Stepped Wedge Cluster Randomized Trials

STAT 548 Qualifying Paper

Kenny Chiu

November 29, 2021

Abstract. The work of Hussey and Hughes (2007) can be viewed as an entry-point to the study of stepped wedge cluster randomized trials. We summarize and comment on their work and address the shortcomings of their paper, which include missing technical details, a simulation study with a vague procedure, and a limited discussion of how their approach may be extended to more complicated trial contexts. We also highlight how their work has guided modern developments in the related literature.

1 Introduction

The work of Hussey and Hughes (2007) can be viewed as an entry-point to the study of stepped wedge cluster randomized trials (SW-CRT), which is a type of experimental design that is particularly pragmatic compared to alternative designs that may also be used in similar contexts. In this report, we review the paper by Hussey and Hughes. We summarize the main ideas while filling in missing details, replicate one of the empirical studies with a slight modification, and provide a critique of the paper. We also discuss how the literature on SW-CRTs has developed since the paper was published, and highlight some of the common extensions to the basic model presented in the paper.

This report is organized as follows: Section 2 summarizes the paper and provides additional details where we feel was missed in the original work; Section 3 discusses our perspective and commentary of the paper; Section 4 describes our attempt to replicate the simulation study and our findings; Section 5 presents common modern extensions to the standard SW-CRT model; and Section 6 concludes this report by summarizing our discussion. Appendix A includes our longer technical explanations and derivations deferred from Section 2. Appendix B shows visualizations of example datasets generated through our simulation procedure.

2 Summary and additional details

In this section, we summarize the main ideas of the paper by Hussey and Hughes (2007) and provide additional details where we feel are absent in the original paper. Our longer technical explanations and derivations are included in Appendix A to avoid disrupting the flow of the summary.

2.1 Context and motivation

Cluster randomized trials (CRT) are characterized by the randomization to interventions being done at the group or cluster-level rather than at the individual-level, and it is typically assumed that the individuals within a cluster are correlated. CRTs are considered when it is inconvenient or not appropriate to administer an intervention to single individuals. Hussey and Hughes (2007) comment that the majority of CRT designs studied and employed (at the time of the paper) featured parallel designs where approximately half of the clusters are simultaneously given one intervention and the other half are simultaneously given another. While these parallel CRTs are convenient analytically, they may present problems in practice if, for example, there are logistical constraints that make delivering the intervention simultaneously across multiple clusters difficult. Other potential issues of parallel designs include ethical concerns where if there is an expectation that a new intervention improves on an existing one, then it may not be possible to withhold the new intervention from certain clusters. Our understanding of the main objective in Hussey and Hughes' work is to promote the stepped wedge CRT design as an alternative that addresses the potential issues of the parallel design, and to provide an overview of how the data collected from such a design are analyzed. In addition, Hussey and Hughes also discuss certain statistical considerations of SW-CRT designs such as power and efficiency of estimators, and how these properties are affected by model assumptions and design parameters.

2.2 SW-CRT design

The SW-CRT design is a type of crossover design. However, unlike in standard crossover CRTs where clusters start with potentially different treatments and switch treatments at a determined time point, SW-CRT are characterized by

1. the crossover being unidirectional where all clusters start with the same treatment (the control or an existing treatment) and end with the same treatment (the intervention), and

2. the staggered times at which each cluster switches to the intervention (with the times being randomized across clusters).

Figure 1 (taken from the original paper) clearly illustrates the differences between the discussed CRT designs.

Parallel		Time
		1
Cluster	1	1
	2	1
	3	0
	4	0

Crossover		Time	
		1	2
Cluster	1	1	0
	2	1	0
	3	0	1
	4	0	1

Stepped wedge		Time				
		1	2	3	4	5
Cluster	1	0	1	1	1	1
	2	0	0	1	1	1
	3	0	0	0	1	1
	4	0	0	0	0	1

Figure 1: Example treatment schedules for parallel, crossover, and stepped wedge CRT designs. The control/existing treatment and the intervention are denoted 0 and 1, respectively. Figure slightly modified from (Hussey & Hughes, 2007).

From Figure 1, it can be seen how the SW-CRT design addresses the practical issues of the parallel design. Rather than simultaneously delivering the intervention to multiple groups, SW-CRTs stagger the delivery to clusters across different times, potentially alleviating logistical concerns. Furthermore, all clusters eventually obtain the intervention, which avoids the problem of withholding the intervention from certain clusters. The SW-CRT design is not without its own complications, however. Staggering the times means that the duration of the study is elongated relative to the parallel and crossover designs. The unidirectional crossover also implies that time may be correlated with the effect of the intervention, which may lead to issues in estimation of the intervention effect when analyzing the data.

Beyond the general characteristics of SW-CRTs described above, other considerations and design parameters (e.g., cluster sizes, number of clusters crossing over at each time point, new individuals in a cluster across time, etc.) will depend on the context of the specific study. Hussey and Hughes examine in detail the model for a SW-CRT under a specific setting. How the model changes to varying trial contexts are only briefly mentioned or, in the case of some variations, not discussed at all. We return to this point in our critique of the paper in Section 3 and again when we discuss model extensions in Section 5.

2.3 Assumed setting and SW-CRT model

The SW-CRT model that Hussey and Hughes (2007) examine in their paper can be considered the "basic" or "standard" model and is based on a particular example trial.

2.3.1 Expedited Partner Treatment trial

The primary SW-CRT setting that Hussey and Hughes work under is based on the context of the *Washington State Community Expedited Partner Treatment (EPT) Trial* that started in 2007 and ended in 2011 (Golden et al., 2015). The hypothesis of interest in this study was whether a public health program that increases the use of EPT decreases the prevalence of chlamydia in young women and the incidence of gonorrhea in WA state. The intervention—promotion of EPT and targeted provision of partner services—was instituted in 23 WA state local health jurisdictions (LHJ) across four waves separated by 6–9 months. Each wave included approximately six LHJs, and the order in which LHJs initiated the intervention were randomly assigned. The measured primary outcomes in the study included the prevalence of chlamydia in women aged 15–25 who tested positive in participating clinics and the incidence of gonorrhea in women as ascertained through public health reporting.

One of the key design aspects to consider for modeling is how the clusters and the individuals are defined. For example, in the chlamydia study of the EPT trial, each LHJ is treated as a cluster, and the women who were tested in a participating clinic during a particular timeframe are the individuals in the cluster. It is important to note that the women within a single cluster differ at different time points over the trial (i.e., the study features a cross-sectional SW-CRT). While the number of women who were tested varied between LHJs and across time within a single LHJ, there were likely more than enough women tested that only a sample of women for each time point and for each LHJ were sufficient for fitting a reasonable model. In such a scenario, it is analytically more convenient to take samples of equal sizes across LHJs and across time. Hussey and Hughes (2007) comment that for the gonorrhea study, because the incidence rates are much lower, taking samples of equal sizes across LHS and time may not be possible.

Another aspect to consider is the form of the outcome. In the EPT trial, the measured outcomes are prevelance of chlamydia and incidence of gonorrhea, both of which are continuous measures. Motivated by the EPT trial, Hussey and Hughes therefore examine the properties of a basic model for a continuous outcome, cross-sectional SW-CRT. Their discussion of the model is primarily under the assumption of equal cluster sizes, though they also examine the case of unequal sizes in their simulation study.

2.3.2 SW-CRT model

A linear mixed effects model (LMM) can be used to model the SW-CRT design described under the setting in the previous section. Assuming that there are I clusters, T time points, and N individuals in each cluster at each time point, the mean for cluster i at time j can be defined as

$$\mu_{ij} = \mu + \alpha_i + \beta_j + X_{ij}\theta$$

where

- μ is the overall mean across clusters and time,
- $\alpha_i \sim N(0, \tau^2)$ is a random effect for cluster $i \in \{1, \dots, I\}$ that captures the correlation between individuals,
- β_j is a fixed effect for time point $j \in \{1, \dots, T-1\}$ (assuming $\beta_T = 0$ for identifiability),
- X_{ij} is a treatment indicator for cluster i at time j with 1 denoting the intervention, and
- θ is the treatment effect of interest.

Using a slightly different notation from Hussey and Hughes, a model at the individual-level is then given by

$$Y_{ijk} = \mu_{ij} + e_{ijk}$$

where $e_{ijk} \sim N(0, \sigma^2)$ are i.i.d. noise, and this leads to a model at the cluster-level given by

$$\bar{Y}_{ij} = \mu_{ij} + \frac{1}{N} \sum_{k=1}^{N} e_{ijk} \ .$$

As mentioned, this model can be considered the standard model for a SW-CRT design as the assumptions it makes are fairly basic. The model can be extended many ways depending on the setting of the specific study. We revisit this point in Section 5.

Aside from the model itself, two other quantities obtained from the model are also commonly referred to in the CRT literature. The intraclass correlation (ICC) $\rho = \frac{\tau^2}{\tau^2+\sigma^2}$ (and its induced variation inflation factor $1 + (N-1)\rho$) and the coefficient of variation (CV) $\frac{\tau}{\mu}$ characterize the effect of the within-cluster correlation on the cluster mean variance and are often the parameters being adjusted in CRT simulation studies. We provide some intuition for these quantities in Appendix A.1.

2.4 Methods and analysis

Hussey and Hughes (2007) discuss several ideas related to estimation and analysis of the simple model for a SW-CRT. We highlight the key points in this section. We also note that Hussey and Hughes generally skip over explanations and derivations when discussing an estimator or one its statistical properties. We try to address these gaps in this section and in Appendix A.

