Sample size and power calculations for open cohort longitudinal cluster randomised trials

Jessica Kasza, {Richard Hooper, Andrew Copas, Andrew B Forbes}

October 9, 2019

Abstract

When calculating sample size or power for stepped wedge or other types of longitudinal cluster randomised trials, it is critical that the planned sampling structure be accurately specified. One common assumption is that participants will provide measurements in each trial period, i.e. there is a closed cohort of study participants. Another common assumption is that each participant provides only one measurement during the course of the trial. However, as has been pointed out by several authors, including Copas et al. (2015), some studies have an "open cohort" sampling structure, where participants may provide variable numbers of measurements. To date, sample size and power calculations for longitudinal cluster randomised trials have not accommodated open cohort sampling structures. Some guidance was provided in Feldman and McKinlay (1994), who stated that the participant-level autocorrelation could be varied to account for the degree of overlap in samples in different periods of the study. However, precisely how this quantity should be varied was not discussed.

We present formulas for sample size and power calculations that allow for the open cohort design, and discuss the impact of the degree of "openness" on sample size and power. We consider open cohort designs where the number of participants in each cluster is expected to be maintained throughout the trial, but individual participants may provide differing numbers of measurements. Our results are a unification of Hooper et al's (2016) results for closed cohorts and repeated cross-sectional sample results, and indicate precisely how Feldman and McKinlay's participant autocorrelation should be varied to account for an open cohort sampling structure. We also discuss different types of open cohort sampling schemes and how an open cohort sampling structure impacts on power in the presence of decaying within-cluster correlation structures and autoregressive participant-level errors.

Keywords: cluster randomised trial, intra cluster correlation, open cohort, power, sample size, stepped wedge

1 Introduction

Cluster randomised trials are randomised trials in which clusters of participants, rather than the participants themselves, are randomised to particular treatments [Murray, 1998]. Longitudinal cluster randomised trials extend standard cluster randomised trials in time: clusters are now randomised to a sequence of treatments, and may switch between intervention and control conditions over the course of the trial [Hooper et al., 2016, Hemming et al., 2019]. Particular examples of such trials include cluster randomised cross-over trials [Arnup et al., 2017], stepped wedge trials [Hussey and Hughes, 2007], or even parallel cluster trial designs in which measurements are taken at several time points throughout the trial. Figure 1 displays the schematic for an example stepped wedge trial with three treatment sequences. It is well known that the grouping of participants within clusters induces dependence between the measurements taken on different participants within the same cluster. This dependence increases the sample size over that which would be required to detect an effect of the same size in an individually-randomised trial [Murray, 1998]. However, longitudinal cluster randomised trials such as the stepped wedge can lead to a reduction in this inflation, by allowing for comparisons within clusters as well as between clusters [Hooper et al., 2016, Matthews and Forbes, 2017].

[Figure 1 about here.]

When calculating required sample sizes or the power of longitudinal cluster randomised trials, whether individual participants are measured only once or multiple times (once in each of a number of distinct trial periods) must be accounted for. To date, sample size calculations for longitudinal cluster randomised trials have assumed either a closed cohort sampling structure, where all participants contribute measurements in all periods of the trial, or that each participant provides only one measurement. However, as pointed out by Copas et al. [2015] in the context of stepped wedge designs, some stepped wedge designs have "open cohort" sampling schemes, where the number of measurements provided by each participant may vary: some participants may provide multiple measurements, and others only one. An example of a stepped wedge design with an open cohort sampling scheme is Tesky et al. [2019]: in that trial, nursing homes are the clusters, and residents of the nursing home are recruited to participate in the study. Recruited participants are replaced with new participants if they leave the nursing home, thus maintaining the same sample size in each cluster in each period of the study. Other studies may involve sampling a fixed number of participants from clusters in each period, for example when the clusters are large communities. When repeated samples are taken from finite populations in this way it is possible that some participants are sampled and provide measurements in more than one period.

Feldman and McKinlay [1994] discussed the possibility of open cohort designs, therein referred to as designs with random overlap, and indicated that the participant autocorrelation (the correlation between mean values of a participant measured at two different time points) could be varied to account for the

degree of overlap (the degree of cohort "openness"). In this paper we show precisely how this participant autocorrelation should be varied for open cohort sampling structures, and that this depends on the expected proportion of participants that will be observed in pairs of treatment periods. We provide sample size formulas for open cohort longitudinal cluster randomised trials, with particular emphasis on the stepped wedge design. For the Hussey and Hughes and block-exchangeable within-cluster correlation structures we provide a design effect that unifies the closed cohort and single-measurement-per-participant design effects of Hooper et al. [2016]. We discuss different open cohort sampling schemes and the impact of open cohorts on sample size calculations when the correlations of measurements taken from the same or different participants decay the further apart in time measurements are taken. Readers can explore our results in an online app written using R Shiny [Chang et al., 2017], available at `https://monash-biostat.shinyapps.io/OpenCohort/`.

2 Sample size formulas for open cohort longitudinal cluster randomised trials

2.1 Model for open cohort cluster randomised trials

We initially consider a model for a continuous outcome with the block exchangeable within-cluster correlation structure: this structure implies that participants measured in the same cluster and the same period of a study have outcomes that are more highly correlated than those of participants measured in the same cluster but in different study periods. While we suppose that the same number of participants are measured in each cluster-period cell of the trial, we do not necessarily suppose that all of the participants provide measurements in all of the periods. Letting Y_{kti} be the outcome for participant i in period t in cluster k,

$$\begin{aligned} &Y_{kti} = \beta_t + \theta X_{kt} + C_k + CP_{kt} + \eta_{ki} + \epsilon_{kti}, \\ &\eta_{ki} \sim N(0, \sigma^2_\eta), \quad \epsilon_{kti} \sim N(0, \sigma^2_\epsilon), \quad C_k \sim N(0, \sigma^2_C), \quad CP_{kt} \sim N(0, \sigma^2_{CP}) \end{aligned} \tag{1}$$

where participant $i = 1, \ldots, m$, period $t = 1, \ldots, T$, cluster $k = 1, \ldots, K$. Fixed effects for each period are included (the β_t), and previous work has shown that for many longitudinal cluster-period trials in which all clusters provide measurements in all periods, the variance of the treatment effect estimator (the key ingredient in sample size and power calculations) is invariant to several choices of parameterisation of these time effects [Grantham et al., 2019a]. Participant-level errors ϵ_{kti} are assumed to be normally distributed, and the participant-level random effect η_{ki} allows for dependence between multiple measurements on the same participant. Cluster-level random effects C_k and cluster-period level random effects CP_{kt} allow for the correlations between participants measured in the same cluster and the same period to differ from the

correlations between participants in the same cluster but different periods. We consider more complex within-cluster correlation structures and autoregressive participant-level errors in Section 2.4.

