Open access • Journal Article • DOI:10.1177/0962280217734981

Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. — Source link

Jessica Kasza, Karla Hemming, Richard Hooper, John N. S. Matthews ...+1 more authors

Institutions: Monash University, University of Birmingham, Queen Mary University of London, Newcastle University

Published on: 01 Mar 2019 - Statistical Methods in Medical Research (Newcastle University)

Topics: Intensive care and Sample size determination

Kasza J, Hemming K, Hooper R, Matthews JNS, Forbes A.

Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials.

***Statistical Methods in Medical Research* (2017)**

DOI:

https://doi.org/10.1177/0962280217734981

Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials

J Kasza[1], K Hemming[2], R Hooper[3], JNS Matthews[4], A Forbes[1] on behalf of the ANZICS Centre for Outcomes & Resource Evaluation (CORE) Committee.

1. Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia
2. School of Health and Population Sciences, University of Birmingham, Birmingham B15 2TT, UK
3. Centre for Primary Care & Public Health, Queen Mary University of London, London, U.K.
4. School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

Corresponding author: Jessica Kasza, Department of Epidemiology and Preventive Medicine, Monash University, Alfred Centre, 99 Commercial Road, Melbourne, Victoria 3004, Australia. Email: jessica.kasza@monash.edu Phone: +61 3 9903 0597

Acknowledgements: This research is supported by Australian National Health and Medical Research Council Centre of Excellence Grant ID 1035261, awarded to the Victorian Centre for Biostatistics (ViCBiostat), and Project Grant ID 1108283.

Abstract word count: 265

Word count: 5,836

Abstract

Stepped wedge and cluster randomised crossover trials are examples of cluster randomised designs conducted over multiple time periods that are being used with increasing frequency in health research. Recent systematic reviews of both of these designs indicate that the within-cluster correlation is typically taken account of in the analysis of data using a random intercept mixed model, implying a constant correlation between any two individuals in the same cluster no matter how far apart in time they are measured: within-period and between-period intra-cluster correlations are assumed to be identical. Recently proposed extensions allow the within- and between-period intra-cluster correlations to differ, although these methods require that all between-period intra-cluster correlations are identical, which may not be appropriate in all situations.

Motivated by a proposed intensive care cluster randomised trial, we propose an alternative correlation structure for repeated cross-sectional multiple period cluster randomised trials in which the between-period intra-cluster correlation is allowed to decay depending on the distance between measurements. We present results for the variance of treatment effect estimators for varying amounts of decay, investigating the consequences of the variation in decay on sample size planning for stepped wedge, cluster crossover and multiple-period parallel-arm cluster randomised trials. We also investigate the impact of assuming constant between-period intra-cluster correlations instead of decaying between-period intra-cluster correlations.

Our results indicate that in certain design configurations, including the one corresponding to the proposed trial, a correlation decay can have an important impact on variances of treatment effect estimators, and hence on sample size and power. An R Shiny app allows readers to interactively explore the impact of correlation decay.

Keywords: exponential decay; intra-cluster correlation; cluster randomised trial; sample size; stepped wedge

Abbreviations: BPICC: between-period intra-cluster correlation; CRT: cluster randomised trial; CRXO: cluster randomised cross over; ICU: intensive care unit; SW: stepped wedge; WPICC: within-period intra-cluster correlation

1. Introduction

Instead of randomly assigning subjects to treatment groups as in an individually-randomised trial, cluster randomised trials (CRTs) randomly assign clusters of subjects to treatment groups, where clusters may be schools, hospitals, geographical regions, or families, for example (1). Given the clustered nature of the data, CRTs will generally require a larger total number of subjects to achieve the same power than required in a comparable individually randomised trial, but may have advantages in situations where individual randomisation is difficult (2). In this paper we consider multiple-period CRTs, where measurements are made on a cluster at multiple pre-defined time points throughout the trial.

There are many types of multiple-period cluster-randomised designs, of which parallel-arm CRTs are the simplest. In parallel CRTs, clusters are randomly allocated to interventions (which we here label treatment or control) at trial commencement. Additionally, parallel CRTs may include one or more pre-intervention measures in each cluster: we refer to such designs as parallel with baseline CRTs. More complex, but with potential gains in power, are designs where clusters may switch between interventions over the course of the trial. Cluster randomised cross-over designs (CRXOs) randomise clusters to receive a sequence of interventions, rather than to receive a particular intervention. In CRXOs, clusters may switch back and forth between interventions one or more times over the course of the trial (3). Stepped wedge (SW) designs are a variant of CRXOs incorporating crossover in one direction only: in the standard SW design, in the first time period all clusters are in the control intervention, and by the last period all clusters are in the treatment intervention (4, 5). All clusters switch from control to treatment at some point over the course of the trial, and will never revert to the control once switched to the treatment. In SWs, it is the time at which the switch from control to treatment occurs that is randomly allocated. More general cluster randomised multi-period designs have been discussed in (6).

SWs have potential gains in power over similar parallel designs in certain circumstances, but are never as powerful as similar CRXOs when intra-cluster correlations are non-zero (4, 7). SWs are particularly useful in situations in which discontinuation of the intervention is difficult or some carry-over of treatment effect in one direction only (from treatment to control) is expected, in which case the CRXO design may not be applicable. The SW design is

also useful to assess the impact of a program destined to be rolled out across all clusters, or when the intervention is seen as desirable, in which case clusters may be more inclined to participate with the promise of eventually receiving the intervention during the course of the trial.

It is now almost universally accepted that cluster randomised trials must acknowledge the effect of clustering at both the design and analysis stage. For single-period parallel CRTs, this acknowledgement is achieved through specification of the intra-cluster correlation, which represents the correlation between two observations in the same cluster. However, in a multiple-period cluster randomised trial, instead of a single intra-cluster correlation, the situation is more complicated: it may be appropriate for correlations between observations from the same cluster to be specified differently if they are in the same or different periods. We use the term "within-period intra-cluster correlation" (abbreviated as WPICC) to describe the correlation between two observations in the same cluster and in the same period; and the term "between-period intra-cluster correlation" (abbreviated as BPICC), to describe the correlation between two observations in the same cluster, but in different time periods. We note that other authors have used different terminology: (8), for example, use the term "intra-cluster correlation" for WPICC, and "inter-period correlation" for BPICC. For example, in CRXOs with two periods, the usual assumption is that the WPICC and BPICC are different, e.g. (8, 9). For CRTs with more than two periods, there may in fact be several BPICCs, corresponding to the correlation between observations taken in each pair of periods. However, the most-widely used model for the design of SW designs trials, the Hussey and Hughes random intercept mixed model for repeated cross-sectional data, implies the equality of within-period and between-period intra-cluster correlations (10, 11).

The implausibility of the assumption that intracluster correlations do not depend on the time between observations has been discussed (e.g. (12-14)), and models proposed in (13) and (4) include separate within-period and between-period intra-cluster correlations, albeit with the assumption of invariance of BPICCs across time. In some situations, it may be more plausible to assume that the correlation between observations from the same cluster *decays* the further apart in time those observations were made. In this paper we consider a more general within-cluster correlation structure, in which the BPICC is allowed to vary

depending on the distance between measurement periods, implying that the BPICC is not constant within clusters.

The specific example we will consider is a proposed CRT in Australian intensive care units (ICUs): data from the Australian and New Zealand Intensive Care Society Adult Patient Database, (15) indicates that an exponential decay within-cluster correlation structure, analogous to an autoregressive structure of order 1, which allows the correlations between observations to decay over time, provides a reasonable approximation to the data. Hence, in the planning of this trial the sensitivity of the required sample size to the usual assumption of a constant correlation within clusters (i.e. the Hussey and Hughes model) requires investigation. We explore the impact of an exponential decaying within-cluster correlation structure on the variance of treatment effects for repeated cross-sectional multiple-period parallel, CRXO and SW CRTs, and hence the impact on required sample sizes to detect given treatment effects. Each subject is present in one time period only, which obviates the need for consideration of participant-level random effects from regression models. Alternative designs, in which subjects may contribute measurements in more than one period, are considered in the Discussion.