2.4.1 Estimation of the treatment effect

The general objective when analyzing data from a SW-CRT is to estimate and test the treatment effect θ. When the variance components τ^2 and σ^2 are known, a cluster-level estimation of θ is possible using weighted least squares (WLS). While this approach is useful for conducting a pre-trial power analysis, it is generally the case in practice that τ^2 and σ^2 are unknown. When the variance components are unknown, an individual-level analysis using generalized linear mixed effects models (GLMM) or generalized estimating equations (GEE) will likely be the preferred approach. If there is a possibility of misspecifying the covariance structure, GEE would be preferred over GLMM as its parameter estimates remain consistent as long as the mean is correctly specified (Diggle et al., 2002). Hussey and Hughes caution that the LMM, GLMM, and GEE approaches all rely on asymptotic results, and so an analysis of a SW-CRT involving few clusters or time points may produce misleading findings.

Hussey and Hughes note that when there are no time effects on the outcome (i.e., when $\beta_j = 0$ for all j), estimation of the treatment effect θ can be done using a within-cluster analysis (an analysis based on comparing the control and intervention time periods for each cluster). This case also allows testing of the treatment effect using a paired t-test where the two groups correspond to the control and the intervention time periods. However, if it is incorrectly assumed that the time effects are trivial, the estimator for θ will be biased. We provide additional details about the estimator and the bias in Appendix A.2.

Hussey and Hughes also discuss the relative efficiency of the WLS estimator of the treatment effect compared to the within-cluster estimator. When there are no time effects, the WLS estimator is always more efficient than the within-cluster estimator. When there are time effects, the within-cluster estimator is more efficient (but likely biased). Note that there is an error in the efficiency given by Hussey and Hughes (2007) as pointed out by Liao et al. (2015), but the above statements still hold. Hussey and Hughes (2007) also imply that an exception to the WLS estimator being more efficient in the no time effect-case is when $\tau^2 = 0$, but the efficiency proof by Liao et al. (2015) appears to hold regardless. We provide further details for the efficiency and the proof for the no time effect-case in Appendix A.3.

2.4.2 Power and relevant factors

To obtain an approximate power for the study, Hussey and Hughes (2007) prescribe using a Wald test to test the hypothesis $H_0 : \theta = 0$ against a simple alternative $H_a : \theta = \theta_a$. The power for a two-tailed test of size α given by Hussey and Hughes (2007) is

$$\text{power} \approx \Phi\left(\frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}} - Z_{1-\frac{\alpha}{2}}\right)$$

where Φ is the cumulative distribution function of the standard normal and $Z_{1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)$-th quantile of the standard normal. We note that this approximation implicitly makes the assumption that $\theta_a > 0$ and is not too small. This approach of computing power is applicable to any estimator that is normally distributed or based on large-sample statistics. We provide additional details about the Wald test and the power calculation in Appendix A.4.

Hussey and Hughes also discuss how the power decreases with fewer measured time points and with delays in the treatment effect. Both of these factors are more of a concern when there are constraints on the duration of the trial. The decrease in power can be mitigated by allowing for more time points (i.e., fewer clusters randomized to each time point) and by increasing the length of each time interval to allow for the treatment effect to realize over fewer time points. We provide an explanation for these statements in Appendix A.5.

2.5 Simulation study of analysis methods

Hussey and Hughes (2007) performed a simulation study to compare the power of the test $H_0 : \theta = 0$ versus $H_a : \theta \neq 0$ in a SW-CRT analysis using LMM, GEE, and GLMM. The case of equal cluster sizes and the case of unequal cluster sizes were both investigated. Their results suggest that LMM has greater power than the other two methods when the cluster sizes are equal, and otherwise GEE and GLMM have greater power. They explain that when the cluster sizes are unequal, the clusters need to be weighted in LMM, but the weights will depend on the true (unknown) variance components.

We comment that certain details of the simulation study are ambiguous based on Hussey and Hughes' description. We try to clarify these details through our replication of their study in Section 4.

3 Critical appraisal

We comment on our perspective of the paper by Hussey and Hughes (2007) and critique the paper's strengths, limitations and weaknesses.

From our understanding of the context of the paper, the paper is clearly aimed at addressing a common problem (the use of parallel designs in CRTs even in the presence of logistical or other concerns) and making a case for the SW-CRT design as a solution to the problem. While the design itself was not new, it is said that the design was infrequently employed in CRTs at the time of the paper and that analyses of such designs generally varied from trial to trial (Brown & Lilford, 2006). From this perspective, the main contribution of Hussey and Hughes is the overview of SW-CRTs detailing the motivation, design, analysis, statistical properties, and possible extensions/issues of consideration that is presented in a relatively succinct and accessible format. Looking at the number of citations that the paper has (approximately 970 at the time of this report) and how the literature has developed since then, the paper successfully achieved its purpose of promoting SW-CRT designs and being an entry-point for those unfamiliar with the design.

The main limitation of the paper is the limited breadth and minimal depth it provides on the technical details of the SW-CRT. This limitation was likely the tradeoff on keeping the paper accessible, which would be an important consideration of the authors as those who actually implement and would be interested in such trials are mostly limited to policy makers and researchers who may have limited statistical background. For this reason, the paper focuses mainly on the technical aspects that are of practical relevance such as analysis and power. Even then, the technical discussion is restricted to primarily the basic model of focus. Derivations are skipped over entirely, and the discussion of extended models are generally left to other references.

From our own attempt at replicating their simulation study, we found that certain details of the simulation are unclear from Hussey and Hughes' description. For example, it is mentioned that the powers in the results are calculated using the standard variance. It may be reasonable to assume that the standard variance refers to the standard error returned in the output of the used function. However, this estimate would be only approximate in the case of LMM where binary data are aggregated for cluster-level analysis, and whether this is the case in the simulation is left unaddressed. Other details of the simulation similarly are missing and must be inferred from the context.

In our opinion, the weaknesses of the paper mainly relate to its presentation and its organization. From the writing itself, it is not entirely obvious which aspects of the paper are novel contributions. For example, the LMM is presented in the paper as a common model in CRTs, but the paper is often cited in the literature when the LMM is referenced (e.g., Bowden et al., 2021; Davis-Plourde et al., 2021; Harrison et al., 2020). It can be unclear from the presentation of the paper whether the LMM section is introducing background or a proposed model. Our understanding is that most of the content are synthesized overviews of existing ideas with the exception of the power calculation procedure, the factors affecting power, and the simulation study. Related to the presentation, the organization of the paper can also be a source of confusion as the statistical issues subsections seemingly jump from one idea to another (sometimes under disconnecting assumptions, e.g., when no time effects assumed in one section but not the following). Finally, there are also a number of typos throughout the paper. While most are minor where the intended idea can be inferred from a second glance, one major typo (or error) is the missing factor in the denominator of the relative efficiency of the estimators as discovered by Liao et al. (2015). As the derivation is not included with the paper, this error would be difficult to notice otherwise.

As part of the objectives for this report, we address the main limitation of the paper by developing the discussion on breadth and depth of SW-CRT designs. Our summary of the paper in Section 2 included notes of where we determined technical details could be further developed with our added details given in Appendix A. The following section describes our attempts to replicate the simulation study and what we found as a result. Section 5 addresses the breadth of extensions to the standard SW-CRT model that are now commonly seen in modern SW-CRT literature.

4 Simulation investigation

In this section, we describe our attempt to replicate the simulation by Hussey and Hughes (2007) and to clarify the missing or ambiguous details. The motivation of their simulation study was to compare the power for testing the hypothesis $H_0 : \theta = 0$ between analyses using LMM, GEE and GLMM. The data used in the study was simulated based on the EPT trial under either equal and unequal cluster sizes (which are said to mimic the sampling plans for comparing chlamydial and gonorrheal rates, respectively).

4.1 Simulating the data

We explain how we simulated the data in our attempt to replicate the study. In each of the 1000 simulations,

- the number of clusters is $I = 24$,
- the number of measured time points is $T = 5$ (four randomization times with six clusters crossing over at each),

- the baseline prevalence of disease is $\mu = 0.05$, and
- the between-cluster variance is assumed to be $\tau^2 = 0.000225$.

In the equal size case, the number of individuals in each cluster at each time point is $N = 100$. In the unequal size case, for each simulation, the parameter of a multinomial distribution is sampled from a Dirichlet distribution with parameters $(1, \ldots, 1)$. A multinomial distribution with this parameter is then used to randomly assign $(N-1)I = 2376$ individuals across I clusters. One individual is added to each cluster afterwards to avoid empty clusters, making the total number of individuals across clusters equal to $NI = 2400$. Each cluster retains the same number of individuals across time points.

Once the cluster sizes are determined in a simulation, individual-level data is simulated for each cluster at each time point. Hussey and Hughes state that they simulate individual-level data using the linear model, which we interpret as each individual in cluster i at time j being simulated from a Bernoulli distribution with probability p_{ij} given by

$$p_{ij} = \max(0, \mu + \alpha_i + X_{ij}\theta) \ .$$

The cluster random effects α_i are sampled from a Normal$(0, \tau^2)$ distribution and are resampled in every simulation. The treatment indicator X_{ij} for each time point j is determined after each cluster is randomly assigned a crossover time, which are also reshuffled in every simulation. The treatment effect θ is determined by the risk ratio (RR) chosen for the study where $\theta = \mu(\mathrm{RR} - 1)$. The values of RR examined by Hussey and Hughes include $\{0.5, 0.6, 0.7, 1\}$, with $\theta = 0$ for RR $= 1$ and $\theta = -0.025$ for RR $= 0.5$. The max is taken for p in case $p < 0$ due to the sampled cluster effect. It is assumed that there are no time effects.

The generated dataset in each simulation has 12,000 rows. Examples of generated (aggregated) datasets are visualized in Appendix B. We note that while we tried to use the same 1000 datasets (or the same first 100 when fewer simulations are run) to obtain the results for each RR, we are unable to guarantee this as our code relies on a set seed but is run multiple times with under different parameter configurations that may alter the control flow of the script. Given more time and computational resources, we would first generate the datasets and save them in order to reuse them across configurations.

