Sample size calculation for stepped wedge and other longitudinal cluster randomised trials

Richard Hooper,[a] Steven Teerenstra,[b] Esther de Hoop,[c] Sandra Eldridge[a]

[a] Centre for Primary Care & Public Health, Queen Mary University of London, London, UK

[b] Radboud Institute for Health Sciences, Radboud University Medical Centre, Nijmegen, Netherlands

[c] Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, Netherlands

Address for correspondence:

Dr Richard Hooper,
Centre for Primary Care & Public Health,
Yvonne Carter Building,
58 Turner Street,
London E1 2AB,
UK

phone: 020 7882 7324
fax: 020 7882 2552
email: r.l.hooper@qmul.ac.uk

Word count: 4209 (manuscript) + 213 (abstract)

Abstract

The sample size required for a cluster randomised trial is inflated compared to an individually randomised trial because outcomes of participants from the same cluster are correlated. Sample size calculations for longitudinal cluster randomised trials (including stepped wedge trials) need to take account of at least two levels of clustering: the clusters themselves, and times within clusters. We derive formulae for sample size for repeated cross-section and closed cohort cluster randomised trials with normally distributed outcome measures, under a multi-level model allowing for variation between clusters and between times within clusters. Our formulae agree with those previously described for special cases such as cross-over and ANCOVA design, though simulation suggests that the formulae could underestimate required sample size when the number of clusters is small. Whether using a formula or simulation, a sample size calculation requires estimates of nuisance parameters, which in our model include the intracluster correlation, cluster autocorrelation and individual autocorrelation. A cluster autocorrelation less than 1 reflects a situation where individuals sampled from the same cluster at different times have less correlated outcomes than individuals sampled from the same cluster at the same time. Nuisance parameters could be estimated from time series obtained in similarly clustered settings with the same outcome measure, using analysis of variance to estimate variance components.

Keywords: clinical trial design, cluster randomized trial, intracluster correlation, sample size, stepped wedge

1. Introduction

Cluster randomised trials take as their unit of randomisation a cluster or collective of individual participants [1]. This is typically for pragmatic reasons: the intervention might be delivered at cluster level, for example, or there might otherwise be a risk of contamination between treatments delivered to participants in the same cluster. The sample size required for a cluster randomised trial is inflated compared to an individually randomised trial because outcomes of participants from the same cluster are correlated [2]. A number of articles [3–9] have discussed the calculation of sample size for cluster randomised clinical trials in which two or more independent cross sections are taken from each cluster at given time intervals, with all the participants at any given time in any given cluster receiving either the experimental or the control treatment. The analysis of such a trial should ideally take account of two levels of clustering: the clusters themselves and the cross-sections within clusters. Sample size calculations assuming this kind of hierarchical multi-level model have been described for a general family of repeated cross-section designs including parallel group and stepped wedge designs [8].

Stepped wedge designs randomise clusters to trial arms with varying delays in switching from the control to the experimental intervention [10–12] – see Figure 1 for an example. There is growing interest in the use of stepped wedge trials to evaluate service delivery and other health interventions delivered at an organisational or institutional level, particularly when a policy decision to implement the intervention across a number of clusters has already been made [7, 13]. In such cases stepped wedge designs have a practical advantage over parallel group designs when there are only sufficient resources to "switch on" the

intervention in a small number of clusters at any given time [10, 11, 14], and they may also have a statistical power advantage [6, 15, 16].

Longitudinal cluster randomised trials need not involve taking repeated cross-sections. A recent review [17] has described a broad typology of stepped wedge designs, differentiated according to how individuals are exposed, whether the same individuals are exposed to both the control and the intervention, and how outcome measurements are obtained. One alternative to a repeated cross-section design is a closed cohort design in which participants are all identified at the start of the trial and then followed over time. No general approach to calculating sample size for a closed cohort cluster randomised trial has been described in the literature [7, 9], though Girling & Hemming, in their study of relative efficiency and optimal design [18], have demonstrated the common ground between closed cohort and repeated cross-section designs, building on work done in the special case of the ANCOVA design (a parallel group design with a single baseline and single follow-up assessment) [5]. In this article we establish a common framework for sample size calculations for longitudinal designs which includes a number of previously published results as special cases.

In Section 2 we review the process of sample size calculation for a repeated cross-section cluster randomised trial, which has been described elsewhere [8], and in Section 3 we show how to adapt this calculation to the closed cohort situation. In Section 4 we present an example of a sample size calculation for a closed cohort cluster randomised trial, using our derived formula and using simulation. Finally in Section 5 we discuss our findings and some suggestions for further work.

2. Repeated cross-section cluster randomised trials

2.1. Statistical model

We consider a continuous, normally distributed outcome measure. To keep things simple we assume the same number of individuals is sampled in each cross-section of each cluster. In addition, all the designs considered in this article will have the same number of clusters randomised to each arm of the trial. We assume cross-sections are taken within clusters at predefined, discrete times following randomisation. Note this is qualitatively different to a design involving continuous recruitment of participants over time [17], in which the sample size at each step would depend on its duration and the rate of recruitment.