In Section 2 we present general formulae for sample size calculations for multiple-period cross-sectional CRTs, first re-stating the main variance expression of Hussey and Hughes (10), and then extending it to more general within-cluster correlation structures. We apply these formulae to explore the consequences of the exponential decay within-cluster correlation structure on the variance of the estimated treatment effect in Section 3. In Section 4 we explore the consequences of the assumption of a constant BPICC structure (e.g. assuming the model of Hussey and Hughes (10), or that of (13) and (4)) instead of an exponential decay BPICC structure on the variance of the estimated treatment effect. Results, extensions, and limitations are discussed in Section 5. Although our results and discussion are framed in terms of parameter estimates corresponding to the proposed ICU trial, we provide an R Shiny web app, (16), at https://monash-biostat.shinyapps.io/NonUniformCorrelation to allow readers to explore the consequences of an exponential-decay within-cluster correlation structure for a range of design choices and parameter values. Users can also up-load their own design matrices. R code for local

implementation of the app is available in the online supplementary material, and at https://github.com/jkasza/NonUniformCorrelation.

2. Sample size calculations for multiple-period CRTs

2.1 Uniform correlation structure

If Y_{itk} denotes the outcome for subject $k = 1, \dots, m$ in cluster $i = 1, \dots, K$, during period $t = 1, \dots T$, the usual Hussey and Hughes model, proposed by (10), is given by

$$Y_{itk} = \mu + \beta_t + X_{it}\theta + \alpha_i + e_{itk}, \qquad \alpha_i \sim N(0, \tau^2), \qquad e_{itk} \sim N(0, \sigma_e^2). \quad (1)$$

We assume that the number of subjects per cluster-period, m, is constant In this model μ is the overall mean outcome, β_t is the fixed effect corresponding to period t, with $\beta_1 = 0$ for identifiability; X_{it} is the intervention variable, equal to 0 (1) if cluster i is in the control condition (treatment) at period t, θ is the intervention effect of interest, α_i is the random effect of cluster i, and e_{itk} is the subject-level residual. As reported in recent systematic reviews, this model is the usual model assumed for cross-sectional stepped-wedge design sample size calculations (17-19).

A key assumption of this model is that the correlation between observations from the same cluster is time-shift invariant: the correlation between any pair of observations in the same cluster remains constant no matter how far apart in time those two observations are $corr(Y_{itk}, Y_{itl}) = corr(Y_{itk}, Y_{isl}) = \frac{\tau^2}{\tau^2 + \sigma_e^2}$; an assumption which may not be plausible (12, 13).

2.2 Non-uniform correlation structure

Here we extend the model in Equation (1) to allow a more general correlation structure between individuals within the same cluster. We assume the model

$$Y_{itk} = \mu + \beta_t + X_{it}\theta + CP_{it} + e_{itk}, \quad \boldsymbol{CP_i} \sim N_T(0, \sigma_{CP}^2 \boldsymbol{R}), \qquad e_{itk} \sim N(0, \sigma_e^2), \quad (2)$$

where terms in common with Equation (1) have the same interpretation as in that model, CP_{it} is the cluster-period random effect for cluster i in period t with $\boldsymbol{CP_i} = (CP_{i1}, \dots, CP_{iT})'$; we assume that the CP_{it} have common variance σ_{CP}^2 and $\sigma_{CP}^2 \boldsymbol{R}$ is the covariance matrix of

$\boldsymbol{CP_i}$. We assume that subjects within a cluster observed in the same period have an exchangeable covariance structure $cov(Y_{itk}, Y_{itl}) = var(CP_{it}) = \sigma_{CP}^2$, and that the covariance between any pair of subjects from the same cluster but different periods is given by $cov(Y_{itk}, Y_{isl}) = cov(CP_{it}, CP_{is})$. We parameterise $\boldsymbol{R}$ using $R_{ts} = r_{ts}$, where $\sigma_{CP}^2 r_{ts} = cov(CP_{it}, CP_{is})$. The within-period intra-cluster correlation is $corr(Y_{itk}, Y_{itl}) = \frac{\sigma_{CP}^2}{\sigma_{CP}^2 + \sigma_e^2} \coloneqq \rho_0$, and the between-period intra-cluster correlation structure is given by $corr(Y_{itk}, Y_{isl}) = \frac{\sigma_{CP}^2}{\sigma_{CP}^2 + \sigma_e^2} r_{ts} = \rho_0 r_{ts}$ for $t \neq s$.

A useful and quite general between-period intra-cluster correlation structure is the Toeplitz structure, $r_{ts} = r_{|t-s|}$, with the restriction being that the choice of r_{ts} must ensure a positive definite $\boldsymbol{R}$. An exponential decay structure (analogous to an autoregressive order 1 correlation matrix) is returned if $r_{ts} = r^{|t-s|}$. A more general version of the autoregressive structure is obtained taking

$$\boldsymbol{R} = \boldsymbol{R}(r_0, r) = \begin{pmatrix} 1 & r_0 r & r_0 r^2 & \dots & r_0 r^{T-1} \\ r_0 r & 1 & r_0 r & \dots & r_0 r^{T-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_0 r^{T-1} & r_0 r^{T-2} & r_0 r^{T-3} & \dots & 1 \end{pmatrix}$$

The Hussey and Hughes model is returned when $\boldsymbol{R} = \boldsymbol{R}(1,1)$. A model with constant between-period intra-cluster correlations, $corr(Y_{itk}, Y_{itl}) \neq corr(Y_{itk}, Y_{isl})$ but $corr(Y_{itk}, Y_{isl}) = \frac{\sigma_{CP}^2}{\sigma_{CP}^2 + \sigma_e^2} r_0$, $t \neq s$, r_0 a constant, analogous to that of Hooper et al (13) (and that of Girling and Hemming in (4)), is returned if $\boldsymbol{R} = \boldsymbol{R}(r_0, 1)$: we refer to this model as the Hooper/Girling model. This is equivalent to imposing an exchangeable correlation structure on $\{CP_{it}\}$. We note that the model we refer to as the Hooper/Girling model has the same within-cluster correlation structure as the model in Section 3.2 of (20), and is a special case of the model presented in (21). We will consider the autoregressive correlation matrix with $\boldsymbol{R} = \boldsymbol{R}(1, r)$, the Hussey and Hughes, and the Hooper/Girling models in our comparative studies in Sections 3 and 4.

2.3 Variance of the treatment effect estimator

Let $\bar{Y}_{it} = \sum_{k=1}^{m} \frac{Y_{itk}}{m}$ denote the mean outcome in cluster i in period t. If $\overline{\boldsymbol{Y}}_{\boldsymbol{i}} = (\bar{Y}_{i1}, \ldots, \bar{Y}_{iT})'$ is the vector of the period means for cluster i, then the variance-covariance matrix of $\overline{\boldsymbol{Y}}_{\boldsymbol{i}}$ in Equation (2) can be written as

$$Cov(\overline{\boldsymbol{Y}}_{\boldsymbol{i}}) = \boldsymbol{V} = \left(\sigma_{CP}^2 + \frac{\sigma_e^2}{m}\right)(\omega \boldsymbol{R} + (1-\omega)\boldsymbol{I}_T)$$

where $\omega = \sigma_{CP}^2/(\sigma_{CP}^2 + \frac{\sigma_e^2}{m})$ and $\boldsymbol{I}_T$ is the $T \times T$ identity matrix. If $\boldsymbol{X}_{\boldsymbol{i}} = (X_{i1}, \ldots, X_{iT})'$ is the column vector of length T of treatment indicators for cluster i, then we show in Appendix A that

$$var(\hat{\theta}) = \left(\sum_{i=1}^{K} \boldsymbol{X}_{\boldsymbol{i}}' \boldsymbol{V}^{-1} \boldsymbol{X}_{\boldsymbol{i}} - \frac{1}{K} \left(\sum_{i=1}^{K} X_{i1}, \ldots, \sum_{i=1}^{K} X_{iT} \right) \boldsymbol{V}^{-1} \begin{pmatrix} \sum_{i=1}^{K} X_{i1} \\ \vdots \\ \sum_{i=1}^{K} X_{iT} \end{pmatrix} \right)^{-1},$$

where the vector $(\sum_{i=1}^{K} X_{i1}, \ldots, \sum_{i=1}^{K} X_{iT})$ is the number of treated clusters in periods 1 to T. Here we assume that $Cov(\overline{\boldsymbol{Y}}_{\boldsymbol{i}})$ is constant across clusters. In models where the variance matrices differ across clusters this expression will not hold: for example, this would be the case when clusters are of different sizes.