We require that m participants are included in each period in each cluster, however, we do not require that all participants contribute measurements in all periods: there is expected to be some flow of participants into and out of each cluster at each period. Such a situation may be expected in longitudinal cluster randomised trials conducted in schools or residential care facilities, or when cluster members are sampled at each period. In these settings, clusters are expected to maintain a relatively stable cluster size throughout the trial duration, but some participants may leave their cluster and be replaced by new participants during the trial. We do not consider the situation in which participants move from one trial cluster to another: in this situation, there would no longer be independence of outcomes between clusters. We suppose that missing observations from participants who do not provide measurements in all periods are missing completely at random. This implies that we do not consider the implications of informative participant departure, where the very fact that a participant is no longer contributing measurements may provide information about those measurements, where past measurements may be informative of participant departure, or where survivor average causal effects may be of interest.

Collapsing to cluster-period means, $\bar{Y}_{kt\bullet} = \frac{1}{m}\sum_{i=1}^{m} Y_{kti}$, gives:

$$\begin{aligned}\bar{Y}_{kt\bullet} &= \beta_t + \theta X_{kt} + C_k + CP_{kt} + \eta_{k\bullet} + \epsilon_{kt\bullet},\\ &\eta_{k\bullet} \sim N(0, \sigma_\eta^2/m), \quad \epsilon_{kt\bullet} \sim N(0, \sigma_\epsilon^2/m), \quad C_k \sim N(0, \sigma_C^2), \quad CP_{kt} \sim N(0, \sigma_{CP}^2).\end{aligned} \tag{2}$$

Considering the variances and covariances of cluster-period means shows how this model depends on the open cohort sampling structure:

$$var\left(\bar{Y}_{kt\bullet}\right) = \sigma_C^2 + \sigma_{CP}^2 + \frac{\sigma_\eta^2}{m} + \frac{\sigma_\epsilon^2}{m}, \quad cov\left(\bar{Y}_{kt\bullet}, \bar{Y}_{ks\bullet}\right) = \sigma_C^2 + \sigma_\eta^2 \frac{n_k(t,s)}{m^2}$$

where $n_k(t,s)$ is the number of participants in cluster k that provide measurements in both periods t and s, $n_k(t,s) = n_k(s,t)$, $n_k(t,s) \leq m$ for all period pairs t, s, and $n_k(t,t) = m$. In order to be valid, each triple of periods t, s, and u must have overlapping numbers of participants that satisfy the following inequality: $n_k(t,u) + n_k(u,s) \leq n_k(t,s) + m$. A proof of this requirement is provided in the Appendix.

In some situations researchers may be aware of how many participants are expected to provide measurements in all pairs of periods of a trial, and may be able to specify $n_k(t,s)$ for all clusters k and period pairs (t,s). However, when there is uncertainty regarding which participants will be present in each pair of trial periods, researchers may instead have some idea of the rate of participant retention, or equivalently, of participant attrition. The rate of attrition is sometimes referred to as the churn rate, where the churn rate

from period t to period s in cluster k is the proportion of participants in period t who do not also appear in period s:

$$\chi_k(t,s) = 1 - \frac{n_k(t,s)}{m} = 1 - r_k(t,s),$$

where $r_k(t,s)$ is the retention rate in cluster k between periods t and s. The covariance between any pair of cluster-period means depends on the churn rate:

$$cov\left(\bar{Y}_{kt\bullet}, \bar{Y}_{ks\bullet} | \chi_k(t,s)\right) = \sigma_C^2 + \sigma_\eta^2 \frac{1}{m}(1 - \chi_k(t,s)). \tag{3}$$

In some situations, it is reasonable to suppose that $\chi_k(t,s) = \chi_k$ for all period pairs t, s. In the next subsection we will discuss when this assumption will be appropriate through a discussion of open cohort sampling processes.

2.2 Open cohort sampling schemes

There are many different ways in which an open cohort sampling scheme can be realised, and here we discuss three such schemes. Figure 2 displays four different schemes for a four-period design. We will discuss each scheme in greater detail, but first summarise each briefly: the "core group" scheme involves a core group of participants in each cluster who provide measurements in each of the periods of the study, complemented by participants who provide measurements in only one period; the "closed population" scheme involves repeated sampling from a closed population of potential study participants; and "rotation sampling" schemes place an upper limit on the number of consecutive periods in which participants will provide measurements and specify a replacement fraction in each period. The replacement fraction is the proportion of participants who will be be replaced by new participants at the start of each new study period.

[Figure 2 about here.]

In the core group scheme, the churn rate is constant for each pair of periods, and can take values $\chi_k(t,s) = \chi_k \in \{0, 1/m, 2/m, \ldots, (m-1)/m, 1\}$. Such a scheme may be appropriate when calculating sample sizes or power for trials taking place in schools or nursing homes, where it is expected that most participants remain in the school or nursing home for the entire duration of the trial, while some may only be present in one trial period. The closed population scheme will be appropriate whenever taking repeated samples from each cluster in each trial period, where some participants may be sampled in multiple periods. Since the sampling is random, at the planning stage of the trial, the expected churn rate is most informative, and will be constant for any pair of periods, $E[\chi_k(t,s)] = E[\chi_k]$. If the total population is of size M, the expected overlapping number of participants between any two periods will be $\frac{m^2}{M}$, giving an expected churn rate of $E[\chi_k] = 1 - \frac{m}{M}$.