Suppose, then, that in each of L trial arms l=1,…,L there are K clusters k=1,…,K, each of which has cross-sections taken at times t=0,1,2,…T after randomisation, with each cross-section consisting of m individuals, i=1,…,m. We assume a model in which the outcome for individual i at time t in cluster k, arm l is

$$Y_{itkl} = \gamma + \theta_t + A_{lt}\delta + \xi_{kl} + \eta_{tkl} + \varepsilon_{itkl}, \qquad (1)$$

where

$$\varepsilon_{itkl} \sim \mathrm{N}(0, \sigma^2_{\text{error}})$$

$$\eta_{tkl} \sim \mathrm{N}\left(0, \sigma^2_{\text{time|clust}}\right)$$

$$\xi_{kl} \sim \mathrm{N}(0, \sigma^2_{\text{clust}}),$$

with the ε_{itkl}, η_{tkl} and ξ_{kl} all independent of one another, and

$$A_{lt} = \begin{cases} 1 & \text{if arm } l \text{ is receiving the experimental treatment at time } t \\ 0 & \text{if arm } l \text{ is receiving the control treatment at time } t. \end{cases}$$

This hierarchical multi-level model includes random effects which model variation between clusters (ξ_{kl}) and also variation between times within a cluster (η_{tkl}). The parameter δ is the treatment effect, which we assume is maintained once the intervention has been introduced. The model also includes a fixed effect of time, θ_t, which must be estimated independently of the treatment effect so that a systematic change over time is not mistaken for an effect of treatment. For identifiability we set θ_0=0. Note that in cluster randomised trials with a stepped wedge design it is quite usual for all the clusters to be randomised simultaneously, so that an effect of time since randomisation is the same as an effect of calendar time – an equivalence rarely found in individually randomised trials [19].

It will be convenient to transform the nuisance parameters σ^2_{error}, $\sigma^2_{\text{time|clust}}$ and σ^2_{clust} into three new parameters

$$\sigma^2 = \sigma^2_{\text{error}} + \sigma^2_{\text{time|clust}} + \sigma^2_{\text{clust}}$$

$$\rho = \left(\sigma^2_{\text{time|clust}} + \sigma^2_{\text{clust}}\right) / \left(\sigma^2_{\text{error}} + \sigma^2_{\text{time|clust}} + \sigma^2_{\text{clust}}\right)$$

$$\pi = \sigma^2_{\text{clust}} / \left(\sigma^2_{\text{time|clust}} + \sigma^2_{\text{clust}}\right)$$

where σ^2 is the total variance and ρ is the intra-cluster correlation – *i.e.* the correlation between assessments of two individuals sampled from the same cluster at the same time. The interpretation of π is discussed below.

2.2. Sample size calculation

We use the following notation for different settings: SI = Single cross-section, Individually randomised; SC = Single cross-section, Cluster randomised; RC = Repeated cross-section, Cluster randomised; CI = Closed cohort, Individually randomised; CC = Closed cohort, Cluster randomised.

We define n_{SI} to be the total sample size required to detect treatment effect δ^* with power $1-\beta$ at two-sided significance level α, by comparing two independent groups of equal size, randomised at the individual level and assessed once. n_{SI} can be calculated using the approximate formula [20]:

$$n_{\mathrm{SI}} = 4\left(\frac{\sigma}{\delta^*}\right)^2 \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \tag{2}$$

where z_p is the 100pth centile of a standard normal distribution. Alternatively this sample size can be obtained from standard tables or software.

Using a single cross-section cluster randomised (SC) design the total number of clusters required, allowing for the clustering [2], becomes

$$N_{\mathrm{SC}} = \mathrm{Deff}_{\mathrm{C}}(m,\rho) \times \frac{n_{\mathrm{SI}}}{m}$$

where $\mathrm{Deff}_{\mathrm{C}}$ is the design effect due to cluster randomising

$$\mathrm{Deff}_{\mathrm{C}}(m,\rho) = 1 + (m-1)\rho. \tag{3}$$

If the same clusters are assessed in repeated cross sections then fewer clusters are required to achieve the same power. The efficiency of a repeated cross-section design under model (1) can be determined by finding the linear unbiased estimator for the treatment effect that has smallest variance. Formulae for sample size have been derived elsewhere [8]. The total number of clusters required is

$$N_{\mathrm{RC}} = \mathrm{Deff}_{\mathrm{R}}(r_{\mathrm{RC}}) \times \mathrm{Deff}_{\mathrm{C}}(m, \rho) \times \frac{n_{\mathrm{SI}}}{m} \tag{4}$$

where r_{RC} is the correlation between two sample means of *m* participants from the same cluster in different cross-sections

$$r_{\mathrm{RC}} = \frac{m\rho\pi}{1 + (m-1)\rho} \tag{5}$$

and $\mathrm{Deff}_{\mathrm{R}}(r)$ is the design effect due to repeated assessment for correlation *r*

$$\mathrm{Deff}_{\mathrm{R}}(r) = \frac{L^2(1-r)(1+Tr)}{4(LB - D + (B^2 + LTB - TD - LC)r)}$$

with constants *B*,*C* and *D* defined from matrix **A** in (1)

$$B = \Sigma_{tl}A_{lt},\ \ C = \Sigma_l(\Sigma_t A_{lt})^2,\ D = \Sigma_t(\Sigma_l A_{lt})^2.$$

Each cluster has $m(T+1)$ participants; hence the total number of participants required for a repeated cross-section cluster randomised trial is

$$n_{\mathrm{RC}} = \mathrm{Deff}_{\mathrm{R}}(r_{\mathrm{RC}}) \times \mathrm{Deff}_{\mathrm{C}}(m, \rho) \times (T+1) n_{\mathrm{SI}}. \quad (6)$$

In (5) the parameter π is the limit of r_{RC} as $m\rho \to \infty$, and can thus be interpreted as the correlation between two population means from the same cluster at different times. We refer to π as the cluster autocorrelation [5].