When decay is exponential, with $\boldsymbol{R} = \boldsymbol{R}(1, r), r \neq 1$, or the more complex $\boldsymbol{R}(r_0, r), r \neq 1$, simple closed form expressions for the variance of the treatment effect are difficult to obtain, requiring the inversion of tridiagonal matrices with non-standard forms, hence in our comparative studies we numerically invert the variance matrix V using the standard matrix inversion command "solve" in R. For the Hussey and Hughes ($\boldsymbol{R} = \boldsymbol{R}(1, 1)$) and Hooper/Girling models ($\boldsymbol{R} = \boldsymbol{R}(r_0, 1)$), closed-form expressions for $var(\hat{\theta})$ are available, e.g. (6).

3. Consequences of a more general within-cluster covariance matrix structure

3.1 A proposed intensive care trial

As part of the planning for a potential four period CRXO trial of the effect of overnight placement of earplugs in intensive care patients on hospital length of stay, data from four

six-month periods over 2012 and 2013 from 33 ICUs contributing to the Australian and New Zealand Intensive Care Society's Adult Patient Database (15) was used to investigate the empirical correlation structure over these 4 periods. Ethical review was not required. There were on average 700 patients per ICU per 6 month period. We fit linear mixed models to the logarithm of length of stay with exponential decay and Toeplitz correlation structures for the cluster-period random effects, using the `hpmixed` and `mixed` procedures in SAS, respectively, which allow for the inclusion of correlation structures in the `random` statement. We found that an exponential decay correlation structure provided a reasonable approximation to the data, with $\sigma_{CP}^2 = 0.039$, $\sigma_e^2 = 1.09$ (implying that $\rho_0 = 0.035$), and $\hat{r} = 0.95$, implying a 5% decay per period in the correlation between observations from the same intensive care unit.

3.2 Design of the comparative study

We first consider the impact of assuming an exponential decay within-cluster correlation structure on the variance, power, and design effects of various types of standard multi-period CRTs, where the design effect is the ratio of the variance of the treatment effect for the CRT relative to the variance for an individually-randomised trial of the same size. To provide an example of specific designs, we compare the choices of the four-period trial designs shown in Figure 1 and the eight-period designs in Appendix Figure 1, calculating the variance of the estimator of the treatment effect and associated quantities by fixing the total variance at unity (i.e. $\sigma_{CP}^2 + \sigma_e^2 = 1$ so that the within-period intra-cluster correlation $\rho_0 = \sigma_{CP}^2$), and assuming the within-cluster covariance matrix $R = \boldsymbol{R}(1, r)$: i.e. $\boldsymbol{R}$ has (t, s) element $r_{ts} = r^{|t-s|}$. Since the total variance has been fixed at 1, results are obtained by varying the WPICC $\sigma_{CP}^2 = \rho_0$, and the parameter associated with the BPICC, r. For easier interpretation, we define decay as $d = 1 - r$: $d = 0$ thus implies no decay in correlation over time, while $d = 1$ implies a total decay in correlation over time with independence between observations from the same cluster in different periods. Larger values of the decay, *d*, are not realistic and hence our attention focusses on values less than 0.5.

In presenting the results of this comparative study we focus on the scenario corresponding to the proposed intensive care trial, assuming 500 patients per ICU (allowing for a reduction from 700 per period in the considered dataset due to loss to follow up, patient refusal, etc.)

for each of four (or eight) periods with $\rho_0 = 0.035$. We also present results for designs with 50 subjects per cluster per period.

The four-period stepped wedge design consists of 3 randomised sequences, while the four-period parallel, parallel with baseline and CRXO designs consist of 4 sequences each. Similarly, the eight-period stepped wedge design consists of 7 sequences, while the eight-period parallel, parallel with baseline and CRXO designs consist of 8 sequences each. While designs in practice may assign more than one cluster to each treatment sequence, in our comparative study we assume that there is one cluster per sequence: thus, to allow comparability of the variance of the treatment effect across these designs with differing numbers of clusters (3 for the four-period SW versus 4 for all other four-period designs), we scale the variance associated with the SW design by $\frac{K}{K+1} = \frac{\#\text{clusters}}{\#\text{clusters}+1}$ to ensure comparability.

We calculate each of the following: the variance of the treatment effect; the design effects; and the power of each of the designs to detect specific user-specified effect sizes. Scaling the variance ensures that these subsequent quantities are comparable between the designs with differing numbers of clusters. Interested readers can interactively explore the impact of alternative design configurations through our R Shiny web app. Although we do not include the results in the main paper, the R Shiny app also allows each of these quantities to be calculated for the Hooper/Girling model with $\boldsymbol{R}(r_0 = 1 - \alpha, 1)$ for some user-specified value of $\alpha \in [0,1]$, analogous to the decay parameter d associated with the exponential decay model.

3.3 Results of the comparative study and consequences for the proposed ICU trial

Figure 2 presents the variance for four-period designs (top row) and for eight-period designs (bottom row), for design parameters that align with those of the ICU trial ($\rho_0 = 0.035$ and 50 subjects in each cluster period (left column) and 500 subjects in each cluster-period (right column), for values of the decay up to 0.5). When the decay = 0 (returning the standard Hussey and Hughes model) the BPICC is maximised (and equal to the WPICC), and the amount of information available from within-cluster comparisons is maximised. In this case, it is known that the CRXO design is optimal, e.g. (4), which is reflected in Figure 2. As the decay term increases, the BPICC reduces, and Figure 2 indicates that the variance associated

with the CRXO, parallel with baseline and SW designs increase, while the variance associated with the parallel CRT design reduces. As an aside, we note that for small values of the decay, there is often a large reduction in the variance of the treatment effect when a baseline measurement is added to the parallel design.

Comparing the variance plots for four and eight periods indicates that the relationship between the variance and the decay parameter is consistent across the different designs as the number of periods increases. For the parallel design, as the decay increases, and dependence between observations in different periods decreases, the variance also decreases, while an increasing relationship between the decay and the variance is observed for the CRXO design. For the SW and parallel with baseline designs, although the variance initially increases with increasing decay, for each design the variance peaks at a particular value of the decay before decreasing again (for the SW design this value of the decay is outside of the displayed range).

Figure 3 displays the 'design effect' associated with each of the designs, for each of the considered design configurations: the design effect re-scales the variances of Figure 2 to be relative to that of an individually-randomised trial of the same size. Figure 4 displays the power to detect an effect size of 0.2 associated with each design, and this is as expected given the variances in the left column.

For small values of the intra-cluster correlation (ρ_0 less than about 0.01), for particular design configurations (e.g. 50 subjects per cluster-period with 4 periods for particular values of the decay, see Appendix Figure 2) the parallel CRT estimates the treatment effect more precisely than the SW-CRT does: this is known to occur in the exchangeable setting (4).

