Power, sample size and sampling costs for clustered data

K. Tokola[a,*], D. Larocque[b], J. Nevalainen[c], H. Oja[a]

[a] *Tampere School of Public Health, University of Tampere, Tampere FI-33014, Finland*
[b] *HEC Montréal, Quebec, Canada*
[c] *Statistics/Department of Social Research, University of Turku, Finland*

Abstract

The data collected in epidemiological or clinical studies are frequently clustered. In such settings, appropriate variance adjustments must be made in order to estimate the sufficient sample size correctly. This paper works through the sample size calculations for clustered data. Importantly, our explicit variance expressions also enable us to optimize the design with respect to the number of clusters and number of subjects; the objective could be either to maximize the power or to minimize the costs with given costs on the clusters and on the individuals. In our approach, units on different levels and treatment groups can have different costs, but the members of the same cluster are assumed to belong to the same treatment group. Design considerations in the health coaching project TERVA are used as motivating examples. R-functions for carrying out the presented computations are provided.

Keywords: Sample size, Clustered data, Cost optimization

1. Introduction

This paper is motivated by the health coaching project (TERVA), which is a clinical study aiming to demonstrate the impact of a health coach program on the general mental and physical health, risk factor behavior modification and ability to self-manage their conditions in participants with heart failure, coronary heart disease or type 2 diabetes. Subjects are randomly allocated to two groups, treatment or control, at a 2 : 1 ratio. By second stage random

[*] corresponding author
Email address: kari.tokola@uta.fi (K. Tokola)

allocation, each subject in the treatment arm is subjected to a 12 months intervention program supervised by a health coach, whereas the subjects in the control arm are merely followed up for the same period of time. Because the coaches may have different effects on the success of the intervention, results of the subjects allocated to the same coach tend to be positively correlated. One way to view this particular design is to see the coaches as clusters, so that the treatment group consists of clusters of equal size, but no clustering is present in the control group (or, in other words, the cluster size is one). As the cost of the study depends not only on the total number of subjects, but also on the number of coaches (clusters), it is also important to consider design cost minimization strategies.

Analysis of clustered data has a central role in biomedical research, because the collected data often involves clustered units such as patients of the same hospital, or siblings. For this reason, sample size and power calculations for clustered data has been an important field of research in applied statistics until recently (Donner et al., 1981; Gangnon and Kosorok, 2004; Heo and Leon, 2008, 2009; Hoover, 2002; Eldridge et al., 2001, 2006; Kerry and Bland, 2001; Liu et al., 2002; Rotondi and Donner, 2009; Tu et al., 2004). Previous work involves various type of responses, designs and models of differing complexity (*e.g.* repeated measurements, cluster vs. individual randomization, multilevel data) and even software (Bauer and Sterba, 2008; Campbell et al., 2004; Hayes and Bennett, 1999; Lee and Thompson, 2005; Roberts and Roberts, 2005; Rotondi and Donner, 2009).

Sample size and economical issues of cluster randomized trials have been reviewed by Klar and Donner (2001) and Campbell et al. (2007). Economical issues are indeed one the most fascinating issues in planning the collection of clustered data. For instance, as excellently outlined by Flynn et al. (2002), staff training, data collection costs, travel costs and management costs are attributed differently to clusters and subjects within the clusters, or even to different treatment groups. In a well planned trial those costs should be minimized while the trial meets a chosen type I and II error rates (or size and power). Alternatively, the trial may be given a fixed budget and the trial should collect data efficiently, that is, with maximal power. Interest towards this type of approach has been expressed under particular settings (Headrick and Zumbo, 2005; Liu, 2003; McKinlay, 1994; Moerbeek, 2006; Moerbeek and Wong, 2008; Raudenbush, 1997; Raudenbush and Liu, 2000), but a general treatment seems to be missing from the literature.

The purpose of the present paper is three-fold. First, we outline how to compute sample size and power with an adjusted t-test for clustered data. Explicit but general variance expressions are worked out. This allows for allocation of treatments on the cluster or subject level, or even their mixtures. Second, we demonstrate how to optimize designs either to minimize costs or to maximize power, with respect to the number of clusters and subjects within the clusters. In the optimization problem we assume that the members of the same cluster also receive the same treatment. Importantly, however, costs can be different from treatment group to another, both on the cluster level and on the subject level. Third, we offer interested readers the option to download R-functions for performing these calculations.

The paper is organized as follows. Section 2 introduces the notation and the assumed model. Section 3 derives the limiting distributions of the test statistic, which are then used in section 4 to derive sample size and power formulae applicable for a variety clustered designs. Cost minimization and power maximization are discussed in section 5. The paper ends with demonstrations within the design of the TERVA project along with concluding remarks.