Consider, for example, a stepped wedge design of the form shown in Figure 1, but with an arbitrary number of steps *L*. In this case *T* = *L* and matrix **A** is given by

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & \cdots & 1 & 1 \\ 0 & 0 & 1 & \cdots & 1 & 1 \\ 0 & 0 & 0 & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}.$$

The design effect due to repeated assessment is then

$$\mathrm{Deff}_{\mathrm{R},L\text{-step}}(r) = \frac{3L(1-r)(1+Lr)}{(L^2-1)(2+Lr)} \quad (7)$$

and the total required sample size becomes

$$n_{\mathrm{RC},L\text{-step}} = \frac{3L(1-r_{\mathrm{RC}})(1+Lr_{\mathrm{RC}})}{(L-1)(2+Lr_{\mathrm{RC}})}(1+(m-1)\rho) n_{\mathrm{SI}}. \quad (8)$$

2.3. Related sample size formulae

The general formula (6) includes some previously described formulae as special cases. Consider, for example, a cross-over design with $L = 2$, $T = 1$, and matrix $\mathbf{A}$ given by

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then the design effect due to repeated assessment becomes

$$\mathrm{Deff}_{\mathrm{R,cross}}(r) = \frac{1-r}{2}$$

and the total required sample size becomes

$$n_{\mathrm{RC,cross}} = (1 + (m-1)\rho - m\rho\pi)n_{\mathrm{SI}}$$

which reproduces the formula of Giraudeau and colleagues [4] for a repeated cross-section cluster randomised cross-over trial.

Alternatively, consider the ANCOVA design, with $L = 2$, $T = 1$, and matrix $\mathbf{A}$ given by

$$\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

In this case the design effect due to repeated assessment becomes

$$\mathrm{Deff}_{\mathrm{R,ANCOVA}}(r) = 1 - r^2$$

and the total required sample size becomes

$$n_{\mathrm{RC,ANCOVA}} = 2(1 - r_{\mathrm{RC}}^2)(1 + (m-1)\rho)n_{\mathrm{SI}}$$

which is equivalent to the formula of Teerenstra and colleagues [5] for a repeated cross-section cluster randomised trial with an ANCOVA design.

If we make the simplifying assumption in model (1) that $\sigma^2_{\mathrm{time|clust}} = 0$ (or equivalently that $\pi = 1$), then equation (6) reduces to the sample size formula of Hussey & Hughes [3], and (8) reduces to the formula of Woertman and colleagues [6] for a stepped wedge design. This simplified model with $\pi = 1$ is also the one assumed in Hemming and colleagues' "steppedwedge" sample size procedure in Stata [21].

3. Closed cohort cluster randomised trials

3.1. Statistical model

In a closed cohort cluster randomised trial we follow the same participants over time rather than taking a fresh cross-section from each cluster at each time. We assume that the participants in a given cluster are all identified at the beginning of the trial and assessed at a series of predefined, discrete times following randomisation. For any given assessment all participants from a given cluster are exposed either to the experimental or the control treatment. In a closed cohort design we need to allow for dependence between outcomes assessed in the same participant over time. The simplest way to model this is to assume an

additional random effect of participant [7]. Thus we assume a model in which the outcome at time t=0,…,T for individual i=1,…,m in cluster k=1,…,K, arm l=1,…,L is

$$Y_{itkl} = \gamma + \theta_t + A_{lt}\delta + \xi_{kl} + \zeta_{ikl} + \eta_{tkl} + \varepsilon_{itkl}, \quad (9)$$

where

$$\varepsilon_{itkl} \sim \mathrm{N}(0, \sigma^2_{\text{error}})$$

$$\zeta_{ikl} \sim \mathrm{N}\left(0, \sigma^2_{\text{indiv|clust}}\right)$$

$$\eta_{tkl} \sim \mathrm{N}\left(0, \sigma^2_{\text{time|clust}}\right)$$

$$\xi_{kl} \sim \mathrm{N}(0, \sigma^2_{\text{clust}}),$$

with the ε_{itkl}, η_{tkl}, ζ_{ikl} and ξ_{kl} all independent of one another, and

$$A_{lt} = \begin{cases} 1 & \text{if arm } l \text{ is receiving the experimental treatment at time } t \\ 0 & \text{if arm } l \text{ is receiving the control treatment at time } t. \end{cases}$$

This is no longer a hierarchical model: individuals are not nested within times, nor times within individuals – instead each individual is assessed at each time, and there is a random effect of individual within cluster, ζ_{ikl}, and a random effect of time within cluster, η_{tkl}. This is an example of a cross-classified multi-level model [22]. The random effect of time within cluster might be the result of changes to the way in which treatment is delivered at a given cluster over time, and represents a kind of interaction between time and cluster.

We define

$$\sigma^2 = \sigma^2_{\text{error}} + \sigma^2_{\text{indiv|clust}} + \sigma^2_{\text{time|clust}} + \sigma^2_{\text{clust}}$$

$$\rho = \left(\sigma^2_{\text{time|clust}} + \sigma^2_{\text{clust}}\right) / \left(\sigma^2_{\text{error}} + \sigma^2_{\text{indiv|clust}} + \sigma^2_{\text{time|clust}} + \sigma^2_{\text{clust}}\right)$$

$$\pi = \sigma^2_{\text{clust}} / \left(\sigma^2_{\text{time|clust}} + \sigma^2_{\text{clust}}\right)$$

$$\tau = \sigma^2_{\text{indiv|clust}} / \left(\sigma^2_{\text{error}} + \sigma^2_{\text{indiv|clust}}\right)$$

where, as in model (1), σ^2 is the total variance, ρ is the intracluster correlation (the correlation between assessments of two individuals from the same cluster at the same time) and π is the cluster autocorrelation (the correlation between two population means from the same cluster at different times). τ is the correlation between two assessments of the same individual at different times in a given cluster, or the individual autocorrelation [5].