The rotation sampling scheme has been explored in the context of the design of surveys conducted over multiple time periods, where repeated samples are taken from some population [Steel and McLaren, 2009]. We consider rotation designs where each participant provides measurements in a maximum of p periods each, with $1/p$ participants in period t being replaced by new participants in period $t+1$. This has been referred to as the "in-for-p" design in the context of surveys. Rotation sampling scheme 1 in Figure 2 has $p = 2$: of the m participants who provide measurements in period 1, half will also provide measurements in period 2 (group A), while the other half (group B) will be replaced by group C. Rotation sampling scheme 2 has $p = 3$.

For such rotation sampling schemes as this, the churn rate is non-constant across period pairs. For a "in-for-p" rotation sampling scheme,

$$\chi_k(t,s) = \frac{|t-s|}{p} \text{ for } |t-s| \leq p \text{ and } 1 \text{ for } |t-s| > p.$$

Other more complex rotation sampling schemes are possible, and have been described in the context of repeated surveys in Steel and McLaren [2009]. For example, participants could be sampled for p periods, excluded for p' periods, and then return for p'' periods.

For core group or closed population schemes, the churn rate does not depend on the length of time between periods, but may differ between clusters. However, if researchers expect that the churn of participants will be similar across clusters, then $\chi_k = \chi$ can be substituted into Equation (3). Alternatively, researchers may instead expect that the churn rate associated with a given cluster is drawn from some distribution of churn rates, with some clusters having greater churn than others. If all χ_k are independent and identically distributed with some probability density function $f_X(\chi)$, this implies that

$$cov\left(\bar{Y}_{kt\bullet}, \bar{Y}_{ks\bullet}\right) = \sigma_C^2 + \sigma_\eta^2 \frac{1}{m}(1 - E[\chi_k]).$$

Although the churn rate is, strictly speaking, a discrete random variable, if m is large enough, researchers could suppose that the churns follow a Beta distribution, with first parameter α equal to the number of participants expected to be lost from one period to the next averaged over all clusters, and the second parameter β equal to the number of participants retained, again averaged over all clusters. We are assuming that the total number of participants in each cluster-period is constant, so that $\alpha + \beta = m$. In that case, $\chi_k \sim Beta(\alpha, \beta)$, with $E[\chi_k] = \frac{\alpha}{\alpha+\beta}$ and

$$cov\left(\bar{Y}_{kt\bullet}, \bar{Y}_{ks\bullet}\right) = \sigma_C^2 + \sigma_\eta^2 \frac{1}{m} \frac{\beta}{\alpha+\beta}.$$

The Beta distribution is a convenient choice since it is bounded by 0 and 1, however, all that is required for sample size calculations is the specification of the expected churn rate.

2.3 Design effects

Hooper et al. [2016] provided design effects for longitudinal cluster randomised trials where participants provide either one measurement only or one measurement in each period of a design (a closed cohort). Here we extend those design effects to incorporate the open cohort sampling scheme when $\chi_k(t,s)$ can be replaced by a constant χ. Following Hooper et al. [2016] we define the following parameters:

$$\sigma^2 = \sigma_C^2 + \sigma_{CP}^2 + \sigma_\eta^2 + \sigma_\epsilon^2, \quad \rho = \frac{\sigma_C^2 + \sigma_{CP}^2}{\sigma^2}, \quad \pi = \frac{\sigma_C^2}{\sigma_C^2 + \sigma_{CP}^2}, \quad \tau = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\epsilon^2}. \tag{4}$$

σ^2 is the total variance, the parameter ρ is the usual intra-cluster correlation (the correlation between a pair of participants measured in the same cluster in the same treatment period); π is the cluster autocorrelation (the correlation between two population means from the same cluster measured at different time periods); and τ is the participant autocorrelation (the correlation between two measurements on the same participant in a given cluster).

If θ^* is the treatment effect that the researcher wishes to detect, with power $1-\beta$ and two-sided significance level α, and the $100p$th centile of the normal distribution given by z_p, then standard results imply that the total number of participants required for an individually-randomised trial is

$$n_i = 4\frac{\sigma^2}{(\theta^*)^2}\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2. \tag{5}$$

As has been shown in Hooper et al. [2016], for example, the number of clusters (K_P) required for a parallel cluster randomised trial with one measurement taken from each of m participants within each cluster is given by

$$K_P = \left[1 + (m-1)\rho\right]\frac{n_i}{m}, \tag{6}$$

where the quantity $1 + (m-1)\rho$ is the design effect that accounts for the clustering of participants.

To account for multiple measurements per cluster, Hooper et al. [2016] showed that an additional design effect is required (where the parameter r is defined below), given by

$$DE(r) = \frac{1}{4}\frac{K^2(1-r)[1+(T-1)r]}{KX_{\bullet\bullet} - \sum_{t=1}^{T}(X_{\bullet t})^2 + \left[(X_{\bullet\bullet})^2 + K(T-1)X_{\bullet\bullet} - (T-1)\sum_{t=1}^{T}(X_{\bullet t})^2 - K\sum_{k=1}^{K}(X_{k\bullet})^2\right]r} \tag{7}$$

where K is the total number of sequences, and all clusters are assumed to be assigned to their own sequence,

T is the total number of measurement periods, and

$$X_{\bullet\bullet} = \sum_{k=1}^{K}\sum_{t=1}^{T} X_{kt}, \quad X_{\bullet t} = \sum_{k=1}^{K} X_{kt}, \quad X_{k\bullet} = \sum_{t=1}^{T} X_{kt}.$$

For the open cohort design, when $\chi_k(t,s)$ can be replaced by a constant χ,

$$r = \frac{\sigma_C^2 + \sigma_\eta^2(1-\chi)}{\sigma_C^2 + \sigma_{CP}^2 + \frac{\sigma_\eta^2}{m} + \frac{\sigma_\epsilon^2}{m}}.$$

This can be written in terms of the correlation parameters as

$$r = \frac{m\rho\pi + (1-\rho)\tau(1-\chi)}{1+(m-1)\rho}. \tag{8}$$

This unifies the cross-sectional and closed cohort design effects in Hooper et al. [2016]: when $\chi = 0$ the result for closed cohorts is returned, and when $\chi = 1$, the result when each participant provides only one measurement is returned.

