variation within subjects over time: all subjects are measured at baseline and follow-up, with identical scores — a special case of the cohort design. The situation $0 < \rho_s < 1$ can arise when all subjects are measured twice and have an

autocorrelation smaller than 1 (cohort), or when part of the subjects are measured twice (with an autocorrelation $\leqslant 1$) and the other part is replaced with new subjects at follow-up (mixture of cohort and cross-sectional).

2.2. *Estimators of the treatment effect*

The comparison between treatment groups of the follow-up scores is performed via the between-cluster estimator $\hat{\delta}_{\text{followup}} = y_{g=1,t=1,\bullet,\bullet} - y_{g=0,t=1,\bullet,\bullet}$, where $\bullet$ in the subscript refers to averaging over the corresponding index (here subjects and clusters within each treatment group). Likewise, the comparison of the change from baseline scores is expressed by

$$\hat{\delta}_{\text{change}} = \left[y_{g=1,t=1,\bullet,\bullet} - y_{g=1,t=0,\bullet,\bullet}\right] - \left[y_{g=0,t=1,\bullet,,\bullet} - y_{g=0,t=0,\bullet,\bullet}\right].$$

To arrive at a sample size formula, we will assume that the variances of the random effects (σ_{c}^2, $\sigma_{\text{c}\tau}^2$, σ_{s}^2, $\sigma_{\text{s}\tau}^2$) are known, which is a common assumptions at the planning stage.

If a fraction π_o and π_1 of the clusters are on the control and treatment group, respectively, then these estimators have variance

$$\text{var}(\hat{\delta}_{\text{followup}}) = \left(\sigma_{\text{c}}^2 + \sigma_{\text{c}t}^2 + \frac{\sigma_{\text{s}}^2}{n} + \frac{\sigma_{\text{s}\tau}^2}{n}\right)\left[\frac{1}{\pi_0 I} + \frac{1}{\pi_1 I}\right] = \cdot[1 + (n-1)\rho] \cdot \left(\frac{1}{\pi_0} + \frac{1}{\pi_1}\right) \cdot \frac{\sigma^2}{In}, \quad (3)$$

and

$$\text{var}(\hat{\delta}_{\text{change}}) = 2\left(\sigma_{\text{c}t}^2 + \frac{\sigma_{\text{s}\tau}^2}{n}\right)\left[\frac{1}{\pi_0 I} + \frac{1}{\pi_1 I}\right] = 2 \cdot (1-r)) \cdot [1 + (n-1)\rho] \cdot \left(\frac{1}{\pi_0} + \frac{1}{\pi_1}\right) \cdot \frac{\sigma^2}{In}, \quad (4)$$

where

$\sigma^2 = \sigma_{\text{c}}^2 + \sigma_{\text{c}\tau}^2 + \sigma_{\text{s}}^2 + \sigma_{\text{s}\tau}^2$ is the variance of all subjects over all clusters.

Because of the randomization, the expected means in control and intervention arm are equal. Therefore, both estimators have expectation δ, and so has any combination $r\hat{\delta}_{\text{followup}} + (1-r)\hat{\delta}_{\text{change}}$. The latter combination has the smallest variance when

$$r = \frac{\sigma_{\text{c}}^2 + \sigma_{\text{s}}^2/n}{\sigma_{\text{c}}^2 + \sigma_{\text{c}\tau}^2 + \sigma_{\text{s}}^2/n + \sigma_{\text{s}\tau}^2/n} = \frac{n\rho}{1 + (n-1)\rho}\rho_{\text{c}} + \frac{1-\rho}{1 + (n-1)\rho}\rho_{\text{s}}, \quad (5)$$

where

$$\rho = \frac{\sigma_{\text{c}}^2 + \sigma_{\text{c}\tau}^2}{\sigma_{\text{c}}^2 + \sigma_{\text{c}\tau}^2 + \sigma_{\text{s}}^2 + \sigma_{\text{s}\tau}^2} \quad (6)$$

is the correlation at one time point in the same cluster between two different subjects, that is, the ICC. See the Appendix for the derivation of Equations (3)–(6).

The combined estimator reads $\hat{\delta}_{\text{ancova}} = \left[y_{g=1,t=1,\bullet,\bullet} - r y_{g=1,t=0,\bullet,\bullet}\right] - \left[y_{g=0,t=1,\bullet,\bullet} - r y_{g=0,t=0,\bullet,\bullet}\right]$, that is, a change from adjusted baseline. It has variance

$$\text{var}\left(\hat{\delta}_{\text{ancova}}\right) = \left(1 - r^2\right) \cdot [1 + (n-1)\rho] \cdot \left(\frac{1}{\pi_0} + \frac{1}{\pi_1}\right) \cdot \frac{\sigma^2}{In}, \quad (7)$$

which is smaller than the variance of either $\hat{\delta}_{\text{followup}}$ and $\hat{\delta}_{\text{change}}$.