2. Description of the model

Consider the comparison of two groups (control vs. treatment) with clustered data without particular restrictions on the cluster structure yet. The data set with n clusters will be denoted by

$$X = (X_1, ..., X_n)$$

and observations within each cluster will be denoted by

$$X_i = (X_{i1}, ..., X_{im_i}), \quad i = 1, ..., n$$

where m_i is the number of observations in the i^{th} cluster. To distinguish between the groups, write G_{ij} for treatment indicator taking values 0 or 1 depending on whether subject j in cluster i is in the control or treatment group, respectively. Thus, the total number of subjects is $N = \sum_{i=1}^{n} m_i$, the number of subjects in the treatment group is $N_1 = \sum_{i=1}^{n} \sum_{j=1}^{m_i} G_{ij}$, and the number of subjects in the control group is $N_0 = \sum_{i=1}^{n} \sum_{j=1}^{m_i} (1 - G_{ij})$.

Throughout the paper we use the following distributional assumptions.

Assumption 1. (Distributional assumptions) The random variables X_{ij} are given by

$$X_{ij} = \mu + \Delta G_{ij} + \epsilon_{ij},\ i = 1, .., n;\ j = 1, ..., m_i$$

where

$$E(\epsilon_{ij}) = 0, \quad \text{Var}(\epsilon_{ij}) = \sigma^2, \quad \text{and} \quad E(\epsilon_{ij}^{\nu+2}) < M \quad \text{for some } \nu > 0$$

and

$$\text{Cov}(\epsilon_{ij}, \epsilon_{ij'}) = \rho\sigma^2, \quad \text{for } j \neq j'$$

and

$$\epsilon_{ij} \text{ and } \epsilon_{i'j'} \text{ are independent for } i \neq i'.$$

Here ρ denotes the *intra-cluster correlation* and σ^2 denotes the variation due to random error. Note the flexibility of the model: no normality of the random errors is assumed, nor do we assume normality of random effects as one would commonly do with mixed models. The assumptions are only on the first and second moments of the univariate and bivariate distributions. Also, the assumption that $E(\epsilon_{ij}^{\nu+2})$ is uniformly bounded for some $\nu > 0$ is needed for the asymptotics. This permits the application of the results to follow to a variety of distributions.

3. Test statistic and limiting distribution

The interest is to test the null hypothesis $H_0 : \Delta = 0$ vs. the alternative hypothesis $H_1 : \Delta \neq 0$. Write $g_{ij} = \frac{G_{ij}}{N_1} - \frac{1-G_{ij}}{N_0}$. The estimate of the treatment effect, *i.e* difference of the means between the two groups, can now be written as

$$\hat{\Delta} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} g_{ij}X_{ij}$$

with $E(\hat{\Delta}) = \Delta$ and

$$\begin{aligned}
\text{Var}(\hat{\Delta}) &= \sum_{i=1}^{n}\sum_{j=1}^{m_i} g_{ij}^2 \text{Var}(X_{ij}) + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{j'\neq j} g_{ij}g_{ij'}\text{Cov}(X_{ij}, X_{ij'}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m_i} g_{ij}^2\sigma^2 + \sum_{i=1}^{n}\left(\sum_{j=1}^{m_i} g_{ij}\right)^2 \rho\sigma^2 - \sum_{i=1}^{n}\sum_{j=1}^{m_i} g_{ij}^2\rho\sigma^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m_i} g_{ij}^2\sigma^2 + \sum_{i=1}^{n}\left[\left(\sum_{j=1}^{m_i} g_{ij}\right)^2 - \sum_{j=1}^{m_i} g_{ij}^2\right]\rho\sigma^2.
\end{aligned}$$

Here the first part of the variance demonstrates the variance not depending on the design (cluster structure), and second part of the variance is the result of clustering.

For the limiting distribution of $\hat{\Delta}$ we need the following assumption on the design.

Assumption 2. (Design assumption I)
(i) $m_i \leq m$ for some m.
(ii) There exists a constant λ, $0 < \lambda < 1$ such that $N_1/N \to \lambda$.
(iii) There exist constants d_0 and d_1 such that

$$N \sum_{i=1}^{n} \sum_{j=1}^{m_i} g_{ij}^2 \to d_0 \quad \text{and} \quad N \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{m_i} g_{ij} \right)^2 - \sum_{j=1}^{m_i} g_{ij}^2 \right] \to d_1.$$

Then we get the following.

Lemma 1. *Under assumptions 1 and 2*

$$\sqrt{N}\hat{\Delta} \to_d N\left(\Delta, (d_0 + \rho d_1)\sigma^2\right)$$

as $n \to \infty$.

Proof It is not a restriction to assume that $\mu = 0$ and $\Delta = 0$. We use Corollary 1.9.2.A in Serfling (1980). Write

$$Y_i = n \cdot \sum_{j=1}^{m_i} g_{ij} X_{ij}, \quad i = 1, ..., n,$$

so that

$$\hat{\Delta} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

The Y_i's are independent, $E(Y_i) = 0$, and

$$Var(Y_i) = \sigma_i^2 = n^2 \left[(\sum_{j=1}^{m_i} g_{ij}^2)(1-\rho)\sigma^2 + (\sum_{j=1}^{m_i} g_{ij})^2 \rho\sigma^2 \right].$$

Minkowski's inequality and assumption 1 then gives

$$E\left(|Y_i|^{2+\nu}\right) \leq n^{2+\nu} \left(\sum_{j=1}^{m_i} |g_{ij}| \right)^{2+\nu} M.$$

As $\sum_j |g_{ij}| \leq m_i / \min(N_0, N_1)$,

$$\sum_{i=1}^{n} E\left(|Y_i|^{2+\nu}\right) = \left(\frac{n}{\min(N_0, N_1)}\right)^{2+\nu} N m^{1+\nu}.$$

As $(N/n^2) \sum_i \sigma_i^2 \to (d_0 + \rho d_1)\sigma^2$, we obtain that

$$\frac{\sum_{i=1}^{n} E\left(|Y_i|^{2+\nu}\right)}{(\sum_{i=1}^{n} \sigma_i^2)^{2+\nu}} \to 0 \quad \text{as } n \to \infty$$

and the proof follows from Corollary 1.9.2.A in Serfling (1980).