4.2 Fitting the model and calculating the power

For LMM, the individual-level data is first aggregated into a dataset with 120 rows where each row corresponds to a particular cluster and time point. The LMM for the mean response of a cluster-time that we use is

$$Y_{ij} = \mu + \alpha_i + \beta_j + X_{ij}\theta + \frac{1}{N_i}\sum_{k=1}^{N_i} \epsilon_{ijk}$$

where the parameters are defined as in Section 2.3.2. Notice that we also include time effects β_j in the model as from the analyst's perspective, it can be difficult to fully assume that there are no time effects. For GEE and GLMM, the individual-level dataset is used and we model the response of an individual as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + X_{ij}\theta + \epsilon_{ijk} \ .$$

We note that Hussey and Hughes do not clarify what link function they use in the GEE and GLMM. We would assume that they use the identity link function so that across all three models, the mean response for an individual in the control phase is given by $\mu + \beta_j$ and for an individual in the intervention phase is given by $\mu + \beta_j + \theta$. This allows the treatment effect θ to be interpreted similarly and leads to comparable powers. However, it is also possible that they use the default logit link. We try both link functions in our simulations.

To fit the LMM, GEE and GLMM in `R`, we use the same `R` functions specified by Hussey and Hughes. We use the `lme()` function from the `R` package `nlme`, the `gee()` function from the `gee` package, and the `glmmPQL()`

function from the MASS package. For GEE, the correlation structure is specified to be exchangeable. Parameters aside from the ones related to our models above are otherwise kept at the default setting.

To approximate the power, we compute in each simulation the Wald test statistic given by

$$W = \frac{\hat{\theta}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}}$$

where $\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}$ is taken as the standard error of $\hat{\theta}$ given in the outputs of the functions (which we assume is what Hussey and Hughes refer to as the "standard variance"). For GEE, we use the robust (Huber-White) standard error in the output. We reject the null hypothesis if $|W| > z(0.975)$ where $z(0.975)$ is the 97.5% quantile of the standard normal distribution. The power is then estimated by the proportion of rejections out of the 1000 simulations. In the unequal cluster size case, we found rare instances where the LMM failed to be fit. In this case, the LMM power is estimated by the proportion of rejections out of the number of non-failing simulations.

Hussey and Hughes also calculate the power using a jackknife estimate of the variance. We discuss this component of the simulation in the following section.

4.3 Jackknife estimate for the variance

Due to time constraints on this report, we only consider a small simulation study with jackknife estimates of the variance involving only LMM and GLMM over 100 simulations. As the exact jackknife estimator of the variance used by Hussey and Hughes is unclear, we propose using a delete-a-cluster jackknife estimator for computational reasons. We choose a cluster as the unit of deletion as deleting individuals would be computationally infeasible, and deleting cluster-times may not make sense for a SW-CRT as some clusters only receive the treatment at one time point (and therefore deleting this particular cluster-time would result in the cluster not being able to contribute to the estimation of the treatment effect). In the context of our simulation, we assume that cluster sizes N_i may vary but remains the same over time for each cluster. Therefore, we use a jackknife estimator for the variance that is weighted by the cluster size. Let

- $\mathbf{y}$ denote the data collected from the SW-CRT,
- $\mathbf{y}_{-i}$ denote the data for all clusters but cluster i across all time points,
- $\hat{\theta}(\bullet)$ denote the treatment effect estimator given data $\bullet$, and
- $M = \sum_{i=1}^{I} N_i$ for convenience of notation.

We define cluster pseudo-values where the pseudo-value for cluster i is defined as

$$\hat{\theta}_i = \frac{M\hat{\theta}(\mathbf{y}) - (M - N_i)\hat{\theta}(\mathbf{y}_{-i})}{N_i} \ .$$

The pseudo-value $\hat{\theta}_i$ can be viewed as an averaged estimate of θ contributed by each individual in cluster i. Our jackknife estimator for θ is then the weighted mean of the cluster pseudo-values given by

$$\hat{\theta}_{\mathrm{JK}} = \frac{1}{M}\sum_{i=1}^{I} N_i\hat{\theta}_i = I\hat{\theta}(\mathbf{y}) - \frac{1}{M}\sum_{i=1}^{I}(M - N_i)\hat{\theta}(\mathbf{y}_{-i}) \ .$$

Our estimator for $\mathrm{Var}(\hat{\theta})$ is then the (approximate) variance of the weighted mean given by

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{M^2}\sum_{i=1}^{I} N_i^2(\hat{\theta}_i - \hat{\theta}_{\mathrm{JK}})^2 \ .$$

We use this jackknife estimator for both LMM and GLMM, though note that the data $\mathbf{y}$ and $\mathbf{y}_{-i}$ are at different levels of granularity for the two approaches. Also note that we do not use the jackknife estimate $\hat{\theta}_{\text{JK}}$ for θ when computing the Wald test statistic.

4.4 Results

We reproduce the results from Hussey and Hughes (2007) in Table 1 for ease of comparison.

Risk ratio	Same cluster sizes			Different cluster sizes		
	LMM	GEE	GLMM	LMM	GEE	GLMM
1.0	0.056 (0.057)	0.084 (0.052)	0.076 (0.053)	0.048 (0.038)	0.095 (0.053)	0.069 (0.049)
0.7	0.697 (0.658)	0.719 (0.644)	0.716 (0.580)	0.307 (0.307)	0.703 (0.577)	0.697 (0.559)
0.6	0.907 (0.884)	0.907 (0.866)	0.917 (0.820)	0.487 (0.503)	0.879 (0.807)	0.906 (0.805)
0.5	0.988 (0.984)	0.990 (0.981)	0.992 (0.948)	0.625 (0.653)	0.982 (0.946)	0.986 (0.942)

Table 1: The original results of the simulation study by Hussey and Hughes (2007). The power is computed using both the "standard variance" and a jackknife estimate of the variance (in parentheses).

Our approximate powers calculated using the output standard errors are shown in Table 2. The simulation and power computation procedures are described as in Sections 4.1 and 4.2, respectively. Note that the powers for GEE in the case of differing cluster sizes were approximated over 100 simulations rather than over 1000 due to computational reasons and time constraints.

Risk ratio	Same cluster sizes					Different cluster sizes				
	LMM	GEE		GLMM		LMM	GEE*		GLMM	
		id	logit	id	logit		id	logit	id	logit
1.0	0.050	0.089	0.081	0.066	0.052	0.062	0.11	0.10	0.058	0.053
0.7	0.700	0.736	0.723	0.805	0.711	0.345	0.70	0.68	0.779	0.688
0.6	0.920	0.928	0.920	0.963	0.929	0.536	0.95	0.93	0.951	0.913
0.5	0.981	0.983	0.982	0.994	0.985	0.719**	0.99	0.97	0.997	0.985

Table 2: Proportion of rejections for testing the treatment effect in LMM, GEE, and GLMM across 1000 simulations (*100 simulations for GEE when cluster sizes differ). The Wald test statistic is computed using the standard error provided in the function output (robust SE for GEE). Proportions for GEE and GLMM with identity (id) and logit link functions are reported. **LMM failed to fit the model on two simulated datasets and so this power was approximated over 998 simulations.

The approximate powers for LMM and GLMM calculated using our jackknife estimate of the variance are shown in Table 3. Due to time constraints, all powers in Table 3 are approximated over 100 simulations.

	Same cluster sizes			Different cluster sizes		
Risk ratio	LMM	GLMM		LMM	GLMM	
		id	logit		id	logit
1.0	0.06	0.09	0.07	0.02	0.08	0.07
0.7	0.69	0.70	0.70	0.28*	0.73	0.69
0.6	0.90	0.95	0.91	0.61	0.93	0.89*
0.5	1.00	1.00	1.00	0.66	0.99*	0.93*

Table 3: Proportion of rejections for testing the treatment effect in LMM and GLMM across 100 simulations. The Wald test statistic is computed using our jackknife estimate of the variance. Proportions for GLMM with identity (id) and logit link functions are reported. *These configurations failed to fit the model in one simulation/jackknife iteration and so these powers were approximated over 99 simulations.

On average, fitting a single LMM and GLMM on the simulated data took less than one second using a standard laptop with a 2.3 GHz processor. On the other hand, fitting GEE on clusters of equal size took approximately 10 seconds and on clusters of unequal sizes took approximately 110 seconds. A default tolerance of 0.001 is used and no initial estimates are provided for the parameters (and so a GLMM is used to obtain initial estimates by default). From examining a few model fits, approximately four iterations are needed to fit a GEE. It is unclear why fitting GEE takes much longer in the unequal cluster size case. We assume that the runtime for the algorithm used by `gee()` is non-linear in the individual cluster sizes, and so a few clusters that are larger on average will significantly increase the computation time.

4.5 Discussion

Comparing our results to the ones reported by Hussey and Hughes', it appears that we were able to replicate their general findings but not their exact simulation procedure. Like in the original study, we find that when cluster sizes are equal, the differences in power across LMM, GEE and GLMM are not significantly noticeable (although some jackknife estimate of the variance is likely needed to maintain the size of the test in the case of GEE and GLMM). When cluster sizes are unequal, the power from LMM falters and so GEE and GLMM would be preferred.

We note that our reported powers in Table 2 are generally consistent (up to sampling variation) or otherwise greater than the powers reported by Hussey and Hughes. In particular, our powers for GLMM with an identity link function are generally much greater than theirs when RR < 1 in both the equal and unequal cluster size case. The powers for GLMM with the logit link function are smaller and arguably more consistent with that of Hussey and Hughes'. This leads us to believe that they may have used the default logit link for GEE and GLMM. Even taking this into consideration, there may still be a slight discrepancy in reported powers as seen in the case for LMM with unequal cluster sizes. It is difficult to gauge whether this is due to the additional variation from sampling cluster sizes, or if there is something else in their procedure that they do differently from ours (e.g., simulating the individual-level data in different ways, or using the output t-test rather than the Wald test).