The quantity r is a correlation coefficient: from the first identity in (5) it follows that r is the correlation between a cluster mean at baseline $y_{g,t=0,\bullet,\bullet}$ and at follow-up $y_{g,t=1,\bullet,\bullet}$. Moreover, the second identity of (5) shows that r will be between the cluster autocorrelation ρ_{c} and the subject autocorrelation ρ_{s}, so will be at least larger than the smallest of the two.

In fact, if $n\rho$ is small (small clusters and/or small intracluster correlation), r will be close to the subject autocorrelation, while if $n\rho$ is large (large clusters and/or large intracluster correlation), r will be close to the cluster autocorrelation.

2.3. *Sample size formula*

From the above expressions of the variance, we see that the sample size of a cluster randomized trial analyzed with ANCOVA can be calculated as follows. First calculate the sample size for a *cluster* randomized trial that is analyzed on the follow-up scores (i.e. multiply the sample size according to a t-test

on the follow-up scores with the factor $[1+(n-1)\rho]$). Then multiply this number with the design effect $\left(1-r^2\right)$. For an analysis of the change from baseline scores, the design effect $2\cdot(1-r)$ should be used instead of $\left(1-r^2\right)$.

2.4. *Power calculation*

The power to detect a treatment effect δ given I clusters of size n is

$$power = \Phi_{t,I-2}\left(\frac{\delta}{\sqrt{\text{var}(\hat{\delta})}} - t_{\alpha/2,I-2}\right),$$

where $\hat{\delta}$ is the corresponding estimator (follow-up, change, ANCOVA) and the variance is as given above and $\Phi_{t,n}$ is the cumulative distribution function of the t-distribution with n DOFs.

From the above sample size adjustment factors, a simpler approach to power calculations can be formulated: a cluster trial of $\left(1-r^2\right) I$ clusters analyzed via ANCOVA has the same power as a cluster trial of I clusters analyzed on the follow-up scores. The same holds for a cluster trial of $2(1-r)I$ clusters analyzed via the change from baseline scores.

2.5. *Minimum number of clusters*

From the sample size adjustment factor, the minimum number of clusters for the ANCOVA analysis is a factor r^2 smaller than for an analysis of the follow-up scores. Assuming the clusters are infinitely large, this comes down to

$$\left(1-\rho_c^2\right)\rho I_{\text{indiv}},$$

where $I_{\text{indiv}} = 2\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2\sigma^2/\delta^2$ is the sample size of an individually randomized trial analyzed with a t-test on the follow-up scores. In practice, the number of clusters will need to be larger, because the cluster size is finite.

2.6. *Comparison of the cohort and cross-sectional model*

If we decide to switch from the cohort design to the cross-sectional design, then σ^2 does not change, because σ^2 describes the variance of the subjects over all the clusters at a given time point. However, this variance is differently decomposed in a cross-sectional design. In each cluster, the subjects at baseline are replaced with new subjects at follow-up. Therefore, the variance at subject level, which is $\sigma^2-(\sigma_c^2+\sigma_{c\tau}^2)$, has no time independent component, that is, $\sigma^2_{\text{s,cross-sec}}=0$ and conditional on the cluster there is no correlation between subjects at baseline and follow-up $\rho_{\text{s,cross-sec}}=0$, because at each time point different (independent) subjects are present in the cluster.

Therefore, the correlation $r_{\text{cross-sec}}$ in the cross-sectional design can be expressed in terms of the correlation in the cohort design r as

$$r_{\text{cross-sec}} = \frac{n\rho}{1+(n-1)\rho}\rho_c + 0 = r - \frac{1-\rho}{1+(n-1)\rho}\rho_s \leqslant r$$

where ρ_s is the subject autocorrelation that would be observed in the cohort design.