3.2. Sample size calculation

As noted above, the efficiency of a trial design under a mixed model such as (1) or (9) can be determined by finding the linear unbiased estimator for the treatment effect that has smallest variance [8]. In the present case we can reasonably restrict attention to estimators that are linear combinations of the cluster means at each time, $\bar{Y}_{\cdot tkl}$, since each individual from the same cluster at a given time is assessed under the same conditions, and therefore will contribute equally to the treatment effect estimate. Comparing models (1) and (9) for given K and **A** we see that $\bar{Y}_{\cdot tkl}$ has the same essential structure in both, with variance given by the same expression in m, σ^2 and ρ under both models. The only other thing that matters to the variance of the treatment effect estimator is the correlation, r, between $\bar{Y}_{\cdot t_1 kl}$ and $\bar{Y}_{\cdot t_2 kl}$ for any

k, l, $t_1 \neq t_2$. In other words, for given K, $\mathbf{A}$, m, σ^2, ρ and r the best linear unbiased estimator is the same whether we are considering model (1) or model (9).Consequently the number of clusters per group, K, required to achieve given statistical power for given $\mathbf{A}$, m, σ^2, ρ and r is the same for a closed cohort cluster randomised (CC) design as for a repeated cross-section cluster randomised (RC) design. This equivalence has been noted previously in the special case of the ANCOVA design [5]. Thus from (4) we know that the total number of clusters required to achieve power $1-\beta$ at two-sided significance level α to detect a treatment effect δ^* is

$$N_{\mathrm{CC}} = \mathrm{Deff}_{\mathrm{R}}(r_{\mathrm{CC}}) \times \mathrm{Deff}_{\mathrm{C}}(m, \rho) \times \frac{n_{\mathrm{SI}}}{m}$$

where r_{CC} is the correlation [5] between $\bar{Y}_{\cdot t_1 kl}$ and $\bar{Y}_{\cdot t_2 kl}$ for any k, l, $t_1 \neq t_2$ under model (9):

$$r_{\mathrm{CC}} = \frac{m\rho\pi + (1-\rho)\tau}{1 + (m-1)\rho}. \tag{10}$$

In a closed cohort design there are m participants per cluster, so the total number of participants required for a closed cohort cluster-randomised trial is

$$n_{\mathrm{CC}} = \mathrm{Deff}_{\mathrm{R}}(r_{\mathrm{CC}}) \times \mathrm{Deff}_{\mathrm{C}}(m, \rho) \times n_{\mathrm{SI}}. \tag{11}$$

In the case of a stepped wedge cohort design this agrees with the formula derived by de Hoop and colleagues [23]. Note that for a closed cohort individually randomised (CI) design, with $m = 1$, ρ and π effectively zero, and $r = \tau$, the total number of participants required becomes

$$n_{\mathrm{CI}} = \mathrm{Deff}_{\mathrm{R}}(\tau) \times n_{\mathrm{SI}}.$$

This brings us back to familiar formulae for individually randomised trials with longitudinal designs [24]. For example, in an individually randomised trial with an ANCOVA design the total number of participants required is

$$n_{\mathrm{CI,ANCOVA}} = (1 - \tau^2) \times n_{\mathrm{SI}}.$$

General sample size formulae for repeated cross-section and closed cohort cluster randomised trials are summarised in Table 1, for easy reference. Design effects due to repeated assessment are tabulated by Hooper & Bourke [8] for a variety of designs and families of designs.

4. Sample size calculation in practice

4.1 Example

A recent review of stepped wedge trials published between 2010 and 2014 [17] identified 11 trials with a closed cohort design. We use one of these – an evaluation of the "Girls on the Go!" program to improve self-esteem in young women in Australia [25] – as an exemplar for our approach to sample size calculation. Clusters in this case were schools: primary schools were randomised to a two-step design (with assessments at baseline and after two successive school terms) and secondary schools to a three-step design (with assessments at baseline and after three successive terms). We consider the design of a three-step study. For their power calculation the investigators assumed 10 participants per cluster and an intracluster

correlation of 0.33. The primary outcome measure was the Rosenberg Self-Esteem Scale – a ten-item scale with each item scoring 1–4. Previous research using this scale [26] suggests an individual autocorrelation of around 0.7. There is less basis for the cluster autocorrelation – we will assume a figure of 0.9. Informed by a pilot study the investigators powered their trial to detect a difference of 1.5 standard deviations. This seems optimistic, and we consider instead the sample size to achieve 80% power at the 5% significance level to detect a mean difference of 2 on the Rosenberg Self-Esteem Scale, assuming a standard deviation (σ) of 5.

If this was a single cross-section, individually randomised trial, the total sample size required (equation 2) would be 198. For a single cross-section cluster randomised trial this would need to be multiplied by the design effect due to cluster randomising (equation 3), which in this case is 3.97, giving a total sample size of 786. Some statistical power can be reclaimed, however, with the longitudinal design: the correlation between two cluster sample means from the same cluster at different times (equation 10) is 0.8662; hence the design effect due to repeated assessment in a 3-step stepped wedge design (equation 7) is 0.1178. The total sample size required, according to our formula (equation 11), is therefore 786 × 0.1178 = 93. This sample size requirement would not quite be achieved with 3 clusters in each arm, but is more than met if we include 4 clusters per arm, or 12 clusters (120 individuals) in total (working back from equations (11) and (2) this gives a power of 89.3%).

4.2 Sample size by simulation

Simulation has been recommended as a robust alternative to approximate sample size formulae for stepped wedge trials [9]. For the example above we also determined the required sample size by simulation, using the simsam package in Stata [27], with a bespoke

programme to generate a data-set from our model and analyse it using mixed regression with restricted maximum likelihood (programme and simsam output are available in the Supporting Information for this article). The simsam package confirms that 4 clusters in each arm are needed to achieve 80% power, the estimate of power with 4 clusters per arm being consistent with the value obtained from the formula (99% Monte Carlo confidence interval 88.7% - 89.5%). However, simulation also shows in this case that the Type I error rate exceeds 5% - a general problem affecting small-sample inference using linear mixed models [28]. If we could adjust the test to control this level correctly then the power would be reduced (corrections such as that of Kenward & Roger [28] are available but are computationally intensive, prohibiting large numbers of simulations for estimating power).