In Feldman and McKinlay [1994], a similar unifying model that encompasses cross-sectional and closed cohort designs was presented. In that, the authors stated that by allowing the participant autocorrelation (which we have here denoted by τ) to vary, their model allowed for "randomly overlapping samples", and that in the case of overlapping samples, τ will be "small but positive, depending on the degree of the overlap". Our result shows exactly how the participant autocorrelation should be varied to allow for open cohorts: the participant autocorrelation τ must be multiplied by the expected retention rate or overlap between periods (i.e. proportion of participants expected to be present in both of any pair of periods).

For open cohort longitudinal cluster randomised trials, the number of clusters required is thus given by

$$K_L = DE(r) \times [1+(m-1)\rho]\,\frac{n_i}{m}, \tag{9}$$

where n_i is the total number of participants required for an individually-randomised trial, given by Equation (5), m is the number of participants measured in each cluster in each period, ρ is the intra-cluster correlation for a pair of participants measured in the same cluster-period. $DE(r)$ is given by Equation (7), and depends on the design schematic. The parameter r, given in Equation (8), depends on the correlation parameters and the proportion of participants expected to be present in any pair of periods. Dividing K_L by the number of sequences in the design and rounding up to the nearest integer then gives the minimum number of clusters per sequence required to reach the desired level of power.

2.4 Incorporating decays in between-period correlations and participant-level correlations

We consider a model allowing for more general within-cluster and within-participant correlation structures. This model includes cluster-period random effects and correlated participant-level errors, and has the following form:

$$
\begin{aligned}
&Y_{kti} = \beta_t + \theta X_{kt} + CP_{kt} + \epsilon_{kti}, \\
&\epsilon_{ki} = (\epsilon_{k1i}, \ldots, \epsilon_{kTi})^T \sim N(\mathbf{0}, \sigma^2_{\epsilon D} D_{\epsilon,i}), \quad CP_k = (CP_{k1}, \ldots, CP_{kT})^T \sim N(\mathbf{0}, \sigma^2_{CP,D} D_{CP}),
\end{aligned} \tag{10}
$$

where $D_{\epsilon,i}$ and D_{CP} are symmetric $T \times T$ matrices with diagonal elements all equal to 1. If participant i provides measurements in only T_i periods of the design, then $D_{\epsilon,i}$ has dimension $T_i \times T_i$. We suppose that the elements of $D_{\epsilon,i}$ are common across participants: if both participant i and j provide measurements in periods t and s, then $D_{\epsilon,i}(t,s) = D_{\epsilon,j}(t,s)$, and we remove the participant subscript on D_ϵ. If $D_\epsilon(t,s) = \tau$ and $D_{CP}(t,s) = \pi$ for $t \neq s$ and some constants τ and π (analogous to the participant and cluster autocorrelations in Equation (4)), then Model (1) is returned. Autoregressive errors at the participant level can be obtained by setting $D_\epsilon(t,s) = \tau_D^{|t-s|}$, and the discrete-time decay model of Kasza et al. [2019a] is returned if $D_{CP}(t,s) = \pi_D^{|t-s|}$ for constants τ_D and π_D. Li [2019] presented a similar model for closed-cohort longitudinal cluster randomised trials with autoregressive participant-level errors and decaying between-period correlations.

Collapsing to cluster-period means gives

$$
\begin{aligned}
Y_{kt\bullet} &= \beta_t + \theta X_{kt} + CP_{kt} + \epsilon_{kt\bullet}, \quad CP_k = (CP_{k1}, \ldots, CP_{kT})^T \sim N(0, \sigma^2_{CP,D} D_{CP}), \\
\epsilon_{kt\bullet} &= \frac{1}{m}\sum_{i=1}^{m} \epsilon_{kti}, \quad var(\epsilon_{kt\bullet}) = \frac{\sigma^2_{\epsilon D}}{m}, \quad cov(\epsilon_{kt\bullet}, \epsilon_{ks\bullet}) = \frac{n_k(t,s)}{m^2}\sigma^2_{\epsilon D} D_\epsilon(t,s).
\end{aligned}
$$

As has been shown previously (e.g Kasza et al. [2019b]), the variance of the generalised least squares estimator of θ is given by:

$$
var\left(\hat{\theta}\right) = \left\{ \sum_{k=1}^{K} \mathbf{X}_k^T var(\bar{\mathbf{Y}}_k)^{-1} \mathbf{X}_k - \sum_{k=1}^{K} \mathbf{X}_k^T var(\bar{\mathbf{Y}}_k)^{-1} \left[\sum_{k=1}^{K} var(\bar{\mathbf{Y}}_k)^{-1} \right]^{-1} \sum_{k=1}^{K} var(\bar{\mathbf{Y}}_k)^{-1} \mathbf{X}_k \right\}^{-1},
$$

where $\mathbf{X}_k^T = (X_{k1}, \ldots, X_{kT})$ is the vector of treatment assignments for cluster k, and $var(\bar{\mathbf{Y}}_k)$ is the $T \times T$ is the variance matrix of the vector $\bar{\mathbf{Y}}_k = (\bar{Y}_{k1\bullet}, \ldots, \bar{Y}_{kT\bullet})^T$ with elements

$$
var(\bar{\mathbf{Y}}_{kt}) = \sigma^2_{CP,D} + \frac{\sigma^2_{\epsilon D}}{m}, \quad cov(\bar{\mathbf{Y}}_{kt}, \bar{\mathbf{Y}}_{ks}) = \sigma^2_{CP,D} D_{CP}(t,s) + \frac{n_k(t,s)}{m^2}\sigma^2_{\epsilon D} D_\epsilon(t,s).
$$

The online app allows calculation of sample size and power for the discrete-time decays in correlations

of cluster and participant random effects, with $D_{CP}(t,s) = \pi_D^{|t-s|}$, and $D_{\epsilon}(t,s) = \tau_D^{|t-s|}$ for constant churn rates. The quantities π_D and τ_D are analogous to the parameters π (the cluster autocorrelation) and τ (the participant autocorrelation) in Equation (4). However, π_D and τ_D now represent the decay in correlation between cluster or participant random effects for measurements only one period apart in time, rather than the decay in correlation for any pair of measurements from different periods. Users also input the total variance $\sigma_D = \sigma^2_{CP,D} + \sigma^2_{\epsilon D}$ and the intra-cluster correlation for a pair of measurements in the same cluster in the same period, $\rho_D = \frac{\sigma^2_{CP,D}}{\sigma^2_{CP,D}+\sigma^2_{\epsilon D}}$.