The ratio of the sample size for the cross-sectional design compared with that of the cohort design is

$$\begin{aligned}\frac{\text{var}\left(\hat{\delta}^{\text{cross-sec}}_{\text{ancova}}\right)}{\text{var}\left(\hat{\delta}^{\text{cohort}}_{\text{ancova}}\right)} &= \frac{\left(1-r^2_{\text{cross}}\right)}{\left(1-r^2\right)} = \frac{(1-r_{\text{cross}})}{(1-r)}\cdot\frac{(1+r_{\text{cross}})}{(1+r)} \\ &= \left[1+(1-\rho)\frac{\rho_s}{[1+(n-1)\rho]\cdot[1-r]}\right]\left[1-(1-\rho)\frac{\rho_s}{[1+(n-1)\rho]\cdot[1+r]}\right], \quad (8)\end{aligned}$$

which expresses that the cross-sectional design requires more sample size than the cohort design (first equality). However, if the cluster size is large ($n\rho$ large) and/or the subject autocorrelation ρ_s small, then the loss in efficiency will be small (last equality).

For an analysis of the change of baseline scores, the efficiency of the cross-sectional design to that of the cohort design is

$$\frac{\text{var}\left(\hat{\delta}^{\text{cross-sec}}_{\text{change}}\right)}{\text{var}\left(\hat{\delta}^{\text{cohort}}_{\text{change}}\right)} = \frac{2\,(1-r_{\text{cross}})}{2\,(1-r)} = \left[1+(1-\rho)\frac{\rho_{\text{s}}}{[1+(n-1)\rho]\cdot[1-r]}\right].$$

2.7. Influence of variability in the estimates of the ICC , r_c, r_s

In the planning stage, plausible values of the intercluster correlation, and subject and cluster autocorrelation are assumed to calculate the sample size. In the analysis stage, these correlations will have to be estimated from the data. To assess the influence of the variability in the estimates of these correlations on the sample size, we performed a simulation study. One thousand datasets were generated according to model (1) with the fixed treatment effect $\delta = 0, 0.2, 0.4$, a fixed time effect equal to 2, and variances σ^2_{c}, $\sigma^2_{\text{c}\tau}$, σ^2_{s}, $\sigma^2_{\text{s}\tau}$ derived from (2) and (6) using $\sigma^2 = 1$ and various preset values of r_{s}, r_{c}, ρ. The number of subjects was $n_s = 20$. The total number of clusters to achieve 80% power at a significance level of 0.05 was estimated by first calculating the uncorrected total number of subjects per treatment group using normal percentiles, that is, $4(z_{1-\alpha/2}+z_{1-\beta})^2\sigma^2/\delta^2$, then multiplying this by the design factor $(1-r^2)[1+(n-1)\rho]$ to correct for the cluster ANCOVA design and dividing this by n to obtain the total number of clusters k. A small sample correction was then performed by multiplying k by $(k+1)/(k-1)$ (see p. 118 in [7]) and the result was rounded up to the nearest even integer, to obtain equal sized intervention and control groups.

In each simulated dataset, cluster means at every time point were calculated and the treatment effect was tested using an ANCOVA analysis. Empirical type I error was estimated as the rejection rate for $\delta = 0$, empirical power was estimated as the rejection rate for $\delta = 0.2$ (small effect), and $\delta = 0.4$ (moderate effect).

For $r_{\text{s}} = 0.8$, $r_{\text{c}} = 0.3$, $\rho = 0.05$, the empirical type I error and empirical power were 0.053 and 0.765 for $\delta = 0.2$, while it was 0.047 and 0.792 for $\delta = 0.4$. Ranging $\rho = 0.01, 0.05, 0.10$; $r_{\text{s}} = 0.8$; $r_{\text{c}} = 0.3$, 0.5; $\delta = 0.2, 0.4$ yielded similar results; the type I error did not exceed 0.061 and the power was at most 0.057 lower than predicted. Thus, the influence of the variability in the estimates is rather limited.

3. Example

The two-step procedure in Section 2.4 can be used to calculate the sample size of a cluster randomized trial with pretest–posttest design. Consider, for example, the Community occupational Therapy in Dementia (CoTiD) trial that compares two implementation strategies of occupational therapy for dementia patients and their caregivers (Netherlands Organization for Health Research and Development ZonMw, grant 80-82315-98-090010). One primary outcome measure is adherence of the occupational therapists to the COTiD program. Assuming a moderate effect size $es = \delta/\sigma = 0.5$ [8], then a sample size of 62 per arm results, which is uncorrected for clustering (significance level 0.05, power 0.80). Two occupational therapists (=subjects) are expected per health care institute (=cluster) with an institute-ICC of $\rho = 0.05$, which means that the design effect because of clustering is $[1+(n-1)\rho] = 1.05$. The consistency of an occupational therapist is likely to be strong, that is, a high therapist autocorrelation $\rho_{\text{s}} = 0.7$, while the agreement of the therapists in one institute may be smaller, that is, an institute autocorrelation $\rho_{\text{c}} = 0.5$. This means that the design effect because of the pretest–posttest is $1-r^2 = 0.54$, that is, 46% reduction. Therefore, the sample size is 35 subjects or 17 clusters per arm.