Note that, under the null hypothesis $H_0 : \;\Delta = 0$, the expected value $E(\hat{\Delta}) = 0$ and a consistent estimate of $\mathrm{Var}(\hat{\Delta})$ under the null hypothesis is

$$\widehat{\mathrm{Var}}(\hat{\Delta}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{m_i} g_{ij}(X_{ij} - \bar{X})\right)^2.$$

where $\bar{X} = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij}$. This is because $E(\hat{\Delta}) = 0$ and

$$\begin{aligned}
\mathrm{Var}(\hat{\Delta}) = E(\hat{\Delta}^2) &= E\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{i'=1}^{n}\sum_{j'=1}^{m_{i'}} g_{ij}X_{ij}g_{i'j'}X_{i'j'}\right) \\
&= E\left(\sum_{i=1}^{n}\left(\sum_{j=1}^{m_i} g_{ij}(X_{ij} - \mu)\right)^2\right)
\end{aligned}$$

where $\mu = E(X_{ij})$. Thus

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{m_i} g_{ij}(X_{ij} - \hat{\mu})\right)^2$$

is consistent to $\mathrm{Var}(\hat{\Delta})$ for any consistent estimate $\hat{\mu}$ of μ and $\bar{X}$ is such an estimate. The test statistic for testing the null hypothesis $H_0 : \;\Delta = 0$ is then the standardized treatment effect estimate

$$T = \frac{\hat{\Delta}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\Delta})}}.$$

By Slutsky's lemma, $T \to_d N(0,1)$ and $T^2 \to_d \chi_1^2$ under the null hypothesis.

Consider next the limiting distribution under a sequence of alternatives $H_n : \Delta = \frac{\Delta_0}{\sqrt{N}}$. If $X_{ij}^* = X_{ij} + G_{ij}\frac{\Delta_0}{\sqrt{N}}$, $i = 1, ..., n; j = 1, ..., m_i$, and $\hat{\Delta}^*$ is calculated from the X_{ij}^* observations, then

$$\sqrt{N}\hat{\Delta}^* = \sqrt{N}\hat{\Delta} + \Delta_0,$$

and, using Lemma 1, the following result follows.

Theorem 1. *Under assumptions 1 and 2 and under the sequence of alternative hypotheses $H_n : \Delta = \frac{\Delta_0}{\sqrt{N}}$,*

$$T^2 \to_d \chi_1^2(\delta^2) \text{ , where } \delta^2 = \frac{\Delta_0^2/\sigma^2}{d_0 + \rho d_1}.$$

Remark 1. The estimate $\hat{\Delta} = \sum_i \sum_j g_{ij} X_{ij}$ above uses the weights $g_{ij} = \frac{G_{ij}}{N_1} - \frac{1-G_{ij}}{N_0}$. Note that the above results hold true for any weights g_{ij} which are (i) positive for the treatment group and (ii) negative for the control group, and (iii) the weights sum up to 1 for the treatment group, and (iv) sum up to -1 for the control group. If the $N \times N$ covariance matrix of the observation vector X, say V, were known, one could find the optimal weights using the Lagrange multiplier technique. Let g be the N-vector of weights, and G the N-vector of the treatment indicator. The problem is to minimize the variance $g^T V g$ under the constraints $g^T 1_N = 0$ and $g^T G = 1$ (Lemponen et al., 2011). If a random effects model with multivariate normality of X is used to analyze the data, then the second row of the matrix $(Z^T V^{-1} Z)^{-1} Z^T V^{-1} X$ with $Z = (1_N, G)$ gives $\hat{\Delta}$, and therefore the resulting (optimal) g^T is the second row of $(Z^T V^{-1} Z)^{-1} Z^T V^{-1}$.

This general result in Theorem 1 can be applied to calculate sample sizes under various clustering designs. So far the results have been applicable for various clustered designs, but in the following we focus on the designs used in our motivating TERVA project example. The assumptions then are as follows.

Assumption 3. (Design assumption II) We assume that
(i) all members of the same cluster receive the same treatment,
(ii) the cluster sizes in the treatment group are all m_1, and

(iii) the cluster sizes in in the control group are all m_0.
(iv) For the asymptotic results we also assume that

$$\frac{N_1}{N} \to \lambda \ \in (0,1) \ \text{ as } \ n \to \infty.$$

Table 1: Design constants in a design under Assumption 3

.