From our results in Table 3, it may appear that our jackknife estimate of the variance did not help to maintain the size of the test for GLMM. It is difficult to determine whether this is due to sampling variation (which is further amplified as only 100 simulations were run in each configuration) or if our jackknife estimator was not properly constructed. Given more time, we would repeat this study over more simulations before making concrete conclusions.

5 Extensions of the standard model

The model examined by Hussey and Hughes (2007) can be considered the basic model as it makes relatively simple assumptions about the design and the effects. Modern literature has extended the basic model in various ways that better align with trial conditions that may be seen in practice. We present several common extensions to the model and discuss what they aim to address in this section.

5.1 Unequal cluster sizes

The case of unequal cluster sizes is one of the few variations of SW-CRTs that Hussey and Hughes discuss in their paper. The basic model assumes that all of the clusters have the same size, and this may have been a reasonable assumption for the EPT trial as there were likely more than enough participants in each LHJ during each time period. However, other trials may have a small population of study and may have difficulty achieving even the desired cluster size. In this case, sizes may vary between not only clusters but potentially also between time periods within a single cluster. A review of SW-CRTs published up to March 2016 found that almost half of the considered trials involved unequal cluster sizes (Kristunas et al., 2017).

Unequal cluster sizes does not change the basic model but leads to potentially differing variances across cluster means (given by $\mathrm{Var}(\bar{y}_{ij.}) = \tau^2 + \frac{\sigma^2}{N_{ij}}$). This affects cluster-level estimation of the model parameters as the correct weights in WLS depend on the true, unknown variance components. This consequently affects the power for testing the treatment effect, and Hussey and Hughes (2007) found that unequal cluster sizes resulted in the LMM method having weaker power compared to the GLMM and GEE methods in their simulation study.

Examples of recent literature that studied the impact of unequal cluster sizes on SW-CRT analysis include the works by Martin et al. (2019), Ouyang et al. (2020), Kasza et al. (2021), and Tian et al. (2021). While these studies all examine differing trial contexts, the general findings echo that of Hussey and Hughes' in that unequal cluster sizes affect the power (and efficiency of estimators), and that using methods that account for the unequal cluster sizes may partially recover some of the lost power (efficiency).

5.2 Delayed treatment effect

The case where there are delays in the treatment effect is also one of the variations that Hussey and Hughes discuss. The basic model assumes that once the intervention is applied to a cluster, its full effect on the measured response is immediate and visible. In reality, depending on what the intervention is, its effect on the response may only be partial in the short term and may take a longer time period in order for its full effect to manifest. Examples of trials where a delayed effect may be expected include the implementation of infant Hepatitis B vaccination programs for reducing incidences in the population (Hughes et al., 2015), and the use of repellents for reducing malaria in villages (Agius et al., 2020).

It is notable that one can modeling delayed treatment effects by allowing the time intervals in the SW-CRT to be long enough for the full treatment effect to be applied. However, if logistical constraints make extending the duration of the trial infeasible, then delayed effects can be modeled by modifying the parameters of the basic model. The most general delayed effects model takes the treatment indicator X_{ij} to be in $[0, 1]$ rather than as binary, in which case the treatment effect θ is then interpreted as the full effect. Further complications can be considered in the model, such as the length of the delay and whether the delay is known or needs to be estimated. These considerations would generally affect the model parameterization.

Hussey and Hughes (2007) empirically show that a delayed effect reduces the power (and we show this theoretically in Appendix A.5.2 using the within-cluster estimator as an example). While delayed treatment effects are a common point of consideration in the literature (in both the analysis of real trials and the development of SW-CRT theory), there does not appear to be much recent work where the primary objective is dedicated to studying its effect.

5.3 Non-normal response

The basic model assumes that the individual-level responses (and therefore the cluster-level responses) are normally distributed, but it may be the case that the individual-level responses are non-normal. Hussey and Hughes examine the EPT trial where the individual-level responses are binary (though their model assumes normal responses), and the responses may be generalized to distributions in the exponential family by modeling the mean response using a GLMM or GEE fitted with a particular link function. An example of a SW-CRT trial that involved non-normal outcomes is the *DECIDE-LVAD* trial that measured the effect of educational materials on patient decision making using a 10-point Likert scale (Allen et al., 2018).

As mentioned, the general approach for extending the basic model to non-normal responses is to use a GLMM or GEE with a link function g for the mean response, i.e., to use the model

$$g(\mathbb{E}[Y_{ijk}]) = \mu_{ij} \ .$$

The link function g is chosen based on how the mean response is to be modeled. For simple outcomes (e.g., the binary response in the EPT trial), cluster-level analyses may still be possible and meaningful. However, more complex outcomes will likely require individual-level analyses. Therefore, SW-CRTs featuring a non-normal response will generally use GLMM or GEE for fitting the model.

A recent review paper commented that there is still much room for development of SW-CRTs with non-normal responses in the literature (Li et al., 2021). Recent work have proposed formulas for power and sample size calculations in SW-CRT designs with binary outcomes (Wang et al., 2021) and discrete outcomes (Xia et al., 2021). From our perspective, it appears that there are still gaps with this topic in the literature. In particular, non-normal and non-discrete responses (e.g., survival times) in SW-CRT designs seem to be one direction that has not been explored much in the literature.

5.4 Cohort designs

The EPT trial is an example of a cross-sectional SW-CRT where the participants in a cluster at a particular time point are different from the participants in the same cluster at another time point. A cohort SW-CRT design differs from the cross-sectional design in that a cluster tracks the same participants over multiple time points. An open cohort design (e.g., ongoing recruitment of participants) tracks the same participants over a partial duration of the trial, while a closed cohort design tracks the same participants over the entire duration of the trial (except in the case of dropouts). An example of a cohort SW-CRT is the *INSTTEPP* trial that investigated if a boot camp translation process for self-management support (1) improved outcomes in chronic patients and (2) improved clinician and staff attitudes toward self-management (Nease Jr et al., 2018). The trial employed an open cohort design for the patients (where new patients may join the study part-way through) and a closed cohort design for the clinicians and staff (with the same participants being evaluated throughout the study).

As cohort designs imply potential repeated measurements for each participant, the general approach for extending the basic model to account for this is to add individual-level random effects, i.e.,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ik} + X_{ij}\theta + e_{ijk}$$

where $\gamma_{ik} \sim N(0, \nu^2)$ and with each distinct participant in a cluster i receiving their own index k. With the added individual-level random effects, an individual analysis using the GLMM or GEE approach would be the preferred approach as otherwise carrying out a WLS with the correct covariance structure would likely be difficult.

Recent work in the literature have examined various properties of cohort designs in varying contexts. For example, Kasza et al. (2020) proposed sample size and power formulas for general open cohort longitudinal CRTs. Li (2020) proposed sample size and power calculation procedures for cohort SW-CRTs with a decay correlation structure (e.g., when participants are expected to change over time or participants in each cluster are slowly replaced by new ones). The commentary by Hooper and Copas (2019) has called for a more fundamental change in modeling SW-CRTs with continuous recruitment and proposed thinking about time as a continuous variable.

5.5 Hierarchical designs

The SW-CRT designs discussed so far in this report can all be considered as a hierarchical design with two levels: the clusters and the participants within a cluster. It is possible that the data collection procedure in a trial may lead to more than two levels of clustering. For example, the CHANGE trial (Lescure et al., 2021) that investigated the effect of education and awareness activities on hand hygiene in nurses involved four levels of clustering. Data were collected on hand hygiene opportunities (level 1) for the nurses (level 2) in certain wards (level 3) of several nursing homes (level 4).

Following the idea of the basic model where random effects are used to model the correlation between the units within a cluster, the correlation due to higher-level clustering can also be incorporated in the model by introducing additional random effects. For example, if we have a three-level trial measuring patients (level 1) in a ward (level 2) in a hospital (level 3) and the visited ward is different at every time point, then a suitable model for this trial may be

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + X_{ij}\theta + e_{ijk}$$

where $\gamma_{ij} \sim N(0, \nu^2)$ (i.e., a cluster-time random effect). Depending on the complexity of the hierarchy, cluster-level analyses may be possible but otherwise an individual analysis using GLMM or GEE would be preferred.

Teerenstra et al. (2019) showed that the general impact of the multi-level structure is the inflation of the cluster mean variances at each level. They also proposed sample size and power calculation formulas for SW-CRTs with more than two levels, which were recently further developed in the work by Davis-Plourde et al. (2021).

5.6 Bayesian approaches

While the approaches for analyzing SW-CRTs have been predominantly frequentist, a number of Bayesian perspectives have appeared within the last few years. The Bayesian approach is a possible alternative to the standard approach for most SW-CRT designs, with the main difference being that priors are placed on what would be the fixed effects and hyperparameters of the basic model. Fitting the Bayesian model would also require a different approach from the frequentist one. Gibbs sampling, a Markov chain Monte Carlo method, is commonly used to fit the model in the Bayesian SW-CRT literature (Cunanan et al., 2016; Zhan et al., 2021).

The *British Columbia Telehealth Trial* is an example of one trial that has been analyzed through a Bayesian approach (Cunanan et al., 2016). The trial was a cross-sectional study that evaluated the effect of telehealth support on improving patient compliance with recommended disease-specific action plans. Cunanan et al. fitted a model using largely uninformative priors on the parameters and computed the power and size through a simulation study. It is noted that one of the strengths of Bayesian approaches over frequentist ones is the access to a full posterior distribution for inference as opposed to relying on point estimates and asymptotic results.

The more recent work by Zhan et al. (2021) explored how informative priors on the time effects in Bayesian models may avoid the overestimation of a required sample size in the design of SW-CRTs. Their results suggest that sample size calculations are consistent with that of frequentist calculations when uninformed priors are used, and that well-specified informative priors reduce the needed sample size. When the priors are misspecified, the introduced bias stays relatively small unless the prior is greatly misspecified. The ability to specify a prior is both a strength and a common point of criticism of Bayesian methods: informative priors may improve the efficiency of estimators relative to that of frequentist approaches, but the choice of a prior is subjective and misspecified priors may instead result in worse efficiency.