4.3 Sensitivity of sample size to assumptions

One advantage of the formula over simulation is that it allows us to relate the total required sample size to the sample size per cluster and the various nuisance parameters. Figure 2 shows the overall design effect – defined here as the product of the design effect due to repeated assessment and the design effect due to cluster randomising – for differing values of the intracluster correlation, cluster autocorrelation, individual autocorrelation, and sample size per cluster, in the example of a 3-step closed cohort stepped wedge design. The overall design effect determines the required sample size for given effect size, significance and power.

Required sample size increases with sample size per cluster. When the intracluster correlation is close to 1 and the cluster autocorrelation is close to 0, the required sample size is relatively insensitive to the intracluster correlation, cluster autocorrelation or individual autocorrelation.

However, for larger sample sizes per cluster, and for individual autocorrelations close to 1, the required sample size is particularly sensitive to the intracluster correlation when the latter is close to 0, and particularly sensitive to the cluster autocorrelation when the latter is close to 1. The sensitivity to the cluster autocorrelation is also magnified at larger intracluster correlations (note that the intracluster correlation and cluster autocorrelation both increase from right to left on the graphs, to aid three-dimensional visualisation of the surfaces).

In practical applications the intracluster correlation is often assumed to be small, and a simplified model with a cluster autocorrelation of 1 may also be assumed (Cf Hussey & Hughes [3]), but we would do well to be conservative and overestimate the intracluster correlation, and underestimate the cluster autocorrelation, given the sensitivities noted above. In the "Girls on the Go!" example the value assumed for the intracluster correlation is relatively high, so the sample size is particularly sensitive to an over-optimistic estimate of cluster autocorrelation. Note that in stepped wedge trials clusters are usually randomised in one go, at the start of the trial, so there is no possibility of modifying the sample size of a closed cohort part-way through using interim estimates of nuisance parameters.

5. Discussion

We have shown how to calculate sample size for a longitudinal cluster randomised trial with a repeated cross-section or closed cohort design. Randomising in clusters reduces statistical power, but assessing the same individuals or clusters under the control and experimental condition at different times can be an efficient approach to design. In choosing a design for any trial, investigators must weigh the competing costs associated with numbers of clusters, individual participants, assessments, and time-points. In this article we have only attempted

to consider the question of sample size: methodological problems such as how to prevent attrition bias in longitudinal cluster randomised trials, and how and when to obtain consent will need further practical investigation [7, 14].

The general formula given here for the design effect due to repeated assessment applies to any "complete" design [29] – that is, one where every trial arm includes an assessment at every time-point. Formulae for incomplete designs such as the dog-leg design can be derived as a separate exercise [8, 30]. Our models assume the same number of clusters in each arm, though the results are easily generalised to other cases by subdividing unequally sized arms into smaller arms which all have the same number of clusters (or simply by re-defining each cluster to be an arm). The most efficient distribution of clusters between arms in a complete stepped wedge design with a given number of steps has been described elsewhere [18, 31]. Our formulae are derived for continuous, normally distributed outcome measures, but can be extended naturally to binary outcomes by extending the definitions of intracluster correlation and individual autocorrelation in some appropriate way [32]. We have not considered the issue of variable cluster size [33].

Asymptotic sample size formulae will underestimate required sample size when the number of clusters is small. More research is needed to determine rules of thumb for correcting the sample size in this case. Simulation and formulae may both turn out to have a useful role in planning longitudinal cluster randomised trials. More work is also needed on analysis: our model for closed cohort cluster randomised trials requires an analysis of cross-classified random effects, methods for which are available in existing statistical computing packages such as Stata, SAS, SPSS, R and MLwiN. In the case of designs with more than two time-points our random effects models effectively assume an "exchangeable" correlation structure

within clusters and individuals over time, that is one in which the cluster or individual autocorrelation is the same whichever two time-points we consider. Exchangeability assumptions are common to most existing approaches to sample size calculation for longitudinal cluster randomised trials. How reasonable they are depends on the outcome measure, the interval between assessments, the nature of the intervention, and the nature of the clusters and participants. Analysis using alternative correlation structures and the robustness of an exchangeability approach to sample size calculation in these cases warrant further investigation, as do methods to combat Type I error rate inflation when the number of clusters is small.

Whether we use a formula or simulation to determine sample size for a repeated cross-section or closed cohort cluster randomised trial, we need estimates of the nuisance parameters. These could, in principle, be estimated from time series obtained in similarly clustered settings with the same outcome measure, using analysis of variance to estimate variance components. The intracluster correlation can be estimated from a single cross-section, and this parameter is already widely reported for a variety of outcomes and settings. The individual autocorrelation of an outcome measure may be known from validation studies. The hardest parameter to quantify is likely to be the cluster autocorrelation. It is tempting to set the latter to 1, but this will lead to an over-estimate of the correlation between sample means from the same cluster at different times, and hence an under-estimate of the required sample size – that is an underpowered study. A cluster autocorrelation less than 1 allows us to model situations where the correlation between outcomes of two individuals sampled from the same cluster at different times ($\pi\rho$) is smaller than the correlation between outcomes of two individuals sampled from the same cluster at the same time (ρ). This is well appreciated in the context of cluster randomised cross-over trials, where Giraudeau and colleagues [4] have

suggested assuming a cluster autocorrelation of 0.5 in the absence of other guidance, but less so for stepped wedge designs: of ten reports of stepped wedge trials published between 2010 and 2014 [34], none included a component of variance between times within clusters in the analysis. In a repeated cross-section design some of the variation between times within a cluster arises because the different cross-sections come from different cohorts, and we might expect closed cohort studies, by definition, to have higher cluster autocorrelations than repeated cross-section studies. Such general rules of thumb are speculative, however. More work is urgently needed to evaluate and report cluster autocorrelations from real-life time series. Perhaps authors of longitudinal cluster randomised trials should be encouraged to report estimates of cluster autocorrelations, just as intracluster correlations are already routinely reported.