4. Sample size comparison with other designs

Below, we will compare our sample size formula for the cluster randomized pretest–posttest design with sample size formulas of other designs. Two comparisons are most important: that to the individually randomized pretest–posttest design (see Section 4.1), which shows that our formula is a generalization of the ANCOVA sample size for individually randomized trials. Second, that to other sample size formulas for ANCOVA of the cluster randomized pretest–posttest design, which illustrates the differences between our formula and existing ones for this design (see Section 4.3).

4.1. *Individually randomized pretest–posttest designs*

Our sample size formula generalizes the sample size formula in [9] for ANCOVA of individually randomized trials. Actually, for cluster size $n = 1$

$$r = \rho\rho_c + (1-\rho)\rho_s = \frac{\sigma_c^2 + \sigma_s^2}{\sigma_c^2 + \sigma_s^2 + \sigma_{ct}^2 + \sigma_{st}^2}.$$

where $\sigma_c^2 + \sigma_s^2$ is the time-invariant subject variance and $\sigma_{ct}^2 + \sigma_{st}^2$ is the time varying variance at subject level in model (1) for $n = 1$. Indeed, no nesting of subjects within clusters is present, which can be phrased as $\sigma_s^2 = \sigma_{st}^2 = 0$. Thus, for $n = 1$, r is the correlation between baseline and follow-up for a subject and our sample size formula comes down to that in [9].

Another approach to analyze the pretest–posttest design is to consider the follow up measurement and the baseline measurement as repeated measures, under the restriction of equal baseline means because of the randomization (constrained longitudinal data model). Lu *et al.* calculated the sample size for constrained longitudinal data models with one or more post-baseline measurements under a monotone missing data pattern of the post-baseline measurements [10]. Their variance formula for the pretest–posttest design gives the same sample size calculation as in [9] in the absence of missing data ($n_{j1} = n$). In the presence of missing post-baseline measurements, their effective sample size consists of only those patients that have complete data at baseline and follow up.

4.2. *Cluster randomized pretest–posttest designs: change from baseline analysis*

Feldman and McKinlay developed the sample size for change from baseline analysis in terms of the variances at cluster and subject level (Equation (19) in [2]). Therefore, they did not express it using a sample size adjustment factor (design effect) or the autocorrelation r between cluster means. By comparing (4) to (7), we see that a change from baseline analysis is a factor $2/(1 + r)$ less efficient than an analysis of covariance. In another parametrization, Preisser *et al.* arrived at the same formula for a change from baseline analysis for the pretest–posttest cross-sectional design (see formula (7) and the comments following on p. 1247 in [11]).

4.3. *Cluster randomized pretest–posttest designs: analysis of covariance*

The sample size requirements of an ANCOVA of cluster randomized trials were also investigated by Bloom *et al*. [4] (see their Equation (4)), by Raudenbush [3] (see his Equation (12)), and by Moerbeek [12] (see her Equation (7)). Their approach is more general in that they consider an arbitrary covariate, not per se a baseline measurement of the outcome. Common to their approaches is that they express the sample size in terms of the variance at cluster and at subject level. The influence of adjusting for covariate(s) is then described as a reduction of these two variances. Bloom *et al*. [4] express these reductions using R_c^2 and R_i^2, the proportion of variance explained by the covariate(s) at cluster level and individual level, respectively. To describe these reductions, Moerbeek [12] uses ρ_W and ρ_B, the within-cluster and between-cluster residual correlations between the posttest outcome and the covariate as defined in [13], while Raudenbush [3] directly starts from the unexplained variances that remain at cluster and subject level after adjusting.

The difference between the sample size formulas in [3, 4, 12] and that in Section 2.3 is mainly a matter of parametrization. First, because the sample size formulas in [3, 4, 12] describe the effect of including a covariate by changes of the variance of cluster and subject level separately, while Section 2.3 describes a change of the total variance (see (7)). Second, because the definitions of the parameters used in the sample size formulas differ. After appropriate translations of the parametrizations, they would be expected to produce similar sample sizes, because all use the same analysis model (including the baseline outcome as covariate) and equivalent estimation methods (maximum likelihood estimators vs minimum variance estimators in balanced designs).

Because their formulas are expressed as a sum of the reduced variances at cluster level and at subject level, they do not arrive at a multiplicative form, that is, an unadjusted sample size times a design effect. Consequently, they need direct specification of the variances at subject and cluster level, which makes their formulation less easy to discuss for input with trialists and applied researchers.