2.5 Sample size and power for rotation "in-for-p" open cohort sampling schemes

When the open cohort sampling scheme has an in-for-p structure, $var(\bar{\mathbf{Y}}_k)$ is the $T \times T$ is the variance matrix of the vector $\bar{\mathbf{Y}}_k = (\bar{Y}_{k1\bullet}, \ldots, \bar{Y}_{kT\bullet})^T$ with elements

$$var(\bar{\mathbf{Y}}_{kt}) = \sigma^2_{CP,D} + \frac{\sigma^2_{\epsilon D}}{m}, \quad cov(\bar{\mathbf{Y}}_{kt}, \bar{\mathbf{Y}}_{ks}) = \sigma^2_{CP,D} D_{CP}(t,s) + \frac{1}{m}\sigma^2_{\epsilon D} D_{\epsilon}(t,s) \mathbb{1}(|t-s| \leq p)\left(1 - \frac{|t-s|}{p}\right)$$

where $\mathbb{1}(|t-s| \leq p)$ is the indicator function for the event $|t-s| \leq p$. The online app allows calculation of sample size and power when in-for-p sampling schemes are of interest. Users can select the sampling scheme, and when selecting in-for-p, the power or sample size for values of $p = 1, \ldots, T$, where T is the number of periods in the user-input design are graphed.

3 Examples of sample size calculations with open cohorts

3.1 Girls on the Go! example

We consider a specific example inspired by the closed-cohort example described in Hooper et al. [2016]: a stepped wedge trial conducted in Australian primary schools to evaluate the "Girls on the go!" programme aimed at increasing the self-esteem of young women [Tirlea et al., 2013]. The primary outcome was the Rosenberg Self-esteem scale, a continuous measure. Following Hooper et al. [2016], we assume an intracluster correlation of $\rho = 0.33$, a cluster autocorrelation of $\pi = 0.9$, an individual autocorrelation of $\tau = 0.7$, a total variance of 25 units, and a mean difference of interest of 2 units. The standard three-sequence stepped wedge design was implemented, as shown in Figure 1, with two schools assigned to each of the three sequences in the original trial, with 10 students enrolled in each school. We suppose here that four schools were assigned to each of the three sequences: Hooper et al. [2016] showed that such a study would have a power of 89.3%. In reality, this study had a closed cohort sampling scheme, but we investigate the impact of a core group open cohort sampling scheme, assuming that the core group made up 0%, 10%, 20%, …, 100% of the sample in each cluster.

In addition to considering the theoretical power of the "Girls on the go!" study, we also simulate study power when different clusters may have a core group that makes up a slightly different proportion of that cluster's samples. For each expected core group proportion of $r = 0.1, 0.2, \ldots, 0.9$ the proportion of participants made up of the core group is simulated from a Beta distribution with $\alpha = 10 \times (1 - r)$ and $\beta = 10 \times r$; simulations were also conducted for expected core group proportion of $r = 0$ (each participant providing one measurement only) and $r = 1$ (closed cohort). For each value of r, 1000 data sets were simulated and analysed. For each r, the proportion of times the null hypothesis was rejected using a two-sided significance level of 5% was calculated.

Figure 3 displays both the theoretical power and simulated power for the "Girls on the go!" study. As the expected core group proportion increases, so too does the power of the study. This is to be expected since the estimator for the treatment effect that we consider (the generalised least squares estimator) combines both within-cluster and between-cluster comparisons [Matthews and Forbes, 2017]. When the core group proportion increases, within-cluster comparisons contribute increasing amounts of information about the treatment effect.

[Figure 3 about here.]

3.2 Incorporating decaying correlations

We extend the "Girls on the go!" example to include decaying correlations and autoregressive participant-level errors, and consider the impact of participant retention rate on power when there is no decay (the scenario considered in Section 3.1), a decay in the participant-level correlation only, when there is a decay in the cluster-level autocorrelation only, and when there is decay in both the participant- and cluster-level correlations. As above, we consider a three-sequence stepped wedge design with four clusters assigned to each of the three sequences, and 10 participants in each cluster in each period, and a total variance of 25 units with a mean difference of interest of 2 units.

For all four scenarios, the assumed intra-cluster correlation is given by $\rho_D = 0.33$, however, the cluster autocorrelation and participant autocorrelation selected depend on whether there is supposed to be decay in those correlations. The values $\tau = 0.7$ and $\pi = 0.9$ in Section 3.1 were specified under the assumption that these autocorrelations would specify the decay for any pair of periods, no matter how far apart in time these may be. However, as has been shown in Kasza and Forbes [2018], if there is a decay in correlations over time, and a model that does not allow for such decay is specified, the estimate of the cluster autocorrelation will account for the decay that was present in the dataset. Hence, when accounting for a decay in correlations, it is not sufficient to assume the same values of τ and π that were assumed when calculating sample sizes or power when there was no decay. Extending the formulas in Kasza and Forbes [2018], we derive adjusted

values of the cluster and participant autocorrelations. This amounts to solving the equations

$$\sum_{t=1}^{T}\sum_{s=1}^{S}\tau_D^{|t-s|} = \tau T(T-1)+T \text{ and } \sum_{t=1}^{T}\sum_{s=1}^{S}\pi_D^{|t-s|} = \pi T(T-1)+T$$

for τ_D and π_D. Doing this for $T = 4$, $\tau = 0.7$ and $\pi = 0.9$ gives $\tau_D = 0.80$ and $\pi_D = 0.94$. Only when decaying cluster-level autocorrelations are incorporated is the value π_D is assumed, and only when decaying participant-level autocorrelations are incorporated is the value τ_D; otherwise, π and τ are included in calculations.

[Figure 4 about here.]