Funding

The work summarised in this article was not funded by a specific grant. RH is supported by the UK National Institute for Health Research as part of its funding for the Pragmatic Clinical Trials Unit at Barts & The London School of Medicine & Dentistry. EH is supported by the 'MultiFaCT' internal grant from the Julius Center for Health Sciences & Primary Care at the University Medical Center Utrecht.

References

1. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold: London, 2000
2. Kerry SM, Bland JM. The intracluster correlation coefficient in cluster randomisation. *BMJ* 1998;**316**(7142):1455. DOI:10.1136/bmj.316.7142.1455
3. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomised trials. *Contemp Clin Trials* 2007;**28**(2):182–91. DOI: 10.1016/j.cct.2006.05.007
4. Giraudeau B, Ravaud P, Donner A. Sample size calculations for cluster randomized cross-over trials. *Stat Med* 2008;**27**(27):5578–85. DOI: 10.1002/sim.3383
5. Teerenstra S, Eldridge S, Graff M, de Hoop E, Borm GF. A simple sample size formula for analysis of covariance in cluster randomized trials. *Stat Med* 2012;**31**(20):2169–78. DOI: 10.1002/sim.5352
6. Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol* 2013;**66**(7):752–8. DOI:10.1016/j.jclinepi.2013.01.009
7. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis and reporting. *BMJ* 2015;**350**:h391. DOI: 10.1136/bmj.h391
8. Hooper R, Bourke L. Cluster randomised trials with repeated cross-sections: alternatives to parallel group designs. *BMJ* 2015;**350**:h2925. DOI: 10.1136/bmj.h2925
9. Baio G, Copas A, Ambler G, Hargreaves J, Beard E, Omar RZ. Sample size calculation for a stepped wedge trial. *Trials* 2015;**16**:354. DOI: 10.1186/s13063-015-0840-9
10. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol* 2006;**6**:54. DOI: 10.1186/1471-2288-6-54

11. Mdege ND, Man M-S, Taylor CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *J Clin Epidemiol* 2011;**64**(9):936–48. DOI: 10.1016/j.jclinepi.2010.12.003
12. Beard A, Lewis JJ, Copas A, Davey C, Osrin D, Baio G, Thompson JA, Fielding KL, Omar RZ, Ononge S, Hargreaves J, Prost A. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials* 2015;**16**:353. DOI: 10.1186/s13063-015-0839-2
13. Hargreaves JR, Copas AJ, Beard E, Osrin D, Lewis JJ, Davey C, Thompson JA, Baio G, Fielding KL, Prost A. Five questions to consider before conducting a stepped wedge trial. *Trials* 2015;**16**:350. DOI: 10.1186/s13063-015-0841-8
14. Prost A, Binik A, Abubakar I, Roy A, De Allegri M, Mouchoux C, Dreischulte T, Ayles H, Lewis JJ, Osrin D. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies. *Trials* 2015;**16**:351. DOI: 10.1186/s13063-015-0837-4
15. Hemming K, Girling A. The efficiency of stepped wedge vs. cluster randomized trials: stepped wedge studies do not always require a smaller sample size. *J Clin Epidemiol* 2013;**66**(12):1427–8. DOI:10.1016/j.jclinepi.2013.07.007
16. de Hoop E, Woertman W, Teerenstra S. The stepped wedge cluster randomized trial always requires fewer clusters but not always fewer measurements, that is, participants than a parallel cluster randomized trial in a cross-sectional design. *J Clin Epidemiol* 2013;**66**(12):1428. DOI:10.1016/j.jclinepi.2013.07.008
17. Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G, Hargreaves JR. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials* 2015;**16**:352. DOI: 10.1186/s13063-015-0842-7

18. Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Stat Med* 2016;DOI: 10.1002/sim.6850
19. Senn S. Seven myths of randomisation in clinical trials. *Stat Med* 2013;**32**(9):1439–50. DOI: 10.1002/sim.5713
20. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *BMJ* 1995;**311**(7013):1145. DOI: 10.1136/bmj.311.7013.1145
21. Hemming K, Girling A. A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster randomized trials. *Stata J* 2014;**14**(2):363–80.
22. Goldstein H. Multilevel covariance component models. *Biometrika* 1987;**74**(2):430–1.
23. de Hoop E, Moerbeek M, Gerritsen DL, Teerenstra S. Sample size estimation for cohort and cross-sectional cluster randomized stepped wedge designs. In: Oomen-de Hoop E, *Efficient designs for cluster randomized trials with small numbers of clusters: stepped wedge and other repeated measurements designs (doctoral thesis)* [Accessed February 2016]. Available from: http://repository.ubn.ru.nl/bitstream/handle/2066/134179/134179.pdf?sequence=1
24. Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Stat Med* 1992;**11**(13):1685–704. DOI: 10.1002/sim.4780111304
25. Tirlea L, Truby H, Haines TP. Investigation of the effectiveness of the "Girls on the Go!" program for building self-esteem in young women: trial protocol. *SpringerPlus* 2013;**2**:683. DOI: 10.1186/2193-1801-2-683
26. Torrey WC, Mueser KT, McHugo GH, Drake RE. Self-esteem as an outcome measure in studies of vocational rehabilitation for adults with severe mental illness. *Psychiatric Services* 2000;**51**(2):229–233. DOI: 10.1176/appi.ps.51.2.229