6 Conclusion

In this report, we summarized the work by Hussey and Hughes (2007), filled in missing details, provided a commentary and critique of the paper, attempted to replicate their simulation study, and discussed how the SW-CRT literature and the approaches to analysis have developed since the paper.

Our perspective of the paper is that it succeeds in its purpose as being an entry-point to SW-CRT designs and focuses on aspects that would be of practical interest (e.g., power) to someone who employs such trials. As a consequence of this choice, the main limitation is that the majority of technical details are removed from discussion and the considered trial contexts are narrowed in scope. We addressed these points in this report by providing the missing technical details and discussing how the basic model may be extended depending on the context of the trial.

In our attempt to clarify the details of their simulation study, we obtained results that are consistent with their general findings but were seemingly unable to replicate their exact results. The time constraints on this report also made it difficult to fully carry out the simulations as we would like to. We have documented our simulation procedure in Section 4 and created an R script that can be run to replicate our results.

As part of the discussion on extensions to the basic model, we drew on real SW-CRTs as examples and highlighted specific areas of SW-CRT theory that seem to be the focus of recent work. Our perception of the common trend across the SW-CRT literature is that the work by Hussey and Hughes set the foundation to the design and analysis of SW-CRTs, and that the majority of modern developments focus on extending or addressing the shortcomings of their methods.

References

Agius, P. A., Cutts, J. C., Han Oo, W., Thi, A., O'Flaherty, K., Zayar Aung, K., Kyaw Thu, H., Poe Aung, P., Mon Thein, M., Nyi Zaw, N., et al. (2020). Evaluation of the effectiveness of topical repellent distributed by village health volunteer networks against Plasmodium spp. infection in Myanmar: A stepped-wedge cluster randomised trial. *PLoS medicine*, *17*(8), e1003177.

Allen, L. A., McIlvennan, C. K., Thompson, J. S., Dunlay, S. M., LaRue, S. J., Lewis, E. F., Patel, C. B., Blue, L., Fairclough, D. L., Leister, E. C., et al. (2018). Effectiveness of an intervention supporting shared decision making for destination therapy left ventricular assist device: The DECIDE-LVAD randomized clinical trial. *JAMA internal medicine*, *178*(4), 520–529.

Bowden, R., Forbes, A. B., & Kasza, J. (2021). Inference for the treatment effect in longitudinal cluster randomized trials when treatment effect heterogeneity is ignored [PMID: 34569853]. *Statistical Methods in Medical Research*, 09622802211041754. https://doi.org/10.1177/09622802211041754

Brown, C. A., & Lilford, R. J. (2006). The stepped wedge trial design: A systematic review. *BMC medical research methodology*, *6*(1), 1–9.

Cunanan, K. M., Carlin, B. P., & Peterson, K. A. (2016). A practical Bayesian stepped wedge design for community-based cluster-randomized clinical trials: The British Columbia Telehealth Trial. *Clinical Trials*, *13*(6), 641–650.

Davis-Plourde, K., Taljaard, M., & Li, F. (2021). Sample size considerations for stepped wedge designs with subclusters. *Biometrics*. https://doi.org/https://doi.org/10.1111/biom.13596

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Zeger, S., et al. (2002). *Analysis of longitudinal data*. Oxford university press.

Golden, M. R., Kerani, R. P., Stenger, M., Hughes, J. P., Aubin, M., Malinski, C., & Holmes, K. K. (2015). Uptake and population-level impact of expedited partner therapy (EPT) on Chlamydia trachomatis and Neisseria gonorrhoeae: The Washington State community-level randomized trial of EPT. *PLOS medicine*, *12*(1), e1001777.

Harrison, L. J., Chen, T., & Wang, R. (2020). Power calculation for cross-sectional stepped wedge cluster randomized trials with variable cluster sizes. *Biometrics*, *76*(3), 951–962. https://doi.org/https://doi.org/10.1111/biom.13164

Hooper, R., & Copas, A. (2019). Stepped wedge trials with continuous recruitment require new ways of thinking. *Journal of clinical epidemiology*, *116*, 161–166.

Hughes, J. P., Granston, T. S., & Heagerty, P. J. (2015). Current issues in the design and analysis of stepped wedge trials. *Contemporary clinical trials*, *45*, 55–60.

Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary clinical trials*, *28*(2), 182–191.

Kasza, J., Bowden, R., & Forbes, A. B. (2021). Information content of stepped wedge designs with unequal cluster-period sizes in linear mixed models: Informing incomplete designs. *Statistics in Medicine*, *40*(7), 1736–1751.

Kasza, J., Hooper, R., Copas, A., & Forbes, A. B. (2020). Sample size and power calculations for open cohort longitudinal cluster randomized trials. *Statistics in medicine*, *39*(13), 1871–1883.

Kristunas, C., Morris, T., & Gray, L. (2017). Unequal cluster sizes in stepped-wedge cluster randomised trials: A systematic review. *BMJ open*, *7*(11), e017151.

Lescure, D., Haenen, A., de Greeff, S., Voss, A., Huis, A., & Hulscher, M. (2021). Exploring determinants of hand hygiene compliance in ltcfs: A qualitative study using flottorps' integrated checklist of determinants of practice. *Antimicrobial Resistance & Infection Control*, *10*(1), 1–11.

Li, F. (2020). Design and analysis considerations for cohort stepped wedge cluster randomized trials with a decay correlation structure. *Statistics in medicine*, *39*(4), 438–455.

Li, F., Hughes, J. P., Hemming, K., Taljaard, M., Melnick, E. R., & Heagerty, P. J. (2021). Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: An overview. *Statistical Methods in Medical Research*, *30*(2), 612–639.

Liao, X., Zhou, X., & Spiegelman, D. (2015). A note on "Design and analysis of stepped wedge cluster randomized trials". *Contemporary clinical trials*, *45*(Pt B), 338.

Martin, J. T., Hemming, K., & Girling, A. (2019). The impact of varying cluster size in cross-sectional stepped-wedge cluster randomised trials. *BMC medical research methodology*, *19*(1), 1–11.

Nease Jr, D. E., Daly, J. M., Dickinson, L. M., Fernald, D. H., Hahn, D. L., Levy, B. T., Michaels, L. C., Simpson, M. J., Westfall, J. M., & Fagnan, L. J. (2018). Impact of a boot camp translation intervention on self-management support in primary care: A report from the INSTTEPP trial and Meta-LARC consortium. *Journal of Patient-Centered Research and Reviews*, *5*(4), 256.

Ouyang, Y., Karim, M. E., Gustafson, P., Field, T. S., & Wong, H. (2020). Explaining the variation in the attained power of a stepped-wedge trial with unequal cluster sizes. *BMC medical research methodology*, *20*(1), 1–14.

Teerenstra, S., Taljaard, M., Haenen, A., Huis, A., Atsma, F., Rodwell, L., & Hulscher, M. (2019). Sample size calculation for stepped-wedge cluster-randomized trials with more than two levels of clustering [PMID: 31018678]. *Clinical Trials*, *16*(3), https://doi.org/10.1177/1740774519829053, 225–236. https://doi.org/10.1177/1740774519829053

Tian, Z., Preisser, J., Esserman, D., Turner, E., Rathouz, P., & Li, F. (2021). Impact of unequal cluster sizes for GEE analyses of stepped wedge cluster randomized trials with binary outcomes. *medRxiv*.

Wang, J., Cao, J., Zhang, S., & Ahn, C. (2021). Sample size and power analysis for stepped wedge cluster randomised trials with binary outcomes. *Statistical Theory and Related Fields*, *5*(2), 162–169. https://doi.org/10.1080/24754269.2021.1904094

Xia, F., Hughes, J. P., Voldal, E. C., & Heagerty, P. J. (2021). Power and sample size calculation for stepped-wedge designs with discrete outcomes. *Trials*, *22*(1), 1–10.

Zhan, D., Ouyang, Y., Xu, L., & Wong, H. (2021). Improving efficiency in the stepped-wedge trial design via Bayesian modeling with an informative prior for the time effects [PMID: 33813906]. *Clinical Trials*, *18*(3), https://doi.org/10.1177/1740774520980052, 295–302. https://doi.org/10.1177/1740774520980052

A Technical explanations

This appendix section includes missing technical details from the paper by Hussey and Hughes (2007) that we defer from our summary in Section 2 in order to avoid overwhelming the main ideas with the technicalities.

A.1 Intuition behind ICC and CV

The intraclass correlation (ICC) and coefficient of variation (CV) are two quantities commonly used to characterize the effect of within-cluster correlation in CRTs. The ICC is given by

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$$

and can be viewed as comparing the cluster variation (the numerator) to the individual variation (the denominator). The ICC is bounded between 0 and 1 where it is 0 if and only if $\tau^2 = 0$ (there is no variation between clusters) and 1 if and only if $\sigma^2 = 0$ (there is no variation between individuals). A quantity related to the ICC is the variance inflation factor $\nu = 1 + (N-1)\rho \geq 1$, which gets its name from rewriting the cluster mean variance as

$$\begin{aligned}
\mathrm{Var}(\bar{Y}_{ij.}) &= \frac{\sigma^2}{N} + \tau^2 \\
&= \frac{\tau^2 + \sigma^2}{N} + \tau^2 - \frac{\tau^2}{N} \\
&= \frac{\tau^2 + \sigma^2}{N} + \left(\frac{\tau^2(N-1)}{N}\right)\left(\frac{\tau^2 + \sigma^2}{\tau^2 + \sigma^2}\right) \\
&= \left(\frac{\tau^2 + \sigma^2}{N}\right)\left(1 + \frac{(N-1)\tau^2}{\tau^2 + \sigma^2}\right) \\
&= \left(\frac{\tau^2 + \sigma^2}{N}\right)\nu\ .
\end{aligned}$$

When the individuals are all independent and there is no variation between clusters ($\tau^2 = 0$), we have $\rho = 0$, $\nu = 1$, and $\mathrm{Var}(\bar{Y}_{ij.}) = \frac{\sigma^2}{N}$. Hence, any amount of cluster variation $\tau^2 > 0$ then "inflates" the mean variance relative to the independent case.