4. Consequences of specifying a simpler between-period intra-cluster correlation structure

4.1 Motivation

The proposed ICU trial is unusual in that an extensive dataset is available at the planning stages of the trial for the estimation of WPICCs and BPICCs. In situations where data for the estimation of within-cluster correlation structure at the planning stages of a trial is limited, assessing the fit of an exponential decay structure to the data may be impossible and

researchers may prefer to assume a simpler within-cluster correlation structure. In this section we compare the variance of the treatment effect estimator under the exponential decay within-cluster structure with $\boldsymbol{R} = \boldsymbol{R}(1, r)$ (i.e. $corr(Y_{itk}, Y_{isl}) = \rho_0 r^{|t-s|}$) to that obtained assuming the Hussey and Hughes model, with $\boldsymbol{R} = \boldsymbol{R}(1,1)$ (i.e. $corr(Y_{itk}, Y_{itl}) = corr(Y_{itk}, Y_{isl}) = \rho_0$), and to that obtained assuming the Hooper/Girling model with $\boldsymbol{R} = \boldsymbol{R}(1-\alpha, 1)$ (i.e. $corr(Y_{itk}, Y_{itl}) = \rho_0$ and $corr(Y_{itk}, Y_{isl}) = \rho_0(1-\alpha)$, $\forall t \neq s$ for some constant $\alpha \in [0,1]$). Recalling that $d = 1 - r$, the three models are equivalent when $d = \alpha = 0$. We refer to the variance of the treatment effect estimator obtained using the Hussey and Hughes model as $var^{HH}(\hat{\theta})$, that obtained given an exponential decay model with decay d as $var_d^E(\hat{\theta})$, and that obtained given a Hooper/Girling model with parameter α as $var_\alpha^{HG}(\hat{\theta})$.

We follow the design of the comparative study outlined in Section 3.2, calculating $var_d^E(\hat{\theta})$ for a range of decay parameter values d, and in the first instance dividing this quantity by $var^{HH}(\hat{\theta})$. To compare the exponential decay model to the Hooper/Girling model, for each value of $d \in [0, 0.5]$ and each $\alpha \in [0, 0.5]$ we calculate $var_d^E(\hat{\theta})/var_\alpha^{HG}(\hat{\theta})$. We display the results of the exponential and Hooper/Girling comparisons as a contour plot: d on the x-axis, α on the y-axis, with the magnitude of the relative variance displayed using shades of grey.

4.2 Comparison to the Hussey and Hughes model

Figure 5 displays the relative variance of the treatment effect associated with the exponential decay model to that of the Hussey and Hughes model, for each of the four considered design types, for four-period designs (top row); eight-period designs (bottom row); designs with 50 subjects per cluster-period (left column); and designs with 500 subjects per cluster-period (right column). Figure 5 indicates that the variance of the treatment effect will be underestimated by the Hussey and Hughes model when there is an exponential decay structure for the SW and CRXO designs, while the variance is overestimated for the parallel design. For the parallel with baseline design, for most design configurations the Hussey and Hughes model will underestimate the variance, although there do exist design configurations in which the Hussey and Hughes model will

overestimate the variance: this occurs for values of the decay parameter that are larger than usually seen in practice.

Figure 5 indicates that for most design configurations, the impact of assuming the Hussey and Hughes model on the treatment effect variance is exacerbated as the number of periods increases and as the number of subjects per cluster-period increases. An exception occurs with the parallel with baseline design, which for certain values of the decay has a lower relative variance for a larger number of periods and subjects per cluster-period.

The upper right panel of Figure 5 shows that for the proposed intensive care trial, with 5% decay, $\rho_0 = 0.035$ and 500 patients per cluster per period, for a 4-period SW trial, the variance of the exponential decay model is around 1.8 times higher than that of the exchangeable model. For the eight-period SW trial, $var_d^E(\hat{\theta}) \approx 2.8 \times var^{HH}(\hat{\theta})$. Hence, were an exchangeable model assumed in the planning stages of this trial, the sample size required to detect the effect size of interest would be drastically underestimated. Similar results hold for CRXO and parallel-with-baseline designs, while for the parallel design the assumption of a constant correlation results in an increased variance relative to the exponential decay model.

4.3 Comparison to the Hooper/Girling model

Figure 6 displays the results of comparing $var_d^E(\hat{\theta})$, $d \in [0,0.5]$ to $var_\alpha^{HG}(\hat{\theta})$, $\alpha \in [0,0.5]$, for designs with four periods (top row) and eight periods (bottom row). Contours are used to display ranges of values of $var_d^E(\hat{\theta})/var_\alpha^{HG}(\hat{\theta})$: values greater than one indicate that $var_d^E(\hat{\theta})$is greater than $var_\alpha^{HG}(\hat{\theta})$. On each plot $d = \alpha$ is marked with a line.

By examining the edge between the 0.75-1 and 1-2 regions of the contour plots, marked with a bold line on all plots, values of the Hooper-Girling α parameter that result in the same variance as an exponential decay model with a given decay parameter $d = 1 - r$ can be obtained. The top-left plot of Figure 6 indicates that for the SW model on 4 periods, for values of d less than about 0.3, $var_d^E(\hat{\theta}) \approx var_\alpha^{HG}(\hat{\theta})$ for $d = \alpha$. For this scenario and these values of the decay, assumption of either model with $d = \alpha$ will result in similar required sample sizes. However, this is a special case: the bottom left hand plot in Figure 6

indicates that for the Hooper/Girling model to return the same variance as the exponential decay model for a SW design on 8 periods, $\alpha > d$: e.g. when $d = 0.05$, $\alpha \approx 0.088$.

Figure 6 indicates that in general, the relationship between d and α required to return similar values for the treatment effect variances given these two models may be complex: depending on the design and number of periods, simply assuming that $d = \alpha$ could lead to an over- or under-estimation of the treatment effect variance, with a corresponding over- or under-estimation of required sample size. Some general relationships hold: for the CRXO and parallel designs, for $\alpha = d$, $var_d^E(\hat{\theta}) \leq var_\alpha^{HG}(\hat{\theta})$. However, Figure 6 indicates that for the CRXO design, $var_d^E(\hat{\theta}) = var_\alpha^{HG}(\hat{\theta})$ implies that $\alpha \leq d$, while for the parallel design, $var_d^E(\hat{\theta}) = var_\alpha^{HG}(\hat{\theta})$ implies that $\alpha \geq d$.

In Appendix Figure 3 we display extended versions of the plots of Figure 6, with 500 subjects per cluster-period, extending the range of d and α to [0,1]; Appendix Figure 4 displays analogous plots for designs with 50 subjects per cluster-period. Appendix Figure 3 indicates that for the parallel and CRXO designs, for all values of d there exists a value of α such that $var_d^E(\hat{\theta}) = var_\alpha^{HG}(\hat{\theta})$. That this holds for other parameter choices in these designs can be seen by interrogating the R Shiny app. However, for the SW and parallel with baseline designs, there exist values of d such that no value of α exists such that $var_d^E(\hat{\theta}) = var_\alpha^{HG}(\hat{\theta})$. For example, for the SW design on 4 periods with 500 subjects per cluster in each period, for $d \in (0.78, 0.96)$, there does not exist an α such that $var_d^E(\hat{\theta}) = var_\alpha^{HG}(\hat{\theta})$. For the SW design on 8 periods with 500 subjects per cluster per period, there exists no corresponding α for $d \in (0.43, 0.89)$. For the stepped wedge and parallel with baseline designs, there is not a 1-1 correspondence between the Hooper/Girling model and the exponential decay model.

Figure 6 indicates that there can be large differences between $var_d^E(\hat{\theta})$ and $var_\alpha^{HG}(\hat{\theta})$: should the within-cluster correlation structure be misspecified, it is possible to grossly over- or under-estimate the variance of the treatment effect, although Appendix Figure 4 indicates that the over-estimation will not be as extreme for a smaller number of subjects per cluster-period. Hence, at the design stage of any trial, researchers should investigate the sensitivity of the required sample size and/or power of their study to misspecification of the within-cluster correlation structure.