27. Hooper R. Versatile sample size calculation using simulation. *Stata Journal* 2013;**13**(1):21–38.
28. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997;**53**(3):983–997.
29. Hemming K, Lilford R, Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Stat Med* 2014;**34**(2):181–96. DOI: 10.1002/sim.6325
30. Hooper R, Bourke L. The dog-leg: an alternative to a cross-over design for pragmatic clinical trials in relatively stable populations. *Int J Epidemiol* 2014;**43**(3):930–6. DOI: 10.1093/ije/dyt281
31. Lawrie J, Carlin JB, Forbes AB. Optimal stepped wedge designs. *Stat Prob Letters* 2015;99:210–214.
32. Ridout MS, Demetrio CGB, Firth D. Estimating intraclass correlation for binary data. *Biometrics* 1999;**55**(1):137–148. DOI: 10.1111/j.0006-341X.1999.00137.x
33. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomised trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol* 2006;**35**(5):1292–1300. DOI:10.1093/ije/dyl129
34. Davey C, Hargreaves J, Thompson JA, Copas AJ, Beard E, Lewis JJ, Fielding KL. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials* 2015;**16**:358. DOI: 10.1186/s13063-015-0838-3

Figure legends

Figure 1. Schematic illustration of assessment in a stepped wedge trial design with five steps. Assessments in each arm at each of times 0 to 5 are made under either the experimental treatment or the control treatment, and the trial arms differ according to the delay with which clusters move from the control to the experimental treatment.

Figure 2. Overall design effect for a 3-step closed cohort stepped wedge trial design, according to the intracluster correlation (ICC), cluster autocorrelation (CAC), individual autocorrelation (IAC), and sample size per cluster, *m*.

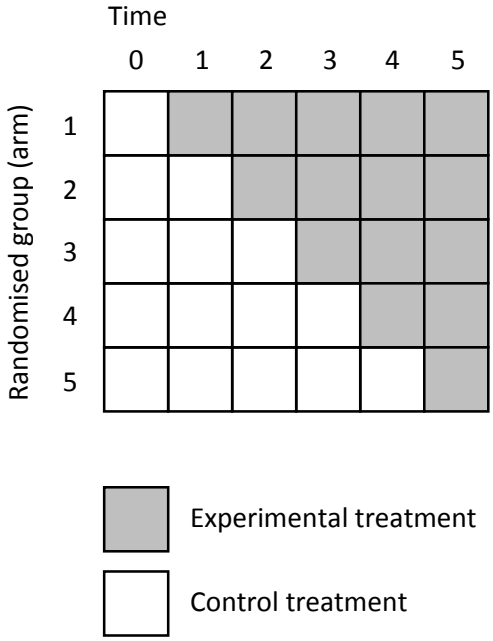
Table 1. Formulae for sample size to achieve given power using repeated cross-section and closed cohort cluster randomised trial designs.

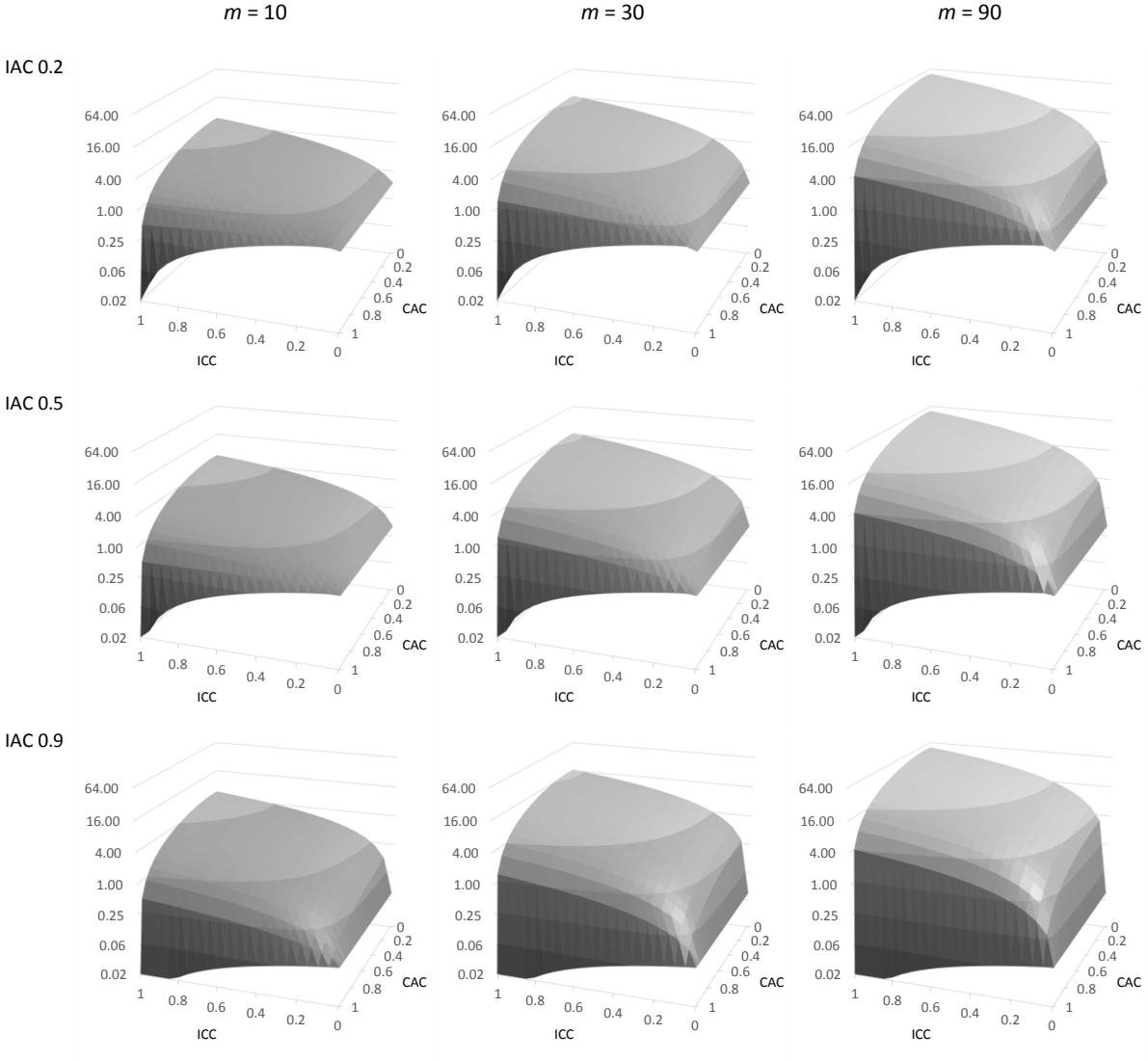
	Repeated cross-section: at each time-point (0,1,2,…,T), m participants are sampled from each cluster	Closed cohort: m participants are sampled from each cluster at baseline and assessed at every time-point
Total number of clusters	$\text{Deff}_\text{R}(r) \times \text{Deff}_\text{C}(m,\rho) \times \frac{n_\text{SI}}{m}$	$\text{Deff}_\text{R}(r) \times \text{Deff}_\text{C}(m,\rho) \times \frac{n_\text{SI}}{m}$
Total number of participants	$\text{Deff}_\text{R}(r) \times \text{Deff}_\text{C}(m,\rho) \times (T+1)n_\text{SI}$	$\text{Deff}_\text{R}(r) \times \text{Deff}_\text{C}(m,\rho) \times n_\text{SI}$
Correlation, r, between two sample means from the same cluster at different times	$\frac{m\rho\pi}{1+(m-1)\rho}$	$\frac{m\rho\pi+(1-\rho)\tau}{1+(m-1)\rho}$