4.4. Cluster randomized posttest designs

Because the posttest is the simplest cluster randomized design, many aspects influencing the sample size of the posttest design have been investigated, including:

(a) What interpretation of the estimate is intended: population-averaged or cluster-specific [11]
(b) The type of randomization: completely randomized , matched or stratified [1, 14]
(c) How the clustering of the outcome is parametrized: via the intracluster correlation [1] or coefficient of variation [15] or variance components [16].
(d) How uncertainty in the ICC is dealt with [17, 18]
(e) How variation of the cluster sizes is accounted for [19, 20].

Each of these aspects for example the uncertainty in the ICC is a subject for further research in case of the analysis of covariance for the pretest–posttest design. However, because the pretest–posttest design has more measurements than the posttest-only design, it would be expected that for example the sample size for ANCOVA accounting for uncertainty in the ICC would still be less than that of the posttest accounting for uncertainty in the ICC. The same would be expected when accounting for any of the other aspects above.

5. Discussion

The sample size formula we derived shows that (compared with an analysis of the follow-up scores) an analysis of covariance using the baseline outcome as covariate reduces the sample size by a factor r^2, where r is the autocorrelation of a cluster mean between baseline and follow-up. The correlation r will lie between the cluster autocorrelation ρ_c and the subject auto correlation ρ_s. In particular, r will be at least as large as the smaller of ρ_c and ρ_s. Thus, if both are at least 0.3, 0.5, or 0.7, then the reduction in sample size will be at least around 10%, 25% or 50%, respectively. Such magnitudes are in line with empirically found reductions [4].

The correlation r of a cluster mean between baseline and follow-up will be more towards the cluster autocorrelation ρ_c when clusters are large and more towards the subject autocorrelation ρ_s when they are small. Intuitively, this can be understood as follows. If the clusters are small, the design effect of a cluster randomized trial $(1+(n-1)\rho))$ is small and the sample will (statistically) behave more like individuals, so that the correlation of subjects between baseline and follow-up predominates. If the clusters are large, the design effect is large, that is, the sample will behave more like separate clusters and the association/connection of individual subjects (within a cluster) at baseline and follow-up is washed out.

For an analysis of change from baseline scores, the sample size differs by a factor $2(1-r)$ from that of an analysis of follow-up scores. Therefore, analyzing the change from baseline scores will only be more efficient, when $r > 0.5$. Even then, however, analysis of covariance is more efficient.

The cohort design is more efficient than the cross-sectional design in case of complete follow-up. However, practical considerations may still favour the cross-sectional design. Substantial loss to follow-up may necessitate costs for oversampling and/or aggressive follow-up, which may outweigh the efficiency gain. Moreover, for large cluster size and/or small subject autocorrelation and/or large intracluster correlation, the efficiency gain of the cohort design over the cross-sectional design is small as can be seen from formula (8).

In addition to the ICC , a sample size calculation for an ANCOVA (and the change from baseline analysis) needs the subject and cluster autocorrelation as input. These can be obtained from previous studies, either indirectly from published variance components or directly from published autocorrelations [2]. In many cases, such results have not been made available. Because cluster and subject autocorrelations have an (clinically) interpretable meaning, a range of plausible values can then be obtained in consultation with clinical experts. The subject autocorrelation is expected high, if the measurement is reproducible. Also, the cluster autocorrelation can be high. Consider for example that the clusters are therapy groups. If the style of the therapist and/or the interaction of subjects within the therapy group has a substantial effect on the subjects' outcomes, then the cluster autocorrelation will be high. This has also been empirically confirmed. For example, Feldman and McKinlay [2] in their Table IV find that the cluster autocorrelation for total cholesterol is high, suggesting that the management by the general practitioner is of important influence.

Another argument for high cluster autocorrelations is provided by Bloom *et al.* [4]. They argued that correlations of aggregates at cluster level may be at least as high as correlations at subject level

(note 4 in [4]), partly because they are expected to be more reliable (p. 31 and note 4). This had also been empirically observed by Klar and Darlington [21, p. 2354].

When the autocorrelations are based on expert judgement, it is sensible to investigate how sensitive the power is to the assumed autocorrelations by varying ρ_s and ρ_c over a range of plausible values. This holds likewise when the autocorrelations are based on previous studies.

The sample size requirements of an ANCOVA of cluster randomized trials were also investigated by Bloom *et al.* [4], by Raudenbush [3] and by Moerbeek [12]. Their approach is more general in that they consider an arbitrary covariate, not per se a baseline measurement of the outcome, but they do not arrive at a design effect (sample size multiplier) and they need direct specification of the variances at subject and cluster level, which makes their formulation less accessible to trialists and applied researchers.