5. Discussion

Although the usual assumption in planning multiple-period CRTs is that the correlation between observations does not depend on the time between these observations, i.e. the assumption of the correlation structure of Hussey and Hughes (10), this assumption may be invalid. In many situations it would be more reasonable to assume that the correlation between observations on subjects from the same cluster will decrease the further apart in time these observations are obtained. This is well appreciated for individual subject longitudinal data, e.g. (22). We have shown that for designs with a treatment switch, even a decay in between-period intra-cluster correlation as small as 5% per period may appreciably inflate the required sample size relative to that required under an assumption of equality of within-period and between-period intra-cluster correlations, and discussed these results in the context of a proposed cluster randomised trial in the intensive care setting. For this proposed SW trial, the impact of a decay in the between-period intra-cluster correlation of 5% per period results in a sample size almost double that required when no decay (i.e. the Hussey and Hughes model) is assumed. The implication is that cluster randomised trials that incorporate treatment switching but do not account for this decay, and instead assume that an equal amount of information is available from each within-cluster between-period comparison, may be dangerously underpowered to detect stated effect sizes, as has been observed in (13).

For parallel cluster multi-period trials, as the decay increases (i.e. as between-period intra-cluster correlation decreases), precision increases: thus, parallel designs that assume no decay in the between-period intracluster correlation may be overpowered to detect the stated effect size and thus may suffer from a lack of efficiency. For the SW and CRXO designs, precision is greatest when there is no decay in the inter-period correlations ($r = 1$ or the decay = 0). The greater the correlation between observations from the same cluster at different time points, the greater the gains of the within-cluster comparisons that the SW and CRXO designs entail. This has been described in the context of cluster-randomised cross over designs in (3). As the decay increases, the correlation between observations from the same cluster but in different periods decreases. For designs such as the SW and CRXO, which capitalise on comparisons within clusters, this leads to an increase in the variance of the treatment effect, while for the parallel design, which does not incorporate any within-

cluster comparisons, independence between observations within the same cluster in different periods (which occurs when the decay = 1) is optimal.

To date, sample size calculations for stepped wedge designs taking into account that within and between correlations differ have assumed that between-period intra-cluster correlations are constant, as in the Hooper/Girling model we considered in Section 4, described in (13), for example. In some situations such an assumption will likely be more appropriate than the assumption of an exponential decay structure the cluster has completely re-organised itself in the interval between consecutive cross-sections. As an example, suppose clusters are schools in which the same year group is assessed over a number of years. This year-group is a completely different group of children each year, so the correlation between pairs of observations from different years may remain constant rather than decaying as the number of years between groups increases. However, we encourage researchers to assess the sensitivity of their power and sample size to any assumption regarding the within-cluster correlation structure. Although it may be convenient for investigators to assume the Hooper/Girling model at the trial design stage, perhaps adopting a conservative value for the between-period intra-cluster correlations, we have demonstrated that, although there exist choices of constant between-period intra-cluster correlations (the Hooper/Girling α parameter) corresponding to many values of exponential decay parameters, such choices are not always available for the stepped wedge and parallel with baseline designs. Moreover, approximating an exponential-decay correlation structure with a Hooper/Girling model with $d = \alpha$ (i.e. assuming that the lag 1 decay persists for all larger lags) may over- or under-estimate the variance of the treatment effect, possibly dramatically so. For SW designs, whether the variance is over- or under-estimated by a given constant between-period intra-cluster correlation approximation depends on the number of periods, the number of subjects per cluster-period, and the within-period intra cluster correlation, ρ_0.

In order to calculate the required sample size for a given trial assuming an exponential-decay between-period intra-cluster correlation structure, an estimate of the decay parameter is required: such an estimate may be difficult to obtain. More generally, the estimate of the parameter ρ_0 (the within-period intracluster correlation) may be based on data from multiple periods, and thus may itself incorporate a decay in between-period intra-

cluster correlation over time: precisely what the impact of such model misspecification is on the estimated decay parameters and on ρ_0 remains unclear. We expect that conducting a sample size calculation on the basis of a value of ρ_0 estimated from a misspecified Hussey and Hughes model, which estimates ρ_0 over all time periods combined rather than within a single period, then using a Hussey and Hughes model to calculate sample size, will lead to an under-estimation of the true variance of the treatment effect. However, if values of the BPICC and WPICC are obtained from a missppecified Hooper/Girling model, it seems that the estimate of the WPICC will incorporate some additional decay over time, and the impact of assuming a Hooper/Girling model on estimated sample size may be lessened. Further research is required to quantify the impact of model misspecification at the trial design stage. As has been pointed out in (13) for the case of a constant between-period intra-cluster correlation, authors of reports of multi-period CRTs should be encouraged to report estimates of within-period and between-period intra-cluster correlations to help guide the choice of such parameters. Although there have been many studies that present within-period intra-cluster correlations for each period of a study separately, for example (23,24), we would recommend that authors also report between-period intra-cluster correlations, as recommended in (25), for example.

In order for a power calculation to be relevant, it is necessary that the analysis strategy applied at the completion of any trial match that used in the sample size calculation, at least approximately. To that end, the Hussey and Hughes model and the Hooper/Girling model have an advantage over the exponential decay model. Many statisticians will be familiar with techniques for fitting these models, however, the exponential decay model we have considered here is not as widely used. For the proposed ICU trial, we applied the `hpmixed` procedure in SAS, incorporating the exponential decay structure on the cluster-level random effects by including the following statement: `random period / sub = cluster type = AR(1)`. Readers may be interested to note that we applied the `hpmixed` procedure due to large cluster sizes requiring extensive computational time with the `mixed` procedure. Parameters for datasets with smaller cluster sizes may be estimated using the conventional `mixed` procedure.

In this paper we have discussed exponential decay structures in the context of repeated cross-sectional multiple-period CRTs: that is, we have supposed that subjects are each

included in a single cluster-period. Multiple measurements for each subject, as occur in so-called closed or open cohort designs, (5), could be accommodated through the inclusion of a subject-level random effect, as in (13). The model proposed in (13) implies that the correlation between two measurements on the same individual does not depend on the time between the measurements: a similar extension to the one we have here proposed for the cluster level random effects would be possible at the subject level. However, there are likely to be difficulties in the estimation of the parameters of such a model with exponential decay at both cluster and individual levels. Further extensions, such as allowing variances to increase over time, for correlations to change within a period, or for correlations to depend on some function of time, are possible, but given the potential difficulties in specifying these at the design stage, we have not explored these here.

In this paper we have focussed on models for normally distributed outcomes: although we expect the conclusions drawn regarding the implications of an exponential decay intra-cluster correlation structure on power to be similar for binary outcomes, particularly in the case of large clusters, further work is required. Both subject-specific and population-averaged models could be investigated: for population-averaged models this could involve extending the working correlation matrix structures for the pre- post-test CRTs examined in (26) to the multiple-period scenario.

The models we have assumed here include a categorical term for period, with a constant effect of period across clusters, that the effect of periods is constant across clusters, and a constant effect of treatment across clusters and periods. Other authors have considered more complex models e.g. (20, 27, 28). For studies conducted over a large number of periods, it may be advantageous to model time using continuous terms; however, specifying the correct functional form for time will likely be difficult at the design stage of a study.

In conclusion, in this paper we have illustrated settings where exponentially decaying correlation can have a substantial impact on the variance of treatment effect estimators associated with multiple period cluster randomised trials, and hence on sample size and power calculations. This is particularly important for stepped wedge trials which incorrectly assume constant between-period intra-cluster correlations. We have described, and have made available in an RShiny app (https://monash-biostat.shinyapps.io/NonUniformCorrelation/), a facility for carrying out comparative

studies of the nature described in this paper: for their particular multiple-period cluster randomised trial, users can investigate the impact of an exponential decay structure on the required sample size and power of their study, either by setting the values of design parameters or up-loading their own design matrix.