The CV is given by $c = \frac{\tau}{\mu}$ and can be viewed as the cluster standard deviation relative to the mean. The CV is unitless and therefore may be useful when comparing the degree of cluster variation in one trial to another without needing to account for scale.

A.2 Within-cluster estimator for treatment effect

Let $t_i \in \{1, \ldots, T-1\}$ be the last time point at which cluster i receives the control/existing intervention. When there are no time effects on the outcome, the within-cluster estimator for the treatment effect given by Hussey and Hughes (2007) is

$$\tilde{\theta} = \frac{1}{I}\sum_{i=1}^{I}\left(\frac{\sum_{j=t_i+1}^{T}\bar{Y}_{ij.}}{T - t_i} - \frac{\sum_{j=1}^{t_i}\bar{Y}_{ij.}}{t_i}\right)\ .$$

Each term in the sum corresponds to a particular cluster, and each term calculates the difference in the mean outcome between the intervention and the control periods of that cluster. Note that the estimator is only unbiased when there are no time effects, as

$$
\begin{align*}
\mathbb{E}[\tilde{\theta}] &= \frac{1}{I}\sum_{i=1}^{I}\left(\frac{\sum_{j=t_i+1}^{T}\mathbb{E}[\bar{Y}_{ij.}]}{T-t_i} - \frac{\sum_{j=1}^{t_i}\mathbb{E}[\bar{Y}_{ij.}]}{t_i}\right) \\
&= \frac{1}{I}\sum_{i=1}^{I}\left(\frac{\sum_{j=t_i+1}^{T}(\mu+\alpha_i+\beta_j+\theta)}{T-t_i} - \frac{\sum_{j=1}^{t_i}(\mu+\alpha_i+\beta_j)}{t_i}\right) \\
&= \frac{1}{I}\sum_{i=1}^{I}\left(\frac{\sum_{j=t_i+1}^{T}(\mu+\alpha_i+\theta)}{T-t_i} - \frac{\sum_{j=1}^{t_i}(\mu+\alpha_i)}{t_i}\right) \tag{*} \\
&= \frac{1}{I}\sum_{i=1}^{I}(\mu+\alpha_i+\theta-\mu-\alpha_i) \\
&= \theta
\end{align*}
$$

where the line $(*)$ follows from the assumption of no time effects. When there are time effects, the estimator is biased with the bias being

$$
\begin{align*}
\operatorname{bias}(\tilde{\theta},\theta) &= \mathbb{E}[\tilde{\theta}] - \theta \\
&= \frac{1}{I}\sum_{i=1}^{I}\left(\frac{\sum_{j=t_i+1}^{T}(\mu+\alpha_i+\beta_j+\theta)}{T-t_i} - \frac{\sum_{j=1}^{t_i}(\mu+\alpha_i+\beta_j)}{t_i}\right) - \theta \\
&= \frac{1}{I}\sum_{i=1}^{I}\left(\frac{\sum_{j=t_i+1}^{T}\beta_j}{T-t_i} + \theta - \frac{\sum_{j=1}^{t_i}\beta_j}{t_i}\right) - \theta \\
&= \frac{1}{I}\sum_{i=1}^{I}\left(\frac{\sum_{j=1}^{T}\beta_j X_{ij}}{T-t_i} - \frac{\sum_{j=1}^{T}\beta_j(1-X_{ij})}{t_i}\right) \\
&= \frac{1}{I}\sum_{j=1}^{T}\beta_j\sum_{i=1}^{I}\frac{t_iX_{ij}-(T-t_i)(1-X_{ij})}{t_i(T-t_i)} \\
&= \sum_{j=1}^{T}\beta_j\sum_{i=1}^{I}\frac{t_i-T(1-X_{ij})}{It_i(T-t_i)}
\end{align*}
$$

with the form in the last line being the one given by Hussey and Hughes.

A.3 Relative efficiency of WLS and within-cluster estimator

The relative efficiency of the WLS estimator $\hat{\theta}$ versus the within-cluster estimator $\tilde{\theta}$ is given by the inverse ratio of their variances, i.e.,

$$
\operatorname{efficiency}(\hat{\theta},\tilde{\theta}) = \frac{\operatorname{Var}(\tilde{\theta})}{\operatorname{Var}(\hat{\theta})}\,.
$$

The WLS estimator is more efficient than the within-cluster estimator if the ratio is greater than 1, and vice versa if the ratio is less than 1. Hussey and Hughes (2007) state that when there are no time effects, the ratio is

$$
\operatorname{efficiency}(\hat{\theta},\tilde{\theta}) = \frac{\sum_{i=1}^{I}\left(\frac{1}{t_i}+\frac{1}{T-t_i}\right)\left((ITU-U^2)\frac{\sigma^2}{N}+IT(TU-V)\tau^2\right)}{I^3\left(\frac{\sigma^2}{N}+T\tau^2\right)}\,.
$$

Liao et al. (2015) have shown that there is a missing factor in the denominator, and that the correct ratio is

$$
\operatorname{efficiency}(\hat{\theta},\tilde{\theta}) = \frac{\sum_{i=1}^{I}\left(\frac{1}{t_i}+\frac{1}{T-t_i}\right)\left((ITU-U^2)\frac{\sigma^2}{N}+IT(TU-V)\tau^2\right)}{I^3T\left(\frac{\sigma^2}{N}+T\tau^2\right)}\,.
$$

We show how this quantity is obtained in the following subsections.

A.3.1 Variance of within-cluster estimator

The variance of the within-cluster estimator $\tilde{\theta}$ of θ is given by

$$\begin{aligned}\operatorname{Var}(\tilde{\theta}) &= \frac{1}{I^2}\sum_{i=1}^{I}\operatorname{Var}\left(\frac{\sum_{j=t_i+1}^{T}\bar{Y}_{ij.}}{T-t_i} - \frac{\sum_{j=1}^{t_i}\bar{Y}_{ij.}}{t_i}\right)\\ &= \frac{1}{I^2}\sum_{i=1}^{I}\operatorname{Var}\left(\frac{\sum_{j=t_i+1}^{T}\left(\beta_j+\theta+\frac{1}{N}\sum_{k=1}^{N}e_{ijk}\right)}{T-t_i} - \frac{\sum_{j=1}^{t_i}\left(\beta_j+\frac{1}{N}\sum_{k=1}^{N}e_{ijk}\right)}{t_i}\right) \qquad (*)\\ &= \frac{\sigma^2}{NI^2}\sum_{i=1}^{I}\left(\frac{1}{T-t_i}+\frac{1}{t_i}\right)\end{aligned}$$

where the line $(*)$ follows because μ and α_i cancel out between the two terms.

A.3.2 Variance of WLS estimator

Let $\hat{\theta}$ denote the WLS estimator of θ extracted from the WLS solution $\hat{\eta} = (\hat{\mu}, \hat{\beta}_1, \dots, \hat{\beta}_{T-1}, \hat{\theta})$. Under the assumption that there are no time effects, the WLS solution has the form $\hat{\eta} = (\hat{\mu}, \hat{\theta})$ and is obtained from

$$\hat{\eta} = (\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{y}$$

where $\mathbf{X}$ is the $IT \times 2$ design matrix given by

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11}\\ \vdots & \vdots \\ 1 & X_{IT}\end{bmatrix},$$

$\mathbf{y}$ is the vector of cluster means of length IT, and $\mathbf{V} = \operatorname{Var}(\mathbf{y})$ is the $IT \times IT$ block diagonal matrix with each $T \times T$ block $\mathbf{V}_i$, $i \in \{1, \dots, I\}$, describing the correlation structure of a cluster over time given by

$$\mathbf{V}_i = \begin{bmatrix} \tau^2+\frac{\sigma^2}{N} & \tau^2 & \dots & \tau^2\\ \tau^2 & \ddots & & \vdots\\ \vdots & & \ddots & \tau^2\\ \tau^2 & \dots & \tau^2 & \tau^2+\frac{\sigma^2}{N}\end{bmatrix}.$$

Let $\mathbf{e}_i$ represent the unit column vector with a 1 in the i-th position. The variance of $\hat{\theta}$ is given by

$$\begin{aligned}\operatorname{Var}(\hat{\theta}) &= \mathbf{e}_2^\top(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}^{-1}\operatorname{Var}(\mathbf{y})\mathbf{V}^{-1}\mathbf{X}^\top(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{e}_2\\ &= \mathbf{e}_2^\top(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{e}_2\end{aligned}$$

A closed-form for the variance of $\hat{\theta}$ is possible for the LMM when $X_{ij} \in \{0, 1\}$ (Hussey & Hughes, 2007). We follow the derivation by Liao et al. (2015) for the closed-form while clarifying their steps in the process.