	Treatment	**Control**	Σ
Number of clusters	n_1	n_0	n
Cluster size	m_1	m_0	
Number of subjects	N_1	N_0	N

See Table 1 for the design constants to be optimized under Assumption 3. As the cluster sizes are constant in both groups, it is clear that the optimal weights for the estimate $\hat{\Delta} = \sum_i \sum_j g_{ij} X_{ij}$ are given by $g_{ij} = \frac{G_{ij}}{N_1} - \frac{1-G_{ij}}{N_0}$. This then gives our estimate, the difference of the sample means, and we have the following result.

Corollary 1. Under assumptions 1 and 3 and under the sequence of alternative hypotheses $H_n : \Delta = \frac{\Delta_0}{\sqrt{N}}$,

$$T^2 \to_d \chi_1^2(\delta^2) \text{ , where } \ \delta^2 = \frac{\Delta_0^2/\sigma^2}{\frac{1-\rho}{\lambda} + \frac{m_1\rho}{\lambda} + \frac{1-\rho}{1-\lambda} + \frac{m_0\rho}{1-\lambda}}.$$

This result can be applied to calculate sample size and power, and to optimize the design.

In practice the cluster sizes may naturally vary for different reasons. One solution could be to use the anticipated average cluster sizes for the treatment and control groups as an approximation. Another possibility is to use the correction methods as suggested by Candel and Van Breukelen (2009), Kerry and Bland (2001) and van Breukelen et al. (2007) to adjust the sample size for variation in cluster sizes.

4. Power and sample size

In the sample size calculations we assume the cluster structure which is displayed in Table 1. We first define a constant $\delta^2_{\alpha,\beta}$, which depends on the type I error rate α and the on power $1-\beta$ as follows.

Definition 1. Let $\delta^2_{\alpha,\beta}$ be defined as the constant for which

$$P\left\{\chi^2_1(\delta^2_{\alpha,\beta}) > \chi^2_{1,1-\alpha}\right\} = 1-\beta,$$

where $\chi^2_1(\delta^2_{\alpha,\beta})$ has a noncentral chi-square distribution with noncentrality parameter $\delta^2_{\alpha,\beta}$ and $\chi^2_{1,1-\alpha}$ is the $(1-\alpha)$-quantile of a (central) chi-square distribution.

The design in Table 1 is fixed if we fix values of N_1, n_1, N_0 and n_0. We wish to do that in an optimal way. Based on Corollary 1, the power of the level α test for an alternative $H_1 : \Delta = \Delta_*$ is approximately $1-\beta$ if N_1, n_1, N_0 and n_0 are chosen so that (approximately)

$$\delta^2(N_1, n_1, N_0, n_0) = \frac{\Delta^2_*/\sigma^2}{\frac{1-\rho}{N_1} + \frac{\rho}{n_1} + \frac{1-\rho}{N_0} + \frac{\rho}{n_0}} = \delta^2_{\alpha,\beta}, \tag{1}$$

or

$$\frac{1-\rho}{N_1} + \frac{\rho}{n_1} + \frac{1-\rho}{N_0} + \frac{\rho}{n_0} = \gamma_{\alpha,\beta},$$

where

$$\gamma_{\alpha,\beta} = \frac{\Delta^2_*/\sigma^2}{\delta^2_{\alpha,\beta}}.$$

Note that the (approximate) power

$$P\left\{\chi^2_1(\delta^2(N_1, n_1, N_0, n_0)) > \chi^2_{1,1-\alpha}\right\}$$

depends on the model parameters through Δ_*/σ (effect size) and ρ (intra-cluster correlation): see Figure 1 for an illustration of this dependence. Once Δ_*/σ and $\delta^2_{\alpha,\beta}$ have been fixed, there is no unique solution in (N_1, n_1, N_0, n_0) that fulfills the condition (1). The selection of the most suitable configuration may be based on practical aspects of the study conduct, or on cost minimization strategies when each unit—treatment cluster, treatment subject, control cluster and control subject—can be assigned a cost. Given a total amount of costs available, one may also be interested in finding a combination (N_1, n_1, N_0, n_0) with the power as high as possible.

5. Cost minimization versus power maximization

A design minimizing the costs of study C, with given power $1-\beta$, can be found by Lagrange's method. A dual problem is the maximization of the power $1-\beta$ given the total costs C. The costs are determined by

- C_1 = the cost of a subject in the treatment group ($C_1 > 0$),
- c_1 = the cost of a treatment cluster ($c_1 > 0$),
- C_0 = the cost of a subject in the control group ($C_0 > 0$), and
- c_0 = the cost of a control cluster ($c_0 > 0$).

On one hand, the total costs of the study are then given by

$$f(N_1, n_1, N_0, n_0) = C_1 N_1 + c_1 n_1 + C_0 N_0 + c_0 n_0.$$

On the other hand, the power of the study depends on (N_1, n_1, N_0, n_0) through the variance expression

$$g(N_1, n_1, N_0, n_0) = \frac{1-\rho}{N_1} + \frac{\rho}{n_1} + \frac{1-\rho}{N_0} + \frac{\rho}{n_0}.$$

Next we consider two settings: the case when the design of the study can be freely chosen, and the case where there are practical restrictions on the cluster sizes.