References

1. Hayes RJ, Moulton LH. Cluster randomised trials: Chapman & Hall/CRC; 2009.
2. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012;345e5661.
3. Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McKenzie JE. Appropriate statistical methods were infrequently used in cluster-randomized crossover trials. J Clin Epidemiol 2016;74:40-50.
4. Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. Stat Med 2016;35:2149-2166.
5. 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.
6. Hooper R, Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. BMJ. 2015;350:h2925.
7. Hemming K, Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. J Clin Epidemiol. 2016;69:137-146.
8. Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. Stat Med. 2008;27:5578-5585.
9. Forbes AB, Akram M, Pilcher D, Cooper J, Bellomo R. Cluster randomised crossover trials with binary data and unbalanced cluster sizes: Application to studies of near-universal interventions in intensive care. Clin Trials. 2015;12:34-44.
10. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 2007;28:182-191.
11. Martin J, Taljaard M, Girling A, Hemming K. Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. BMJ Open. 2016;6:e010166.
12. Taljaard M, Teerenstra S, Ivers NM, Fergusson DA. Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. Clin Trials. 2016.
13. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Stat Med. 2016:35;4718-4728.
14. Martin J, Girling A, Nirantharakumar K, Ryan R, Marshall T, Hemming K. Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care. Trials. 2016;17:402.
15. Stow PJ, Hart GK, Higlett T, George C, Herkes R, McWilliam D, et al. Development and implementation of a high-quality clinical database: the Australian and New Zealand Intensive Care Society Adult Patient Database. Journal of Critical Care. 2006;21:133-141.
16. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. Shiny: web application framework for R. https://CRAN.R-project.org/package=shiny. 2017.

17. Baio G, Copas A, Ambler G, Hargreaves J, Beard E, Omar RZ. Sample size calculation for a stepped wedge trial. Trials. 2015;16:1-15.
18. 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;66:752-8.
19. Barker D, McElduff P, D'Este C, Campbell MJ. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. BMC Med Res Methodol. 2016;16:69.
20. Murray DM, Hannan PJ, Wolfinger RD, Baker WL, Dwyer JH. Analysis of data from group-randomized trials with repeat observations on the same groups. Stat Med. 1998:17;1581-1600.
21. Feldman HA, McKinlay SM. Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. Stat Med. 1994:13;61-78.
22. Diggle PJ, Heagerty PJ, Liang K-Y, Zeger SL. Analysis of longitudinal data. 2nd ed. Oxford, UK: Oxford University Press; 2002.
23. Murray DM, Short B. Intraclass correlation among measures related to alcohol use by school aged adolescents: estimates, correlates and applications in intervention studies. J Drug Education. 1996;26:207–230.
24. Siddiqui O, Hedeker D, Flay BR, Hu FB. Intraclass correlation estimates in a school-based prevention study. Am J Epidemiol. 1996;144:425– 433.
25. Martin J, Girling A, Nirantharakumar K, Ryan R, Marshall T, Hemming K. Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care. Trials. 2016;17:402.
26. JS Preisser, ML Young, DJ Zaccaro, M Wolfson. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. Statistics in Medicine. 2003;22:1235-1254.
27. Hemming K, Taljaard M, Forbes A. Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. Trials. 2017;18:101.
28. Hughes JP, Granston TS, Heagerty PJ. Current issues in the design and analysis of stepped wedge trials. Contemporary Clinical Trials. 2015;45;55-60.

Figure legends

Figure 1. Design matrices corresponding to the 4-period designs considered in the comparative study.

Figure 2: Variance of the treatment effect estimator for multiple-period cluster designs for 50 subjects per cluster-period (left column), and 500 subjects per cluster-period (right column). Variances are shown for the four-period stepped wedge, parallel cluster, parallel cluster with baseline, and cluster cross over designs given in Figure 1 (top row), and for eight-period designs in Appendix Figure 1 (bottom row), for varying values of the exponential decay correlation structure decay parameter. For all designs, the within-period intra-cluster correlation $\rho_0 = 0.035$. For the parallel, parallel with baseline and cross-over designs, results are scaled to allow comparison with the stepped wedge results. Note that the vertical scales for the 4-period and 8-period designs differ.

Figure 3. Design effects for the treatment effect estimator for multiple-period cluster designs for 50 subjects per cluster-period (left column), and 500 subjects per cluster-period (right column). Design effects are shown for the four-period stepped wedge, parallel cluster, parallel cluster with baseline, and cluster cross over designs given in Figure 1 (top row), and for eight-period designs in Appendix Figure 1 (bottom row), for varying values of the exponential decay correlation structure decay parameter. For all designs, the within-period intra-cluster correlation $\rho_0 = 0.035$. For the parallel, parallel with baseline and cross-over designs, results are scaled to allow comparison with the stepped wedge results. Note that the vertical scales for all plots differ.

Figure 4. Power of the multiple-period cluster designs for 50 subjects per cluster-period (left column), and 500 subjects per cluster-period (right column) to detect an effect size of 0.2. The thin line at the top of each plot denotes power = 1. Power is shown for the four-period stepped wedge, parallel cluster, parallel cluster with baseline, and cluster cross over designs given in Figure 1 (top row), and for eight-period designs in Appendix Figure 1 (bottom row), for varying values of the exponential decay correlation structure decay parameter. For all designs, the within-period intra-cluster correlation $\rho_0 = 0.035$. For the parallel, parallel with baseline and cross-over designs, results are scaled to allow comparison with the stepped wedge results. Note that the vertical scales for all plots differ.

Figure 5. Variance of the treatment effect estimator obtained given an exponential decay model (with varying values of decay) relative to the variance of the treatment obtained for the Hussey and Hughes model. The y-axis of each plot is on the log (base 10) scale. The top row gives results for the four-period designs given in Figure 1; the bottom row for eight-period designs in Appendix Figure 1. The left column displays results for 50 subjects per cluster per period, and the right column for 500 subjects per cluster per period. For all designs, the within-period intra-cluster correlation $\rho_0 = 0.035$. For the parallel, parallel with baseline and cross-over designs, results are scaled to allow comparison with the stepped wedge results.

Figure 6. Contour plots of the variance of the treatment effect estimator obtained given an exponential decay model relative to that given the Hooper/Girling model, for varying values of the exponential decay model decay parameter and the Hooper/Girling α parameter. The top row displays results for the four-period designs in Figure 1, and the bottom row displays results for the eight-period designs in Appendix Figure 1. For all designs, the within-period intra-cluster correlation $\rho_0 = 0.035$, with 500 subjects per cluster per period. The line of equality between the two parameters is marked with a thin line, and equality of the variances is marked with thick lines on each plot.