Figure 4 displays the power for each of the four considered correlation structures (no decay; decay in participant autocorrelations only; decay in cluster autocorrelations only; decay in both participant and cluster autocorrelations) for core group proportions from 0 to 1. For all correlation structures, the power increases as the core group proportion increases, with the steepest increases occurring when there is decay in the participant autocorrelation. When a decay in participant autocorrelation is included, the correlation between measurements in successive periods is greater than when there is no decay ($\tau_D = 0.8$ versus $\tau = 0.7$). The greater the autocorrelation between successive measurements on the same participant, the more information there is in the comparison of outcomes from that participant measured under control and intervention conditions. When the core group proportion is higher, the more participants there are that provide measurements under both control and intervention conditions in successive periods, and thus the power to detect a given effect size increases.

Similarly, power is greater when cluster autocorrelations decay over time than when there is no decay, for all core group proportions: this is the case because π_D is greater than π, so when a decaying cluster autocorrelation is accounted for, measurements in the same cluster in adjacent periods are more highly correlated than when there is no decay ($\pi_D = 0.94$ versus $\pi = 0.9$).

4 Discussion

In this paper we have presented formulas for sample size and power calculations for open cohort longitudinal cluster randomised trials, where participants may provide varying numbers of measurements. Design effects were provided for the model with a block-exchangeable within-cluster correlation structure, and a formula for the variance of the treatment effect estimator was provided for when the within-cluster correlation structure is more complex. The design effect unifies the closed cohort and single-measurement design effects provided in Hooper et al. [2016]. We have also provided an online app to allow readers to investigate the impact of varying degrees of cohort openness on the power of their planned studies.

For designs in which some or all clusters switch between treatments, the conservative assumption is that each participant provides one measurement only: this will always lead to larger sample sizes. Hence, researchers may be tempted to conservatively assume a retention rate of 0 (i.e. completely non-overlapping samples at each study period). However, when planning studies, researchers should carefully consider the ethical implications of exposing participants to involvement in a clinical trial, and use a value of the retention rate that accurately reflects what is expected to happen during the trial.

Assuming a common expected retention rate across clusters that does not depend on the time between periods leads to closed-form sample size formulas. In many situations, such as core group and closed population sampling schemes, we would expect that such an assumption would be adequate, but further work is required to assess the impact of varying retention rates. For example, rotation sampling schemes imply that the churn between two periods, $\chi_k(t,s)$ depends on the amount of time between periods, $|t-s|$. Other sampling schemes may also lead to an increase in churn over time. We have only considered three different types of sampling schemes possible in open cohort designs: the core group, closed population, and rotation sampling schemes. Other sampling schemes are indeed possible, and have been explored at length in the repeated survey literature. It seems that further research into the applicability of alternative open cohort sampling schemes in the context of longitudinal cluster randomised trials is necessary. When researchers wish to minimise the burden on participants, rotation sampling schemes may be a good choice, but further work on these and related schemes is required to determine the impact of changing the number of measurements on subjects for various types of longitudinal cluster randomised trials.

Researchers reporting open-cohort longitudinal cluster randomised trials should be encouraged to report the number of participants overlapping for each pair of periods in each cluster, or at least some estimate of the retention rate. If different clusters have different expected rates of retention, upper and lower bounds on required sample sizes can be obtained by assuming the lowest and the highest expected retention rate across all clusters. We also assumed that the number of participants was the same in each cluster-period. If the number of participants differed across cluster-periods, required sample sizes will be inflated, as for the closed-cohort and single measurement sampling structures [Eldridge et al., 2006].

The within-cluster correlation structures we have considered depend on treatment periods, rather than on the specific trial entry time of each participant: time is treated as a discrete phenomenon, taking values $1, 2, \ldots, T$. Recent papers have discussed time as a continuous phenomenon in longitudinal cluster randomised trials, where participants have outcomes that are measured in continuous time, rather than at a set of discrete time points common to all participants. Grantham et al. [2019b] discussed within-cluster correlation structures in the context of continuous time, and Hooper and Copas [2019] discussed the need to clarify sampling schemes and the terminology used to refer to specific sampling schemes. If participants can have their observations recorded at any time, rather than at a set of discrete times common to all participants, then

the correlation structures we have assumed for cluster-level random effects are not likely to be satisfactory: these correlation structures imply that all participants within a period are exchangeable, and that any pair of participants measured in a period are more highly correlated than any pair of participants measured in distinct periods. However, it may be more plausible to assume that participants measured at the start and end of a period have outcomes that are less correlated than participants who are measured at the end of one period and the start of the next. Correlation structures such as that described in Grantham et al. [2019b] imply that outcomes from the same cluster and in the same period are no longer exchangeable. When there is no longer exchangeability within periods, precisely which participants provide measurements in each pair of periods, rather than just the number overlapping in each pair of periods, becomes important for sample size calculations.

In this paper we have provided design effects and sample size formulas for open cohort sampling structures, unifying previous work which provided separate results for closed cohort sampling structures and single-measurement structures. We have considered three different types of open cohort sampling schemes, but there are likely many more that trialists may find useful. Future work should consider alternative open cohort sampling schemes, and the questions of participants moving between clusters and informative participant departure from or entry into clusters. Further, the impact of treatment effect heterogeneity across clusters, and other correlation structures that depend on treatment periods, could be considered in the context of open cohorts.

References

S. Arnup, J. E. McKenzie, K. Hemming, D. Pilcher, and A. B. Forbes. Understanding the cluster randomised crossover design: a graphical illustration of the components of variation and a sample size tutorial. Trials, 18(1):381, 2017.

W. Chang, J. Cheng, J. J. Allaire, Y. Xie, and J. McPherson. shiny: Web application framework for R. https://cran.r-project.org/package=shiny. R package version 1.0.5, 2017.

A. J. Copas, J. J. Lewis, J. A. Thompson, and et al. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. Trials, 16:1–12, 2015.

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

H. A. Feldman and S. M. McKinlay. Cohort versus cross-sectional design in large field trials: Precision, sample size, and a unifying model. Statistics in Medicine, 13, 1994.

K. Grantham, S. Heritier, A. B. Forbes, and J. Kasza. Time parameterisations in cluster randomized trial planning. The American Statistician, page doi: 10.1080/00031305.2019.1623072, 2019a.

K. Grantham, J. Kasza, S. Heritier, K. Hemming, and A. B. Forbes. Accounting for a decaying correlation structure in cluster randomised trials with continuous recruitment. Statistics in Medicine, 38(11):1918–1934, 2019b.