5.1. No restrictions

We are confronted with the following two dual problems:

1. Minimize $f(N_1, n_1, N_0, n_0)$ given $g(N_1, n_1, N_0, n_0) = \gamma_{\alpha,\beta}$. The Lagrange objective function with this side condition is
$$f(N_1, n_1, N_0, n_0) - \kappa(g(N_1, n_1, N_0, n_0) - \gamma_{\alpha,\beta}).$$
2. Minimize $g(N_1, n_1, N_0, n_0)$ given $f(N_1, n_1, N_0, n_0) = C$. The Lagrange objective function with this side condition is
$$g(N_1, n_1, N_0, n_0) - \kappa(f(N_1, n_1, N_0, n_0) - C).$$

Here κ is the Lagrange multiplier. It is straightforward to see that, in both cases the solution is of the form

$$N_1 = d\sqrt{\frac{1-\rho}{C_1}}, \quad n_1 = d\sqrt{\frac{\rho}{c_1}}, \quad N_0 = d\sqrt{\frac{1-\rho}{C_0}}, \quad \text{and} \quad n_0 = d\sqrt{\frac{\rho}{c_0}}.$$

In the cost minimization problem d is chosen so that $g(N_1, n_1, N_0, n_0) = \gamma_{\alpha,\beta}$, and in the power maximization problem so that $f(N_1, n_1, N_0, n_0) = C$.

5.2. Restrictions on cluster sizes

In practice the design often cannot be chosen optimally. It is common, for example, that hospitals in a clinical study can only be expected to recruit a certain number of subjects on average. Another example is the TERVA project, where the control clusters are of size 1, and each coach cannot handle too many subjects. The optimization problem with a fixed cluster size is constrained by another two constraints:

1. Minimize $f(N_1, n_1, N_0, n_0)$ given $g(N_1, n_1, N_0, n_0) = \gamma_{\alpha,\beta}$, $N_1 = n_1 m_1$ and $N_0 = n_0 m_0$ with m_0 and m_1 fixed. The Lagrange objective function with these side conditions is

$$f(N_1, n_1, N_0, n_0) - \kappa_1(g(N_1, n_1, N_0, n_0) - \gamma_{\alpha,\beta}) - \kappa_2(N_1 - n_1 m_1) - \kappa_3(N_0 - n_0 m_0).$$

2. Minimize $g(N_1, n_1, N_0, n_0)$ given $f(N_1, n_1, N_0, n_0) = C$, $N_1 = n_1 m_1$ and $N_0 = n_0 m_0$ with m_0 and m_1 fixed. The Lagrange objective function with these side conditions is

$$g(N_1, n_1, N_0, n_0) - \kappa_1(f(N_1, n_1, N_0, n_0) - C) - \kappa_2(N_1 - n_1 m_1) - \kappa_3(N_0 - n_0 m_0).$$

Here κ_1, κ_2 and κ_3 are again the Lagrange multipliers. The solution is of the form

$$N_1 = d\sqrt{\frac{1-\rho}{C_1 - \kappa_2}}, \quad n_1 = d\sqrt{\frac{\rho}{c_1 + \kappa_2 m_1}},$$
$$N_0 = d\sqrt{\frac{1-\rho}{C_0 - \kappa_3}}, \quad \text{and} \quad n_0 = d\sqrt{\frac{\rho}{c_0 + \kappa_3 m_0}}.$$

In the cost minimization problem d is chosen so that $g(N_1, n_1, N_0, n_0) = \gamma_{\alpha,\beta}$, and in the power maximization problem so that $f(N_1, n_1, N_0, n_0) = C$.

6. Health coaching project TERVA

As a motivating example, we go through the cost optimization process in the TERVA project. The target population in the TERVA study consists of subjects who have one or more of three beforehand defined chronic conditions. All subjects are randomized to either a health coaching group or a control group at a 2 : 1 ratio. Subjects at the health coaching group are assigned to a full time health coach. The subjects within the same coach can be correlated, and should therefore be treated as clusters. This study is an example where the treatment itself generates clusters. Thus, clustering is present only in the health coaching group and therefore there are no cluster costs in the control group. However, subjects in the control group can be treated as clusters of size one ($m_0 = 1$).

The Lagrange objective function to minimize the costs with fixed power $1 - \beta$ simplifies in this case to

$$C_1 N_1 + c_1 n_1 + c_0 n_0 - \kappa_1 \left(\frac{1-\rho}{N_1} + \frac{\rho}{n_1} + \frac{1}{n_0} - \gamma_{\alpha,\beta} \right) - \kappa_3 (N_0 - n_0)$$

and the solution for maximizing power with fixed total costs C is given by the objective function

$$\frac{1-\rho}{N_1} + \frac{\rho}{n_1} + \frac{1}{n_0} - \kappa_1 \left(C_1 N_1 + c_1 n_1 + c_0 n_0 - C \right) - \kappa_3 (N_0 - n_0).$$