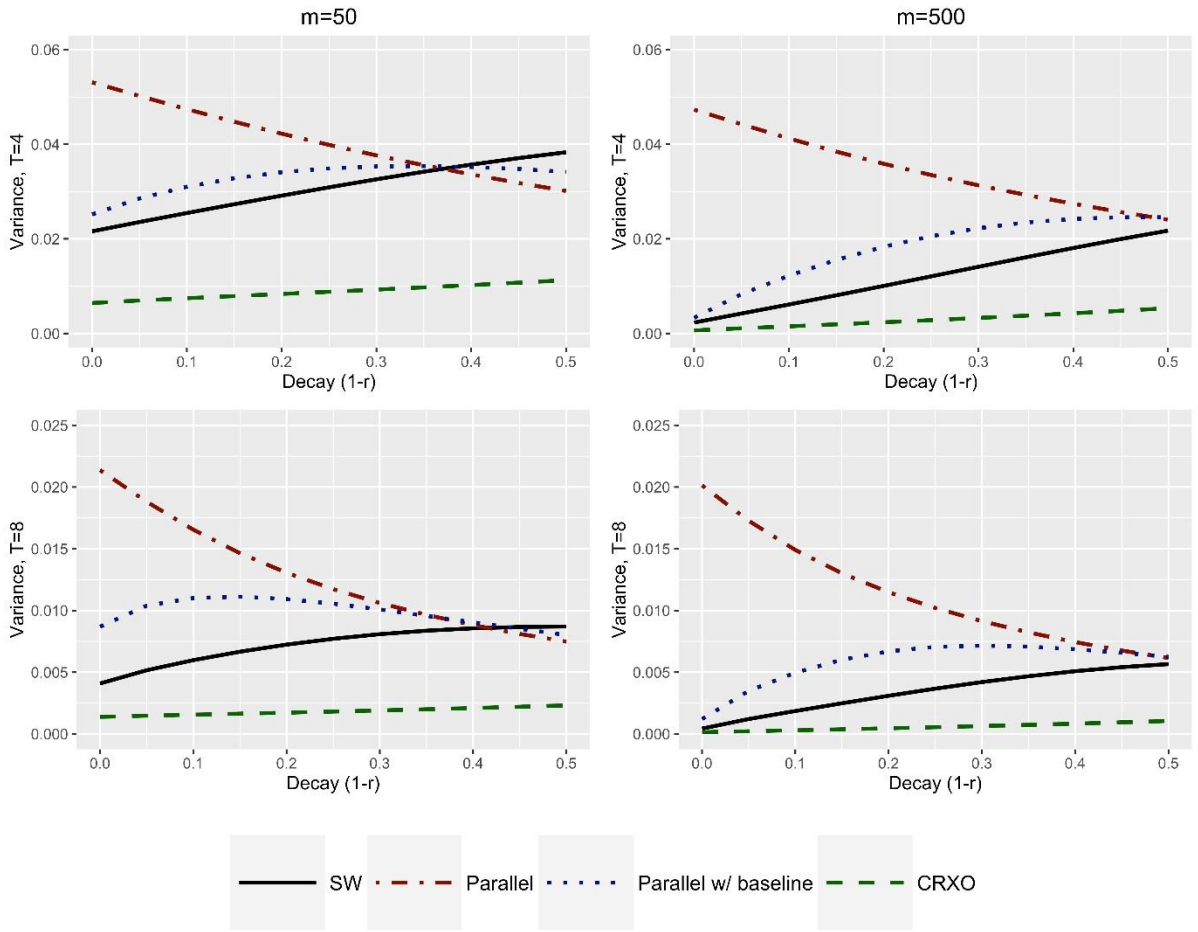
m=50
m=500
Variance, T=4
Variance, T=8
Decay (1-r)
SW
Parallel
Parallel w/ baseline
CRXO

Figure 2

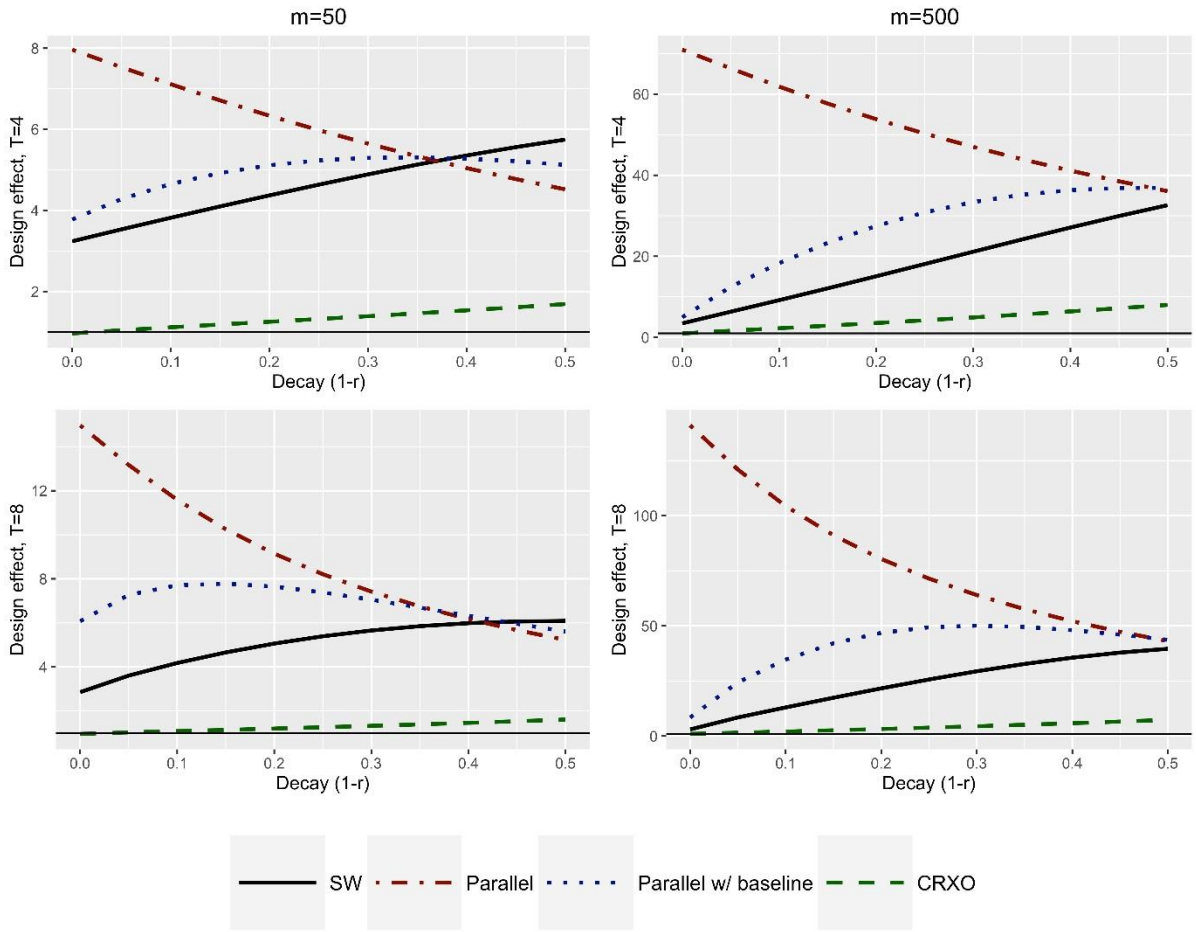
m=50
m=500
Design effect, T=4
Design effect, T=8
Decay (1-r)
SW
Parallel
Parallel w/ baseline
CRXO

Figure 3

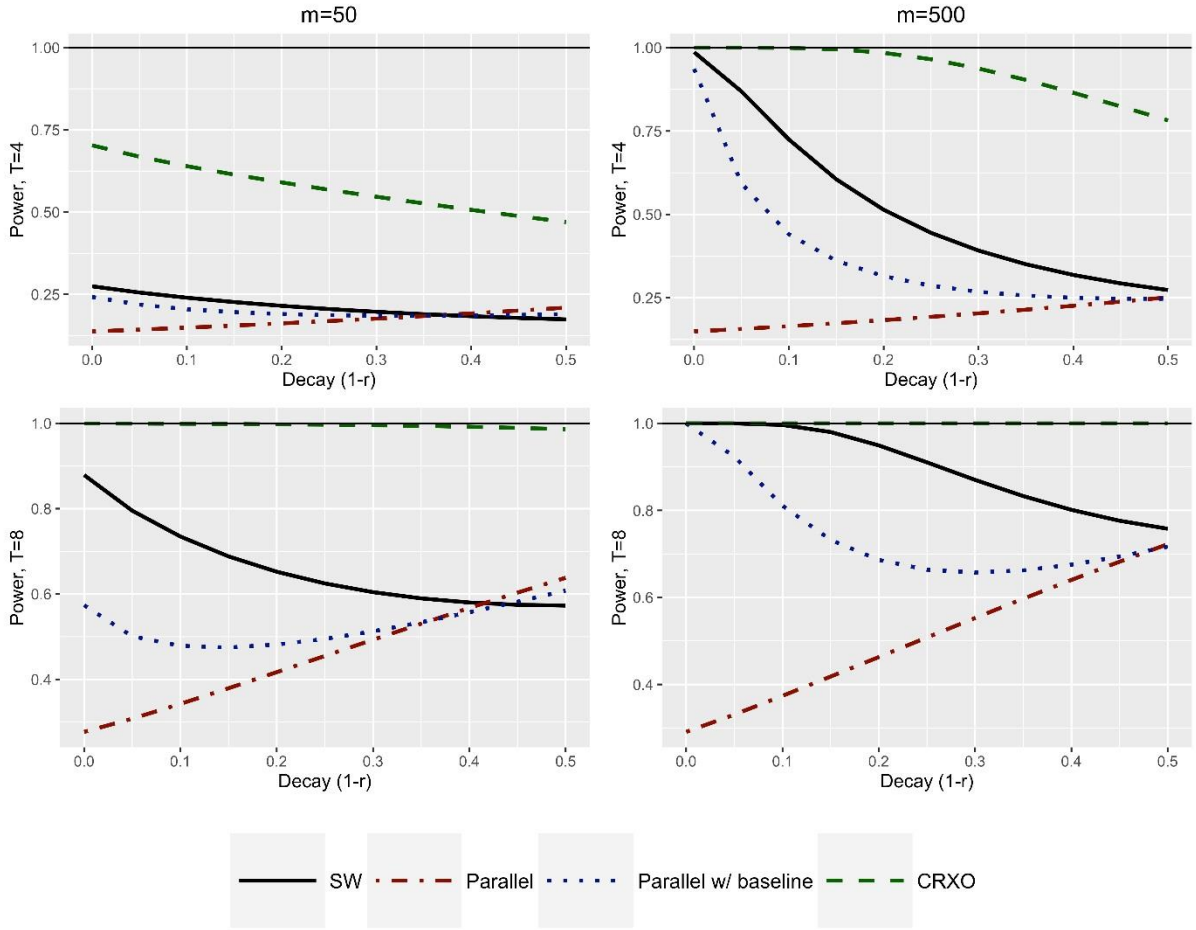
m=50
m=500
Power, T=4
Power, T=8
Decay (1-r)
SW
Parallel
Parallel w/ baseline
CRXO