Our parametrization of sample size formula has several advantages. The sample size formula takes a simple multiplicative form: first calculate the sample size for a cluster randomized trial as usual and then multiply this number by the design factor $(1–r^2)$ for the ANCOVA. Second, the efficiency gain of ANCOVA can be directly recognized as r^2.

Furthermore, the autocorrelation of cluster means r is expressed in terms of Pearson correlations (the ICC , the subject autocorrelation, and the cluster autocorrelation), which have an interpretable meaning to clinical experts. Therefore, if no prior estimates are available, consultation with clinical experts can motivate (a range of) plausible values.

Finally, the (relatively simple) expression of r in terms of ρ, ρ_c, ρ_s and n gives insights on how the gain in efficiency r^2 depends on those factors.

Lastly, one caveat is in order, when applying our formula. The ICC used is the unadjusted ICC: if the ICC already accounts for the outcome at baseline, the usual sample size applies: $[1 + (n - 1)\rho_{adj}]N$ with ρ_{adj} the adjusted ICC and N the sample size according to a t-test on the follow-up scores. Conversely, if such adjusted ICCs are not known, our sample size formula gives a way to input subject matter knowledge to still estimate the reduction in sample size using ANCOVA with baseline scores.

Appendix

If $I_0 = \pi_0 I$, $I_1 = \pi_1 I$ are the number of clusters in the control and intervention group respectively, then

$$\hat{\delta}_{\text{followup}} = y_{g=1,t=1,\bullet,\bullet} - y_{g=0,t=1,\bullet,\bullet} = \delta + \frac{1}{I_1}\sum_{i=I_o+1}^{I_0+I_1}\left[c_i + (c\tau)_{i,t=1} + \frac{1}{n}\sum_{k=1}^{n}(s_{ik} + (s\tau)_{i,t=1})\right]$$
$$- \frac{1}{I_0}\sum_{i=1}^{I_0}\left[c_i + (c\tau)_{i,t=1} + \frac{1}{n}\sum_{k=1}^{n}(s_{ik} + (s\tau)_{i,t=1})\right]$$

and

$$\hat{\delta}_{\text{change}} = \left[y_{g=1,t=1,\bullet,\bullet} - y_{g=1,t=0,\bullet,\bullet}\right] - \left[y_{g=0,t=1,\bullet,\bullet} - y_{g=0,t=0,\bullet,\bullet}\right] = \delta$$
$$+ \frac{1}{I_1}\sum_{i=I_o+1}^{I_0+I_1}\left\{(c\tau)_{i,t=1} - (c\tau)_{i,t=0} + \frac{1}{n}\sum_{k=1}^{n}\left[(s\tau)_{i,t=1} - (s\tau)_{i,t=0}\right]\right\}$$
$$- \frac{1}{I_0}\sum_{i=1}^{I_0}\left\{(c\tau)_{i,t=1} - (c\tau)_{i,t=0} + \frac{1}{n}\sum_{k=1}^{n}\left[(s\tau)_{i,t=1} - (s\tau)_{i,t=0}\right]\right\}$$

(Because of the randomization, the means in control and intervention arm at baseline have the same expectation).

From these expressions it follows that

$$\text{var}(\hat{\delta}_{\text{posttest}}) = \left(\sigma_c^2 + \sigma_{ct}^2 + \frac{\sigma_s^2}{n} + \frac{\sigma_{s\tau}^2}{n}\right)\left[\frac{1}{\pi_o I} + \frac{1}{\pi_1 I}\right],$$

$$\text{var}(\hat{\delta}_{\text{change}}) = 2\left(\sigma_{ct}^2 + \frac{\sigma_{s\tau}^2}{n}\right)\left[\frac{1}{\pi_o I} + \frac{1}{\pi_1 I}\right]$$

and

$$\operatorname{cov ar}\left(\hat{\delta}_{\text{followup}}, \hat{\delta}_{\text{change}}\right) = \operatorname{var}\left(\frac{1}{I_1}\sum_{i=I_o+1}^{I_0+I_1}\left[(c\tau)_{i,t=1} + \frac{1}{n}\sum_{k=1}^{n}(s\tau)_{i,t=1}\right] - \frac{1}{I_0}\sum_{i=1}^{I_0}\left[(c\tau)_{i,t=1} + \frac{1}{n}\sum_{k=1}^{n}(s\tau)_{i,t=1}\right]\right)$$
$$= \left(\sigma^2_{c\tau} + \frac{\sigma^2_{s\tau}}{n}\right)\left[\frac{1}{\pi_o I} + \frac{1}{\pi_1 I}\right].$$

Denote

$$\alpha = \left(\sigma^2_{c\tau} + \frac{\sigma^2_{s\tau}}{n}\right)\left[\frac{1}{\pi_o I} + \frac{1}{\pi_1 I}\right],\ \beta = \left(\sigma^2_{c} + \frac{\sigma^2_{s}}{n}\right)\left[\frac{1}{\pi_o I} + \frac{1}{\pi_1 I}\right].$$

(The part that varies within clusters between time points, the part that varies between clusters at a fixed time point).