K. Hemming, J. Kasza, R. Hooper, A. Forbes, and M. Taljaard. A tutorial on sample size calculation for cluster randomised multiple-period parallel, cross-overand stepped-wedge trials and introduction to the Shiny CRT calculator. Under review, 2019.

R. Hooper and A. Copas. Stepped wedge trials with continuous recruitment require new ways of thinking. Journal of Clinical Epidemiology, page doi: 10.1016/j.jclinepi.2019.05.037, 2019.

R. Hooper, S. Teerenstra, E. de Hoop, and S. Eldridge. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Statistics in Medicine, 35:4718–4728, 2016.

M. A. Hussey and J. P. Hughes. Design and analysis of stepped wedge cluster randomized trials. Contemporary Clinical Trials, 28:182–191, 2007.

J. Kasza and A. B. Forbes. Inference for the treatment effect in multiple-period cluster randomised trials when random effect correlation structure is misspecified. Statistical Methods in Medical Research, page doi: 10.1177/0962280218797151, 2018.

J. Kasza, K. Hemming, R. Hooper, J. N. S. Matthews, and A. B. Forbes. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. Statistical Methods in Medical Research, 28(3):703–716, 2019a.

J. Kasza, M. Taljaard, and A. B. Forbes. Information content of stepped wedge designs when treatment effect heterogeneity and/or implementation periods are present. Statistics in Medicine, in press, 2019b.

F. Li. Design and analysis considerations for stepped wedge cluster randomised trials with decayed correlation. arXiv preprint, page arXiv:1903.09923, 2019.

J. N. S. Matthews and A. B. Forbes. Stepped wedge designs: insights from a design of experiments perspective. Statistics in Medicine, 36:3772–3790, 2017.

D. M. Murray. Design and analysis of group-randomized trials. Oxford University Press, New York, NY, 1998.

D Steel and C McLaren. Design and analysis of surveys repeated over time. Handbook of statistics, pages 289–313, 2009.

V A Tesky, A Schall, U Schulze, U Stangier, F Oswald, M Knopf, J König, M Blettner, E Arens, and J Pantel. Depression in the nursing home: a cluster-randomized stepped-wedge study to probe the effectiveness of a novel case management approach to improve treatment (the DAVOS project). Trials, 20(1):424, 2019.

L. Tirlea, H. Truby, and T. P. Haines. Investigation of the effectiveness of the "Girls on the Go!" program for building self-esteem in young women: trial protocol. SpringerPlus, 2(1):683, 2013.

Appendix

We here prove that if $n_k(t,s)$ is the number of participants providing measurements in both period t and period s, then $n_k(t,u) + n_k(u,s) \leq n_k(t,s) + m$, where m is the number of participants observed in each cluster in each period. We consider a particular cluster and omit the cluster subscripts k. Suppose that in this cluster, a total of M participants provide measurements, and the vector of length M, $x_t = (x_{t1}, \ldots, x_{tM})^T$ indicates whether each participant provides a measurement in period t: $x_{ti} = 1$ if participant i provides a measurement in period t, and $x_{ti} = 0$ if not.

Define $l(t,s) = \sum_{i=1}^{M} |x_{ti} - x_{si}|$: this counts the number of participants that provide measurements in period t only or period s only, and can be thought of as the distance between periods t and s in terms of participants. $l(t,s)$ is commonly referred to as the taxi-cab metric, and satisfies the triangle inequality:

$$l(t,s) \leq l(t,u) + l(u,s).$$

An equivalent definition of $l(t,s)$ is

$$l(t,s) = \sum_{i=1}^{M} (x_{ti} - x_{si})^2 = \sum_{i=1}^{M} \left(x_{ti}^2 + x_{si}^2 - 2x_{ti}x_{si} \right) = 2m - 2\sum_{i=1}^{M} x_{ti}x_{si} = 2m - 2n(t,s).$$

Applying the triangle inequality to this definition gives the result.

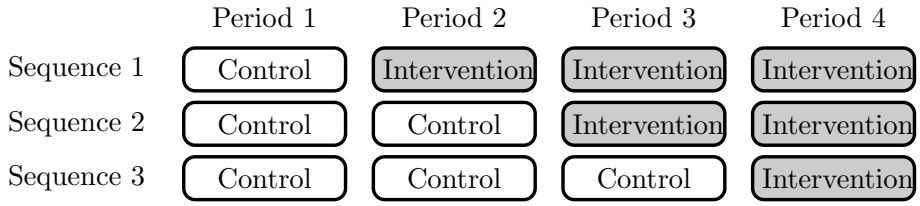

Figure 1: An example stepped wedge schematic, for the stepped wedge design considered in the "Girls on the go!" example in Section 3.1. Multiple clusters may be assigned to each of the treatment sequences.