Note that we can rewrite $\mathbf{V}_i = \frac{\sigma^2}{N}\mathbf{I}_T + \tau^2 \mathbf{1}_T\mathbf{1}_T^\top$. By the Sherman-Morrison formula, we have

$$\begin{aligned}
\mathbf{V}_i^{-1} &= \frac{N}{\sigma^2}\mathbf{I}_T - \frac{\frac{N^2\tau^2}{\sigma^4}}{1 + \frac{N\tau^2}{\sigma^2}\mathbf{1}_T^\top \mathbf{I}_T \mathbf{1}_T}\mathbf{I}_T\mathbf{1}_T\mathbf{1}_T^\top\mathbf{I}_T \\
&= \frac{N}{\sigma^2}\left(\mathbf{I}_T - \frac{N\tau^2}{\sigma^2 + NT\tau^2}\mathbf{1}_T\mathbf{1}_T^\top\right) \\
&= \frac{N^2}{\sigma^4 + NT\sigma^2\tau^2}\begin{bmatrix} \frac{\sigma^2}{N} + (T-1)\tau^2 & -\tau^2 & \dots & -\tau^2 \\ -\tau^2 & \ddots & & \vdots \\ \vdots & & \ddots & -\tau^2 \\ -\tau^2 & \dots & -\tau^2 & \frac{\sigma^2}{N} + (T-1)\tau^2 \end{bmatrix}
\end{aligned}$$

We then have

$$\begin{aligned}
\mathbf{V}^{-1}\mathbf{X} &= \frac{N^2}{\sigma^4 + NT\sigma^2\tau^2}\begin{bmatrix} \frac{\sigma^2}{N} & \left(\frac{\sigma^2}{N} + (T-1)\tau^2\right)X_{11} - \tau^2\sum_{j=2}^T X_{1j} \\ \vdots & \vdots \\ \frac{\sigma^2}{N} & \left(\frac{\sigma^2}{N} + (T-1)\tau^2\right)X_{IT} - \tau^2\sum_{j=1}^{T-1} X_{Ij} \end{bmatrix}, \\
\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X} &= \frac{N^2}{\sigma^4 + NT\sigma^2\tau^2}\begin{bmatrix} \frac{\sigma^2}{N}IT & \frac{\sigma^2}{N}U \\ \frac{\sigma^2}{N}U & W \end{bmatrix}
\end{aligned}$$

with

$$\begin{aligned}
U &= \sum_{i=1}^I \sum_{j=1}^T X_{ij}\ , \\
W &= \sum_{i=1}^I \sum_{j=1}^T \left(\left(\frac{\sigma^2}{N} + (T-1)\tau^2\right)X_{ij}^2 - \tau^2 X_{ij}\left(\sum_{k=1}^T X_{ik} - X_{ij}\right)\right) \\
&= \sum_{i=1}^I \sum_{j=1}^T \left(\left(\frac{\sigma^2}{N} + T\tau^2\right)X_{ij}^2 - \tau^2 X_{ij}\sum_{k=1}^T X_{ik}\right) \\
&= \left(\frac{\sigma^2 + NT\tau^2}{N}\right)\sum_{i=1}^I \sum_{j=1}^T X_{ij}^2 - \tau^2 \sum_{i=1}^I \sum_{j=1}^T \left(X_{ij}^2 + \sum_{k \neq j}^T X_{ij}X_{ik}\right) \\
&= \left(\frac{\sigma^2 + NT\tau^2}{N}\right)U - \tau^2 \sum_{i=1}^I \left(\sum_{j=1}^T X_{ij}\right)^2
\end{aligned}$$

where U in the last line follows because $X_{ij} = X_{ij}^2 \in \{0,1\}$. Let $V = \sum_{i=1}^I \left(\sum_{j=1}^T X_{ij}\right)^2$. Therefore,

$$\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X} = \begin{bmatrix} \frac{NIT}{\sigma^2+NT\tau^2} & \frac{NU}{\sigma^2+NT\tau^2} \\ \frac{NU}{\sigma^2+NT\tau^2} & \frac{NU}{\sigma^2} - \frac{N^2\tau^2 V}{\sigma^4+NT\sigma^2\tau^2} \end{bmatrix}.$$

By the formula for the inverse of a 2 × 2 matrix, we then have

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}) &= \mathbf{e}_2^\top(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{e}_2 \\
&= \left(\frac{NIT}{\sigma^2+NT\tau^2}\right)\left(\left(\frac{NIT}{\sigma^2+NT\tau^2}\right)\left(\frac{NU}{\sigma^2}-\frac{N^2\tau^2V}{\sigma^4+NT\sigma^2\tau^2}\right)-\left(\frac{NU}{\sigma^2+NT\tau^2}\right)^2\right)^{-1} \\
&= \left(\frac{NU\sigma^2+N^2\tau^2(UT-V)}{\sigma^4+NT\sigma^2\tau^2}-\frac{NU^2}{IT(\sigma^2+NT\tau^2)}\right)^{-1} \\
&= \frac{IT\frac{\sigma^2}{N}\left(\frac{\sigma^2}{N}+T\tau^2\right)}{(ITU-U^2)\frac{\sigma^2}{N}+IT(UT-V)\tau^2}\ .
\end{aligned}$$

A.3.3 Relative efficiency

From the previous sections, the relative efficiency of the WLS and the within-cluster estimator is then directly obtained from the ratio

$$\begin{aligned}
\mathrm{efficiency}(\hat{\theta},\tilde{\theta}) &= \frac{\mathrm{Var}(\tilde{\theta})}{\mathrm{Var}(\hat{\theta})} \\
&= \frac{\sigma^2}{NI^2}\sum_{i=1}^{I}\left(\frac{1}{T-t_i}+\frac{1}{t_i}\right)\left(\frac{IT\frac{\sigma^2}{N}\left(\frac{\sigma^2}{N}+T\tau^2\right)}{(ITU-U^2)\frac{\sigma^2}{N}+IT(UT-V)\tau^2}\right)^{-1} \\
&= \frac{\sum_{i=1}^{I}\left(\frac{1}{T-t_i}+\frac{1}{t_i}\right)\left((ITU-U^2)\frac{\sigma^2}{N}+IT(UT-V)\tau^2\right)}{I^3T\left(\frac{\sigma^2}{N}+T\tau^2\right)}\ .
\end{aligned}$$

A.3.4 Time effect on efficiency

We reconstruct the proof for the efficiency of the WLS and within-cluster estimators under the assumption of no time effects in this section.

Proposition 1 (Liao et al. (2015) with minor errors corrected). *Under the model described in Section 2.3.2 and under the assumption that there are no time effects, we have*

$$\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\tilde{\theta})\ .$$

Proof. Note that we can rewrite the quantities

$$\begin{aligned}
U &= \sum_{i=1}^{I}\sum_{j=1}^{T}X_{ij} = \sum_{i=1}^{I}(T-t_i) = IT - \sum_{i=1}^{I}t_i\ , \\
V &= \sum_{i=1}^{I}\left(\sum_{j=1}^{T}X_{ij}\right)^2 = \sum_{i=1}^{I}(T-t_i)^2 = IT^2 - 2T\sum_{i=1}^{I}t_i + \sum_{i=1}^{I}t_i^2\ .
\end{aligned}$$

Hence, we have

$$
\begin{aligned}
\text{efficiency}(\hat{\theta},\tilde{\theta}) &= \frac{\sum_{i=1}^{I}\left(\frac{1}{T-t_i}+\frac{1}{t_i}\right)\left((ITU-U^2)\frac{\sigma^2}{N}+IT(UT-V)\tau^2\right)}{I^3T\left(\frac{\sigma^2}{N}+T\tau^2\right)} \\
&= T\sum_{i=1}^{I}\frac{1}{t_i(T-t_i)}\left(\frac{(U-\frac{U^2}{IT})\frac{\sigma^2}{N}+(UT-V)\tau^2}{I^2\left(\frac{\sigma^2}{N}+T\tau^2\right)}\right) \\
&= T\left(\sum_{i=1}^{I}\frac{1}{t_i(T-t_i)}\right)\left(\frac{\left(\sum_{i=1}^{I}t_i-\frac{\left(\sum_{i=1}^{I}t_i\right)^2}{IT}\right)\frac{\sigma^2}{N}+(UT-V)\tau^2}{I^2\left(\frac{\sigma^2}{N}+T\tau^2\right)}\right) \\
&= T\left(\sum_{i=1}^{I}\frac{1}{t_i(T-t_i)}\right)\left(\frac{\left(\sum_{i=1}^{I}t_i-\frac{\left(\sum_{i=1}^{I}t_i\right)^2}{IT}\right)\frac{\sigma^2}{N}+\left(\sum_{i=1}^{I}t_i-\frac{\sum_{i=1}^{I}t_i^2}{T}\right)T\tau^2}{I^2\left(\frac{\sigma^2}{N}+T\tau^2\right)}\right) .
\end{aligned}
$$

Let $\mathbf{e}_I$ be a vector of ones and $\mathbf{t}=(t_1,\ldots,t_I)$. By the Cauchy-Schwarz inequality,

$$
\sum_{i=1}^{I}t_i^2=\|\mathbf{t}\|^2\geq\frac{|\mathbf{e}_I^\top\mathbf{t}|^2}{\|\mathbf{e}_I\|^2}=\frac{\left(\sum_{i=1}^{I}t_i\right)^2}{I} .
$$

Then

$$
\begin{aligned}
\text{efficiency}(\hat{\theta},\tilde{\theta}) &\geq T\left(\sum_{i=1}^{I}\frac{1}{t_i(T-t_i)}\right)\left(\frac{\left(\sum_{i=1}^{I}t_i-\frac{\sum_{i=1}^{I}t_i^2}{T}\right)\frac{\sigma^2}{N}+\left(\sum_{i=1}^{I}t_i-\frac{\sum_{i=1}^{I}t_i^2}{T}\right)T\tau^2}{I^2\left(\frac{\sigma^2}{N}+T\tau^2\right)}\right) \\
&= \left(\sum_{i=1}^{I}\frac{1}{t_i(T-t_i)}\right)\left(\frac{T\sum_{i=1}^{I}t_i-\sum_{i=1}^{I}t_i^2}{I^2}\right) \\
&= \frac{1}{I^2}\left(\sum_{i=1}^{I}\frac{1}{t_i(T-t_i)}\right)\left(\sum_{i=1}^{I}t_i(T-t_i)\right) .
\end{aligned}
$$

Again, by the Cauchy-Schwarz inequality,

$$
\left(\sum_{i=1}^{I}\frac{1}{t_i(T-t_i)}\right)^{\frac{1}{2}}\left(\sum_{i=1}^{I}t_i(T-t_i)\right)^{\frac{1}{2}}\geq\left|\sum_{i=1}^{I}\sqrt{\frac{t_i(T-t_i)}{t_i(T-t_i)}}\right|=I .
$$

Thus,

$$
\text{efficiency}(\hat{\theta},\tilde{\theta})=\frac{\text{Var}(\tilde{\theta})}{\text{Var}(\hat{\theta})}\geq 1 .
$$

□

Note that even in the case $\tau^2=0$, the above result still appears to hold unlike what Hussey and Hughes (2007) claim.