n_SI is the total number of participants required for a single cross-section, individually randomised design; ρ is the intracluster correlation; π is the cluster autocorrelation; τ is the individual autocorrelation. For definitions of the design effect due to repeated assessment, Deff_R, and the design effect due to cluster randomising, Deff_C, see the text.

Supporting information legend

Sample size calculation by simulation in Stata, for the example given in the article: Stata code and output (pdf)


Time
0
1
2
3
4
5
Randomised group (arm)
1
2
3
4
5
Experimental treatment
Control treatment


m = 10
m = 30
m = 90
IAC 0.2
IAC 0.5
IAC 0.9
64.00
16.00
4.00
1.00
0.25
0.06
0.02
ICC
CAC
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1

SUPPORTING INFORMATION FOR:
Sample size calculation for stepped wedge and other longitudinal cluster randomised trials
Richard Hooper, Steven Teerenstra, Esther de Hoop, Sandra Eldridge

Sample size calculation by simulation in Stata, for the example given in the article

1. Stata code

```
program define s_cohortstep, rclass

*** This programme generates and analyses a data-set
*** from a closed cohort, stepped wedge trial.

      version 12.0
      syntax, SD(real) ICC(real) CAC(real) IAC(real)  ///
            NCLUSPERGRP(integer) CLUSSIZE(integer)  ///
            NSTEP(integer)  ///
            TIMECOEFF(real) TREATCOEFF(real)

      drop _all

      scalar sdclus=`sd'*sqrt(`icc'*`cac')
      scalar sdtime=`sd'*sqrt(`icc'*(1-`cac'))
      scalar sdchar=`sd'*sqrt(`iac'*(1-`icc'))
      scalar sderr=`sd'*sqrt((1-`iac')*(1-`icc'))

      set obs `=`nstep'*`ncluspergrp''
      gen idclus=_n
      gen group=1+mod(_n-1,`nstep')
      gen rand_clus=rnormal(0,sdclus)
      forvalues i=0/`nstep' {
                  gen rand_time`i'=rnormal(0,sdtime)
      }

      expand `clussize'
      sort idclus
      gen id=_n
      gen rand_char=rnormal(0,sdchar)

      reshape long rand_time, i(id) j(time)
      sort idclus id time

      gen treat=(time>=group)
      gen rand_err=rnormal(0,sderr)
      gen y=`timecoeff'*time+`treatcoeff'*treat+  ///
            rand_clus+rand_time+rand_char+rand_err

      xtmixed y i.time treat || idclus: || idclus: R.time || id:, reml
      return scalar p=2*normal(-abs(_b[treat]/_se[treat]))

end

*** NB the data generation step in s_cohortstep assumes
*** a linear effect of time for convenience, but the
*** analysis fits time as a categorical variable.
```

```
*** The simsam command determines sample size by simulation. To
*** download simsam use the command "findit simsam" and follow
*** instructions for installation - a help file is included.

*** NB as specified below simsam takes a very long time to run,
*** because of the high precision specified for the estimates.
*** The package can arrive at a less precise solution more
*** quickly: e.g. try prec(0.05) instead of prec(0.005).

set seed 210815

simsam s_cohortstep ncluspergrp,  ///
      assuming(sd(5) icc(0.33) cac(0.9) iac(0.7)  ///
            clussize(10) nstep(3) timecoeff(0))  ///
      detect(treatcoeff(2))  ///
      null(treatcoeff(0))  ///
      p(.8) start(2) inc(1) prec(0.005)
```

2. Output

```
-----------------------------------------------------
iteration nclusp~p              power (99% CI)
-----------------------------------------------------
        1        2 ..... 0.6100 (0.4765, 0.7327)
        2        4 ..... 0.8910 (0.8632, 0.9150)
        3        4 ..... 0.8859 (0.8775, 0.8940)
        4        4 ..... 0.8908 (0.8868, 0.8946)
        5        3 ..... 0.7922 (0.7871, 0.7973)
-----------------------------------------------------
     null        4 ..... 0.0584 (0.0532, 0.0640)
-----------------------------------------------------

   ncluspergrp = 4
        achieves 89.08% power (99% CI 88.68, 89.46)
          at the 5% significance level
    to detect
    treatcoeff = 2
     assuming
            sd = 5
           icc = 0.33
           cac = 0.9
           iac = 0.7
      clussize = 10
         nstep = 3
     timecoeff = 0

     under null:  5.84% power (99% CI  5.32,  6.40)

If continuing, use prec/inc < 4.9e-02
```