In both cases the solution is

$$N_1 = d\sqrt{\frac{1-\rho}{C_1}}, \quad n_1 = d\sqrt{\frac{\rho}{c_1}}, \quad \text{and} \quad n_0 = d\sqrt{\frac{1}{c_0 + \kappa_3}},$$

where d is chosen so that

$$\frac{1-\rho}{N_1} + \frac{\rho}{n_1} + \frac{1}{n_0} = \gamma_{\alpha,\beta} \quad \text{(the first case)}$$

or so that

$$C_1 N_1 + c_1 n_1 + c_0 n_0 = C \quad \text{(the second case)}.$$

Set, for example, costs in euro (say) to $C_1 = 200$, $c_1 = 30\,000$, $C_0 = 0$ and $c_0 = 50$, $\alpha = 0.05$, $1 - \beta = 0.8$, effect size $\Delta_* / \sigma = 0.2$, and intra-cluster correlation $\rho = 0.05$. Coaches are very expensive compared to subjects, and control subjects are cheaper than subjects under the coaching program. By

Table 2: Costs (in Euros) and power when $\alpha = 0.05$, effect size $\Delta_*/\sigma = 0.2$, and intra-cluster correlation $\rho = 0.05$.

N	n_1	m_1	n_0	m_0	**Costs**	**Power**
2 518	16	54	1 654	1	735 500	0.821
2 517	16	54	1 653	1	735 450	0.821
2 502	16	53	1 654	1	732 300	0.819
2 501	16	53	1 653	1	732 250	0.819
2 464	15	54	1 654	1	694 700	0.799
2 463	15	54	1 653	1	694 650	0.799
2 449	15	53	1 654	1	691 700	0.797
2 448	15	53	1 653	1	691 650	0.797

using the formulas above, the solution for minimum costs (696 658) for the study can be achieved by taking 15.09 treatment clusters, 53.39 subjects per treatment cluster and 1653.48 controls. Total sample size would be 2459.28.

Table 2 gives eight possible designs (nearest integer solutions) around the cost minimum, and the researcher can choose his/her favorite. If we would have chosen to apply the 1 : 1 ratio by using for example 20 treatment clusters of size 38 and 760 controls, we would need only a total of 1 520 subjects to achieve power of 0.801. However, this design would have been much more expensive: it would cost 790 000 due to the high cost of treatment clusters. With the same amount of money we could achieve the power of 0.846 by implementing the power maximization strategy resulting in total sample size of 2 894 ($n_1 = 17, m_1 = 54, n_0 = 1876$). Compared to a design where treatment and control subjects are allocated at 1:1 ratio, the optimal allocation ratio either saves money, or gives more power.

Cost as a function of n_1 and m_1 is also illustrated in Figure 2.

7. Concluding remarks

This paper develops the necessary asymptotic theory for sample size and power calculations for clustered data. Explicit variance formulae allow for cost minimization (with a fixed power) and power maximization (with a given budget). Ready-to-use R functions for statistical software R (R Development Core Team, 2009) are available at
`http://www.uta.fi/∼kari.tokola/optimize/`

for researchers interested in carrying out the design optimization. The optimizing functions return a table with alternative applicable designs close to the optimum.

Sometimes it could be of interest to estimate the sample size for nonparametric tests. The calculations for nonparametric tests on clustered data can be constructed by following the outlines provided in section 3 but by replacing the original observations with score, such as sign or rank. The score has impact on the noncentrality parameter. The sample size calculations for multivariate outcomes and/or other cluster setups could be developed as well. More detailed examination of these issues is reserved for future research.

8. Acknowledgements

The research has been supported by Pfizer, the Finnish Innovation Fund (SITRA) and Päijät-Häme Social Services Health District (PHSSHD) (Kari Tokola), the Academy of Finland (Jaakko Nevalainen and Hannu Oja) and Natural Sciences and Engineering Research Council of Canada (Denis Larocque). We also wish to thank Klaus Nordhausen for his review of the R code.

References

Bauer, D. J., Sterba, S. K., 2008. Evaluating group-based interventions when control participants are ungrouped. Multivariate Behavioral Recearch 43, 210–236.

Campbell, M. J., Donner, A., Klar, N., 2007. Developments in cluster randomized trials and statistics in medicine. Statistics in Medicine 26, 2–19.

Campbell, M. K., Thomson, S., Ramsay, C. R., MacLennan, G. S., Grimshaw, J. M., 2004. Sample size calculator for cluster randomized trials. Computers in Biology and Medicine 34, 113–125.

Candel, M. J. J. M., Van Breukelen, G. J., 2009. Varying cluster sizes in trials with clusters in one treatment arm: Sample size adjustments when testing treatment effects with linear mixed models. Statistics in Medicine 28, 2307–2324.

Donner, A., Birkett, N., Buck, C., 1981. Randomization by cluster: sample size requirements and analysis. American Journal of Epidemiology 20, 367–376.

Eldridge, S., Cryer, C., Feder, G., M., U., 2001. Sample size calculations for intervention trials in primary care randomizing by primary care group: an empirical illustration from one proposed intervention trial. Statistics in Medicine 20, 367–376.

Eldridge, S. M., Ashby, D., Kerry, S., 2006. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. International Journal of Epidemiology 35, 1292–1300.