Figure 4

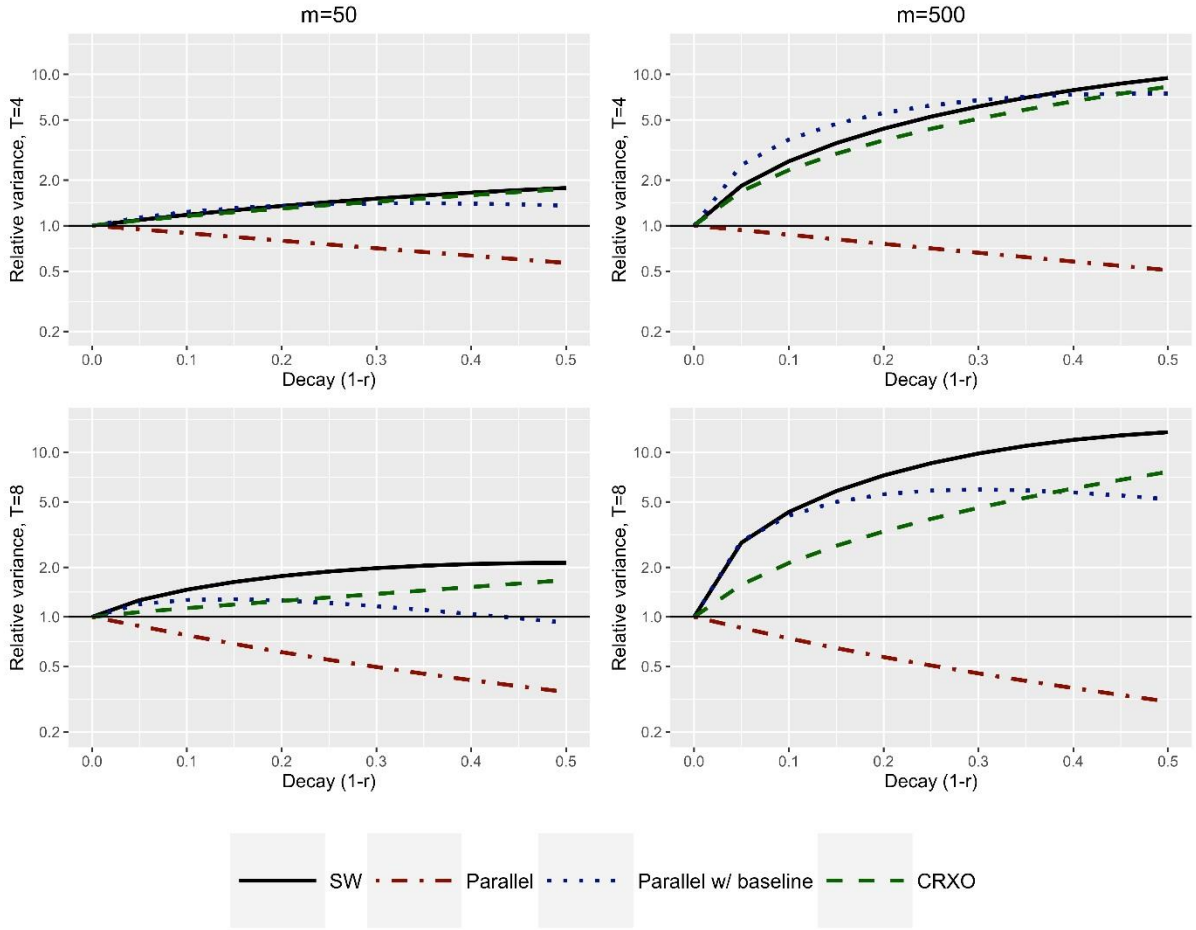

Figure 5

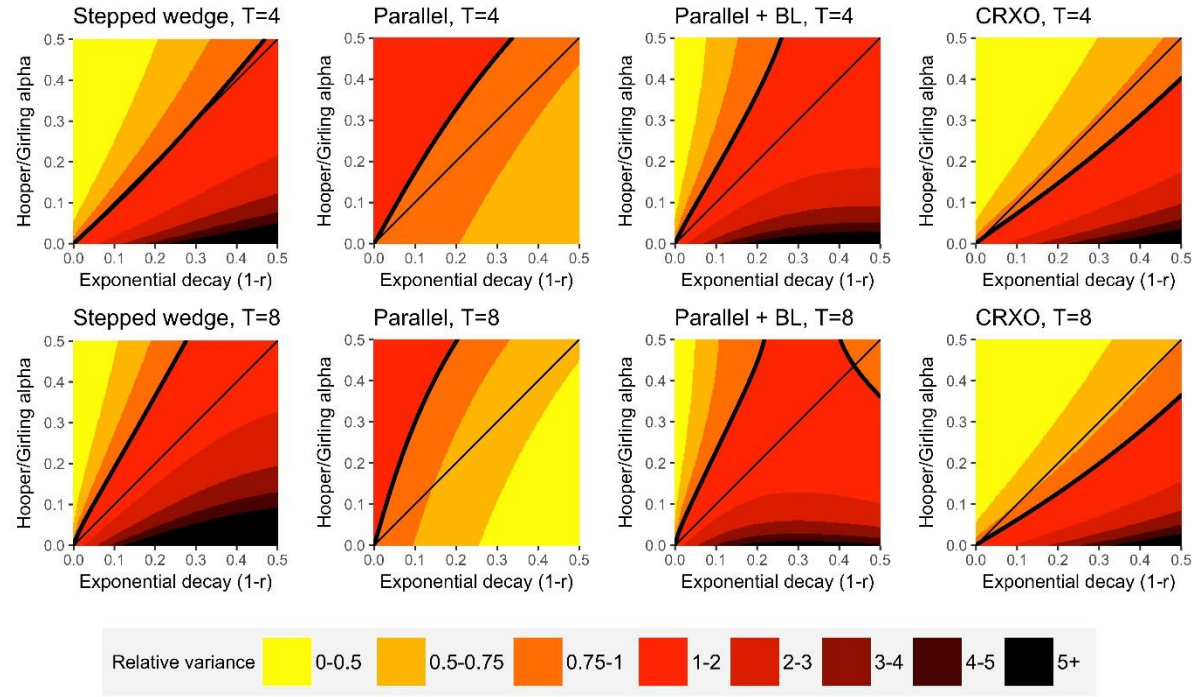

Figure 6