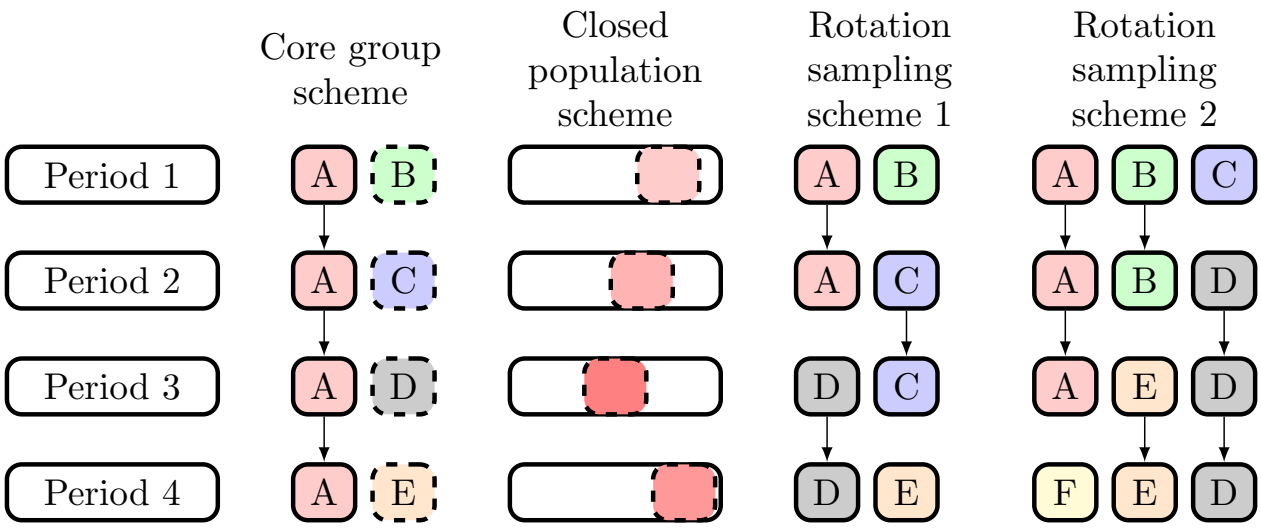

Figure 2: The three variants of open cohort sampling schemes that we will consider, illustrated for a four-period design. Groups of participants measured in multiple periods are denoted with repeated letters and colours. Rotation sampling scheme 1 has an in-for-2 sampling structure, and rotation sampling scheme 3 has an in-for-3 sampling structure.

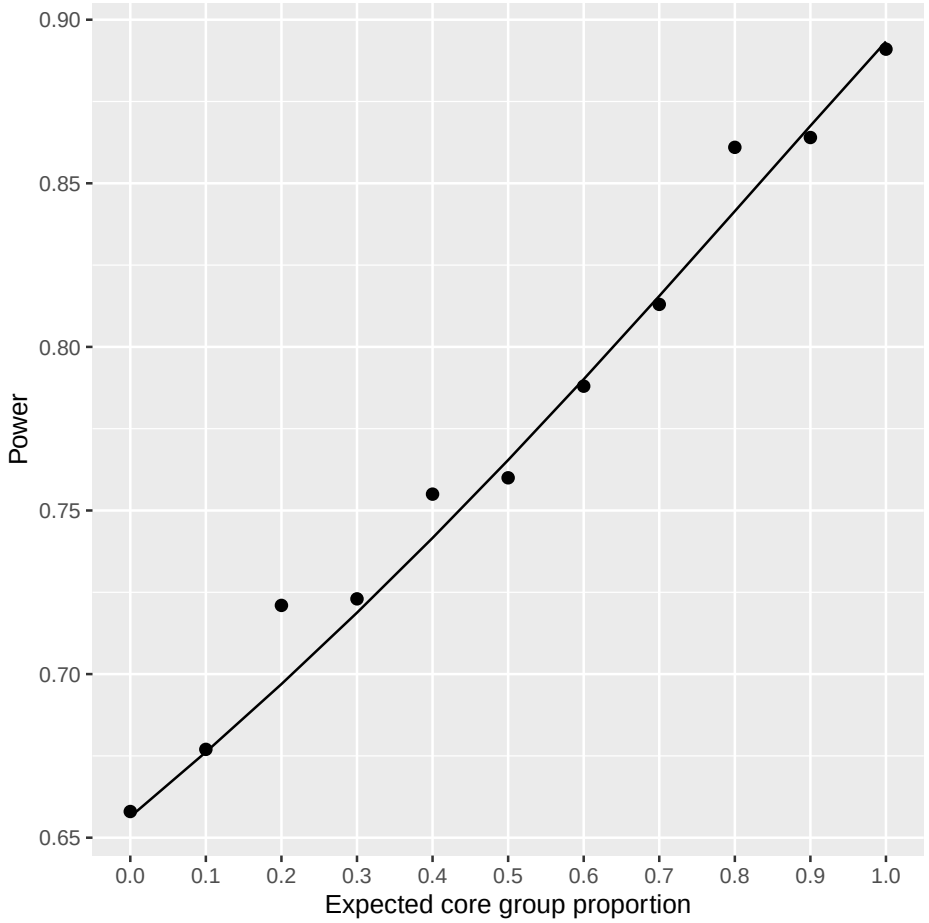

Figure 3: Theoretical (line) and simulated (points) power for the "Girls on the go!" programme, for varying expected core group proportion. When expected core group proportion is 0, each participant provides only one measurement during the trial; when expected core group proportion is 1, each participant provides one measurement in each trial period.

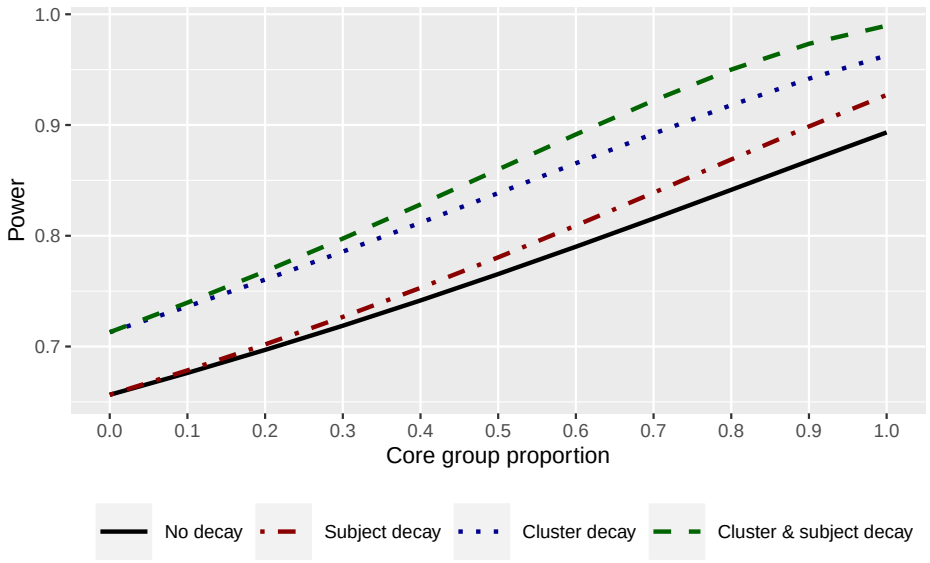

Figure 4: Differences between the theoretical power for the "Girls on the go!" programme without any decay in between-period correlations and participant errors and for models assuming decaying between-period correlations and/or autoregressive participant errors. When expected core group proportion is 0, each participant provides only one measurement during the trial; when expected core group proportion is 1, each participant provides one measurement in each trial period.