Flynn, T. N., Whitley, E., Peters, T. J., 2002. Recruitments strategies in a cluster randomized trial - cost implications. Statistics in Medicine 21, 397–405.

Gangnon, R. E., Kosorok, M. R., 2004. Sample-size formula for clustered survival data using weighted log-rank statistics. Biometrika 91, 263–275.

Hayes, R. J., Bennett, S., 1999. Simple sample size calculation for cluster-randomized trials. International Journal of Epidemiology 28, 319–326.

Headrick, T. C., Zumbo, B. D., 2005. On optimizing multi-level designs: Power under budget constraints. Australian & New Zealand Journal of Statistics 47, 219–229.

Heo, M., Leon, A. C., 2008. Statistical power and sample size requirements for three level hierarchical cluster randomized trials. Biometrics 64, 1256–1262.

Heo, M., Leon, A. C., 2009. Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized trials. Statistics in Medicine 28, 1017–1027.

Hoover, D., 2002. Power for t-test comparisons of unbalanced cluster exposure studies. Journal of Urban Health 79, 278–294.

Kerry, S. M., Bland, J. M., 2001. Unequal cluster sizes for trials in english and welsh general practice: implications for sample size calculations. Statistics in Medicine 20, 377–390.

Klar, N., Donner, A., 2001. Current and future challenges in the design and analysis of cluster randomization trials. Statistics in Medicine 20, 3729–3740.

Lee, C. J., Thompson, S. G., 2005. The use of random effects models to allow for clustering in individually randomized trials. Clinical Trials 2, 163–173.

Lemponen, R., Larocque, D., Nevalainen, J., Oja, H., 2011. Two sample multivariate nonparametric test and estimate for cluster-correlated data. Submitted.

Liu, A., Shih, W. J., Gehan, E., 2002. Sample size and power determination for clustered repeated measurements. Statistics in Medicine 21, 1787–1801.

Liu, X., 2003. Statistical power and optimum sample allocation ratio for treatment and control having unequal costs per unit of randomization. Journal of Educational and Behavioral Statistics 28, 231–248.

McKinlay, S., 1994. Cost efficient designs for cluster unit trials. Preventive Medicine 23, 606–611.

Moerbeek, M., 2006. Power and money in cluster randomized trials: When is it worth measuring a covariate? Statistics in Medicine 25, 2607–2617.

Moerbeek, M., Wong, W. K., 2008. Sample size formulae for trials comparing group and individual treatments in a multilevel model. Statistics in Medicine 27, 2850–2864.

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL http://www.R-project.org

Raudenbush, S., 1997. Statistical analysis and optimal design for cluster randomized trials. Psychological Methods 2, 173–185.

Raudenbush, S. W., Liu, X., 2000. Statistical analysis and optimal design for multisite randomized trials. Psychological Methods 5, 199–213.

Roberts, C., Roberts, S. A., 2005. Design and analysis of clinical trials with clustering effects due to treatment. Clinical Trials 2, 152–162.

Rotondi, M., Donner, A., 2009. Sample size estimation in cluster randomized educational trials: an empirical bayes approach. Journal of Educational and Behavioral Statistics 34, 229–237.

Serfling, R. J., 1980. Approximation Theorems of Mathematical Statistics. John Wiley & Sons.

Tu, X., Kowalski, J., Zhang, J., Lynch, K. G., Crits-Christoph, P., 2004. Power analyses for longitudinal trials and other clusterd designs. Statistics in Medicine 23, 2799–2815.

van Breukelen, G. J. P., Candel, M. J. J. M., Berger, M. P. F., 2007. Relative effiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. Statistics in Medicine 26, 2589–2603.