Then

$$\alpha + \beta = \left[\frac{1}{\pi_o I} + \frac{1}{\pi_1 I}\right]\cdot\left(\sigma^2_{c} + \sigma^2_{c\tau} + \frac{\sigma^2_{s}}{n} + \frac{\sigma^2_{s\tau}}{n}\right) = \left[\frac{1}{\pi_o I} + \frac{1}{\pi_1 I}\right]\cdot\frac{1}{n}\cdot[1 + (n-1)\rho]\cdot\sigma^2$$

In terms of

$$r = \frac{\beta}{\alpha + \beta}$$

we have

$$\operatorname{var}(\hat{\delta}_{\text{posttest}}) = \alpha + \beta = \left[\frac{1}{\pi_o I} + \frac{1}{\pi_1 I}\right]\cdot\frac{1}{n}\cdot[1 + (n-1)\rho]\cdot\sigma^2$$
$$\operatorname{var}(\hat{\delta}_{\text{change}}) = 2\alpha = 2(1-r)(\alpha+\beta) = 2(1-r)\left[\frac{1}{\pi_o I} + \frac{1}{\pi_1 I}\right]\cdot\frac{1}{n}\cdot[1 + (n-1)\rho]\cdot\sigma^2$$
$$\operatorname{cov ar}\left(\hat{\delta}_{\text{posttest}}, \hat{\delta}_{\text{change}}\right) = \alpha = (1-r)(\alpha+\beta) = (1-r)\left[\frac{1}{\pi_o I} + \frac{1}{\pi_1 I}\right]\cdot\frac{1}{n}\cdot[1 + (n-1)\rho]\cdot\sigma^2.$$

The estimator $x\hat{\delta}_{\text{change}} + z\hat{\delta}_{\text{posttest}}$ with $x + z = 1$, has minimum variance (use Lagrange multipliers) for $x = r$, $z = 1 - r$, where $r = \frac{\beta}{\alpha+\beta}$.

Proof

$\operatorname{var}(x\hat{\delta}_{\text{change}} + z\hat{\delta}_{\text{posttest}}) = x^2\,(2\alpha) + z^2\,(\alpha+\beta) + 2xz\,(\alpha)$, and the Lagrange multiplier equations read:

$$0 = \lambda + \frac{\partial Var(xX + zZ)}{\partial x} = \lambda + 2x\,(2\alpha) + 2z\,(\alpha) \tag{1}$$

$$0 = \lambda + \frac{\partial Var(xX + zZ)}{\partial z} = \lambda + 2z\,(\alpha+\beta) + 2x\,(\alpha) \tag{2}$$

Subtracting (1) from (2) gives

$$0 = 2z\,(\beta) + 2x(-\alpha), \text{ that is, } \frac{x}{z} = \frac{\beta}{\alpha}.$$

Because $x + z = 1$, we have $x = \frac{\beta}{\alpha+\beta}$ and $z = \frac{\alpha}{\alpha+\beta}$. □

The variance of this minimum variance estimator is:
(using the relation between r, α, β)

$$\operatorname{var}(x\hat{\delta}_{\text{change}} + z\hat{\delta}_{\text{posttest}}) = r^2 2\alpha + 2r(1-r)\alpha + (1-r)^2(\alpha+\beta) = (1-r)(1+r)(\alpha+\beta)$$
$$= \left(1-r^2\right)\cdot[1+(n-1)\rho]\cdot\left(\frac{1}{\pi_0}+\frac{1}{\pi_1}\right)\cdot\frac{\sigma^2}{In}.$$

Using

$$\sigma_{\text{c}}^2 = \rho_{\text{c}}\rho\sigma^2, \quad \sigma_{\text{ct}}^2 = (1-\rho_{\text{c}})\rho\sigma^2, \quad \sigma_{\text{s}}^2 = \rho_{\text{s}}(1-\rho)\sigma^2, \quad \sigma_{\text{st}}^2 = (1-\rho_{\text{s}})(1-\rho)\sigma^2$$

the correlation r can be rewritten as

$$r = \frac{\sigma_{\text{c}}^2 + \sigma_{\text{s}}^2/n}{\sigma_{\text{c}}^2 + \sigma_{\text{c}\tau}^2 + \sigma_{\text{s}}^2/n + \sigma_{\text{s}\tau}^2/n} = \frac{\rho_{\text{c}}\rho + \rho_{\text{s}}(1-\rho)/n}{\rho_{\text{c}}\rho + (1-\rho_{\text{c}})\rho + \rho_{\text{s}}(1-\rho)/n + (1-\rho_{\text{s}})(1-\rho)/n}$$
$$= \frac{\rho_{\text{c}}\rho + \rho_{\text{s}}(1-\rho)/n}{\rho + 1/n - \rho/n} = \frac{n\rho}{n\rho+(1-\rho)}\rho_{\text{c}} + \frac{1-\rho}{n\rho+(1-\rho)}\rho_{\text{s}}.$$

Acknowledgement

The Netherlands Organization for Health Research (ZonMW) supported the research of Steven Teerenstra for this subject under grant number 9930059.

References

1. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold: London, 2000.
2. Feldman HA, McKinlay SM. Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. *Statistics in Medicine* 1994; **13**:61–78. DOI: 10.1002/sim.4780130108.
3. Raudenbush SW. Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods* 1997; **2**:173–185.
4. Bloom HS, Richburg-Hayes L, Black AR. Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Evaluation and Policy Analysis* 2007; **29**:30–59. DOI: 10.3102/0162373707299550.
5. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *American Journal of Public Health* 2004; **94**:423–432. DOI: 10.2105/AJPH.94.3.423.
6. Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review* 2003; **27**:79–103. DOI: 10.1177/0193841X02239019.
7. Steel RGD, Torrie JH. *Principles and Procedures of Statistics: A Biometrical Approach*, 2nd edition. McGraw-Hill: New York, 1980.
8. Cohen J. A power primer. *Psychological Bulletin* 1992; **112**:155–59.
9. Borm GF, Fransen J, Lemmens WAJG. A simple sample size formula for analysis of covariance in randomized clinical trials. *Journal of Clinical Epidemiology* 2007; **60**:1234–1238. DOI: 10.1016/j.jclinepi.2007.02.006.
10. Lu K, Mehrotra DV, Liu G. Sample size determination for constrained longitudinal data analysis. *Statistics in Medicine* 2009; **28**:679–699. DOI: 10.1002/sim.3507.
11. Preisser JS, Young ML, Zaccaro DJ, Wolfson M. An integrated population-averaged approach to the design,analysis and sample size determination of cluster-unit trials. *Statistics in Medicine* 2003; **22**:1235–1254. DOI: 10.1002/sim.1379.
12. Moerbeek M. Power and money in cluster randomized trials: When is it worth measuring a covariate? *Statistics in Medicine* 2006; **25**:2607–2617. DOI: 10.1002/sim.2297.
13. Snijders TAB, Bosker RJ. *Multilevel Analysis. An introductio to Basic and Advanced Multilevel Modeling*. Sage Publications: London, 1999.
14. Campbell MK, Thomson S, Ramsay CR, MacLennan GS, Grimshaw JM. Sample size calculator for cluster randomized trials. *Computers in Biology and Medicine* 2004; **34**:113–25. DOI: 10.1016/S0010-4825(03)00039-8.
15. Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology* 1999; **28**:319–26. DOI: 10.1093/ije/28.2.319.
16. Hsieh FY. Sample size formulae for intervention studies with the cluster as unit of randomization. *Statistics in Medicine* 1988; **7**:1195–201. DOI: 10.1002/sim.4780071113.
17. Rotondi MA, Donner A. Sample size estimation in cluster randomized trials educational trials: an empirical Bayes approach. *Journal of Educational and Behavioral Statistics* 2009; **34**:229–237. DOI: 10.3102/1076998609332756.
18. Turner RM, Prevost AT, Thompson SG. Allowing for imprecision of the intracluster correlation coefficient in design of cluster randomized trials. *Statistics in Medicine* 2004; **23**:1195–214. DOI: 10.1002/sim.1721.
19. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology* 2006; **35**:1292–300. DOI: 10.1093/ije/dyl129.
20. van Breukelen GJP, Candel M J J M, Berger MPF. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine* 2007; **26**:2589–2603. DOI: 10.1002/sim.2740.
21. Klar N, Darlington G. Methods for modelling change in cluster randomization trials. *Statistics in Medicine* 2004; **23**:2341–2357. DOI: 10.1002/sim.1858.