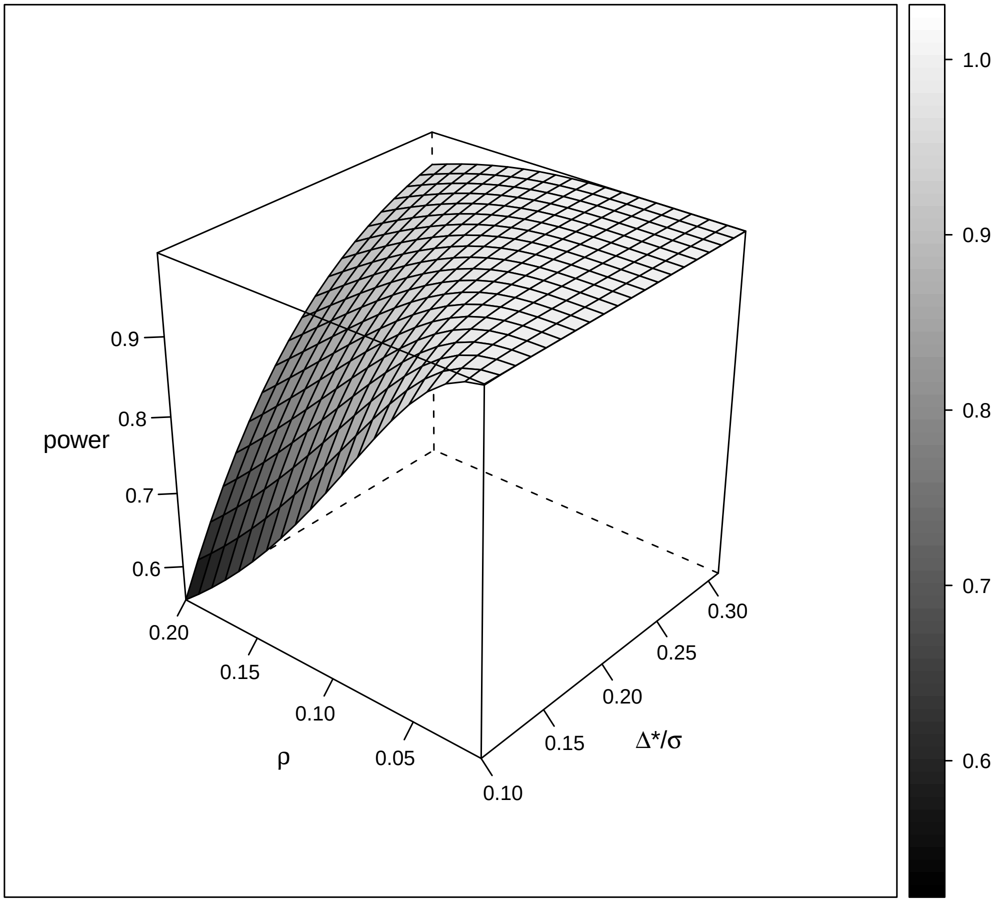

Figure 1: Power as a function of Δ_*/σ and ρ. Design parameters are fixed at $n_1 = 10$, $m_1 = 100$, $n_0 = 500$, $m_0 = 1$ to mimic the TERVA cluster design. The α-level is 0.05.

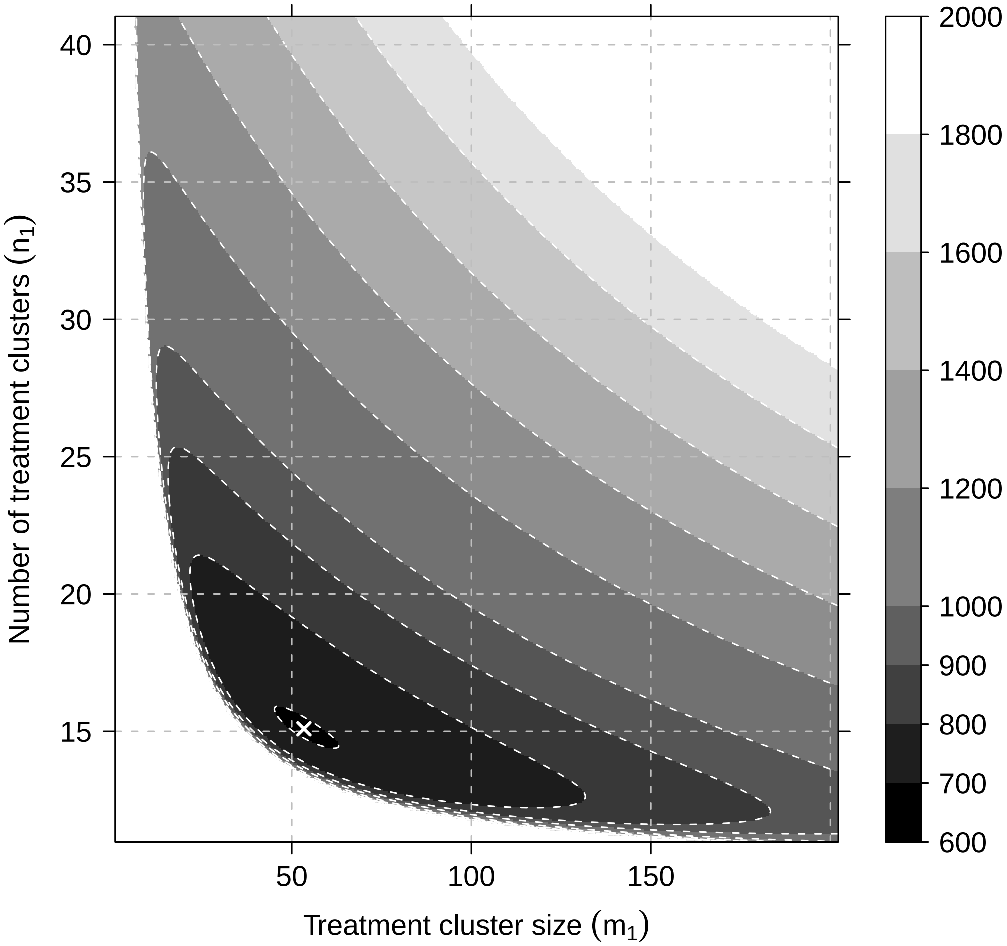

Figure 2: Costs (in thousands) as a function of n_1 and m_1. The number of controls n_0 is changing behind to guarantee a constant power of 0.8. Other parameters are set at $\Delta_*/\sigma = 0.2$, $\rho = 0.05$, $\alpha = 0.05$ and $m_0 = 1$. The minimum is marked with the cross.