Corollary 2. *Under the model described in Section 2.3.2 and under the assumptions that there are no time effects and that $\tau^2=0$, we have*

$$
\text{Var}(\hat{\theta})\leq\text{Var}(\tilde{\theta}) .
$$

Proof. The efficiency when $\tau^2 = 0$ is

$$\text{efficiency}(\hat{\theta}, \tilde{\theta}) = \left(\sum_{i=1}^{I} \frac{1}{t_i(T - t_i)} \right) \left(\frac{T \sum_{i=1}^{I} t_i - \frac{\left(\sum_{i=1}^{I} t_i\right)^2}{I}}{I^2} \right) .$$

Then by the same argument as in Proposition 1,

$$\text{efficiency}(\hat{\theta}, \tilde{\theta}) = \frac{\text{Var}(\tilde{\theta})}{\text{Var}(\hat{\theta})} \geq 1 .$$

□

A.4 Wald test and power

Hussey and Hughes (2007) prescribe using a Wald test to obtain an approximate power for testing the hypothesis $H_0 : \theta = 0$ versus $H_a : \theta = \theta_a$. For some estimator $\hat{\theta}$ of θ that is normally-distributed (either exactly under assumptions or approximately based on large samples), the test statistic in the Wald test is

$$Z = \frac{\hat{\theta}}{\sqrt{\text{Var}(\hat{\theta})}}$$

which has an (approximate) standard normal distribution under H_0. Under H_a, the statistic Z has an (approximate) normal distribution with mean $\frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}}$ and variance 1. Let $Z_{1-\frac{\alpha}{2}}$ be the $\left(1 - \frac{\alpha}{2}\right)$-th critical value of the standard normal distribution for significance level α. The power of the two-tailed test is then

$$\begin{aligned}
\mathbb{P}\left(Z < -Z_{1-\frac{\alpha}{2}} \,\middle|\, H_a\right) + \mathbb{P}\left(Z > Z_{1-\frac{\alpha}{2}} \,\middle|\, H_a\right) &= \mathbb{P}\left(Z - \frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}} < -Z_{1-\frac{\alpha}{2}} - \frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}} \,\middle|\, H_a \right) \\
&\quad + \mathbb{P}\left(Z - \frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}} > Z_{1-\frac{\alpha}{2}} - \frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}} \,\middle|\, H_a \right) \\
&= \Phi\left(-\frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}} - Z_{1-\frac{\alpha}{2}} \right) + 1 - \Phi\left(Z_{1-\frac{\alpha}{2}} - \frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}} \right) \\
&= \Phi\left(-\frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}} - Z_{1-\frac{\alpha}{2}} \right) + \Phi\left(\frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}} - Z_{1-\frac{\alpha}{2}} \right)
\end{aligned}$$

where Φ is the cumulative distribution function of the standard normal. Notice that if $\theta_a > 0$ and is not too small, then the first term is approximately 0 and so the power is approximately

$$\mathbb{P}\left(Z < -Z_{1-\frac{\alpha}{2}} \,\middle|\, H_a\right) + \mathbb{P}\left(Z > Z_{1-\frac{\alpha}{2}} \,\middle|\, H_a\right) \approx \mathbb{P}\left(Z > Z_{1-\frac{\alpha}{2}} \,\middle|\, H_a\right) = \Phi\left(\frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}} - Z_{1-\frac{\alpha}{2}} \right) ,$$

which is the power given by Hussey and Hughes (2007). The other term dominates when $\theta_a < 0$ and is not too small (in absolute value). Note that this calculation is also approximate if Z is only approximately normally distributed or if $\text{Var}(\hat{\theta})$ needs to be estimated.

A.5 Measured time points and delayed treatment effect on power

It can be seen from the power calculation in Appendix A.4 that the power depends on the variance of the estimator $\hat{\theta}$. For estimators that have a relatively large variance, the power decreases to the significance level where

$$\Phi\left(-\frac{\theta_a}{\sqrt{\mathrm{Var}(\hat{\theta})}} - Z_{1-\frac{\alpha}{2}}\right) + \Phi\left(\frac{\theta_a}{\sqrt{\mathrm{Var}(\hat{\theta})}} - Z_{1-\frac{\alpha}{2}}\right) \approx 2\Phi\left(-Z_{1-\frac{\alpha}{2}}\right) = \alpha\ .$$

For estimators that have a relatively small variance, one of the terms dominate and so a power greater than the significance level can be expected.

From the above, it then follows that design factors and assumptions that affect the variance of the estimator will also affect the power of the study. Hussey and Hughes (2007) briefly discuss how the number of measured time points and delays in the treatment effect affect power. We aim to provide more insight on their discussion in the following sections.

A.5.1 Number of time points

Hussey and Hughes (2007) state that the optimal power is achieved when only one cluster crosses over at each time point. We use the within-cluster estimator $\tilde{\theta}$ to illustrate this point. From Appendix A.3.1, the variance of the estimator is

$$\mathrm{Var}(\tilde{\theta}) = \frac{\sigma^2}{NI^2} \sum_{i=1}^{I} \left(\frac{1}{T - t_i} + \frac{1}{t_i}\right)\ .$$

Suppose that all $I > 2$ clusters are assigned an unique crossover time ($t_i = i$ without loss of generality) and that $T = I + 1$. Suppose that in another trial, the clusters are measured over $T - 1$ time points and the I-th cluster shares its crossover time with another cluster $j \in \{1, \ldots, I - 1\}$. The time factor in the estimator variance for this other trial with fewer time points is then

$$\begin{aligned}
\sum_{i=1}^{I-1} \left(\frac{1}{T - t_i - 1} + \frac{1}{t_i}\right) + \frac{1}{T - t_j - 1} + \frac{1}{t_j} &= \sum_{i=1}^{I-1} \left(\frac{1}{I - i} + \frac{1}{i}\right) + \frac{1}{I - j} + \frac{1}{j} \\
&> \sum_{i=1}^{I-1} \left(\frac{1}{I - i} + \frac{1}{i}\right) + \frac{1}{I} + \frac{1}{I} \\
&= \sum_{i=1}^{I} \left(\frac{1}{I - i + 1} + \frac{1}{i}\right) \\
&= \sum_{i=1}^{I} \left(\frac{1}{T - t_i} + \frac{1}{t_i}\right)\ .
\end{aligned}$$

Similar arguments can be made for trials with even fewer time points. Thus, keeping everything but the number of time points fixed, the variance of the estimator is smallest when each cluster crosses over at its own time point. The increase in variance due to a reduced number of time points leads to a decrease in power.

A.5.2 Delayed treatment effect

Hussey and Hughes (2007) state that delays in the treatment effect reduce power. Delayed treatment effects can be modeled by allowing X_{ij} to be in $[0, 1]$. We again use a within-cluster estimator $\tilde{\theta}$ to illustrate this point, though note that the estimator given by Hussey and Hughes needs to be modified to account for the delay in the treatment effect. Suppose that the delays X_{ij} are known and that $X_{ij} \in (0, 1)$ for at least one cluster i and time point j. Then a (modified) unbiased estimator (assuming that there are no separate time effects, i.e., $\beta_j = 0$) is given by

$$\tilde{\theta} = \left(\sum_{i=1}^{I} \sum_{j=1}^{T} \frac{X_{ij}}{T - t_i} \right)^{-1} \sum_{i=1}^{I} \left(\frac{\sum_{j=t_i+1}^{T} \bar{Y}_{ij.}}{T - t_i} - \frac{\sum_{j=1}^{t_i} \bar{Y}_{ij.}}{t_i} \right) .$$

This estimator is unbiased as

$$\begin{aligned} \mathbb{E}[\tilde{\theta}] &= \left(\sum_{i=1}^{I} \sum_{j=1}^{T} \frac{X_{ij}}{T - t_i} \right)^{-1} \sum_{i=1}^{I} \left(\frac{\sum_{j=t_i+1}^{T} \mathbb{E}[\bar{Y}_{ij.}]}{T - t_i} - \frac{\sum_{j=1}^{t_i} \mathbb{E}[\bar{Y}_{ij.}]}{t_i} \right) \\ &= \left(\sum_{i=1}^{I} \sum_{j=1}^{T} \frac{X_{ij}}{T - t_i} \right)^{-1} \sum_{i=1}^{I} \left(\frac{\sum_{j=t_i+1}^{T} (\mu + \alpha_i + X_{ij}\theta)}{T - t_i} - \frac{\sum_{j=1}^{t_i} (\mu + \alpha_i)}{t_i} \right) \\ &= \theta \left(\sum_{i=1}^{I} \sum_{j=1}^{T} \frac{X_{ij}}{T - t_i} \right)^{-1} \sum_{i=1}^{I} \sum_{j=t_i+1}^{T} \frac{X_{ij}}{T - t_i} \\ &= \theta \end{aligned}$$

where the last line follows because $X_{ij} = 0$ for $j \in \{1, \ldots, t_i\}$. It follows from the derivation in Appendix A.3.1 that the variance of the estimator is

$$\mathrm{Var}(\tilde{\theta}) = \frac{\sigma^2}{N} \left(\sum_{i=1}^{I} \sum_{j=1}^{T} \frac{X_{ij}}{T - t_i} \right)^{-2} \sum_{i=1}^{I} \left(\frac{1}{T - t_i} + \frac{1}{t_i} \right) .$$

Note that by assumption,

$$\sum_{i=1}^{I} \sum_{j=1}^{T} \frac{X_{ij}}{T - t_i} = \sum_{i=1}^{I} \sum_{j=t_i+1}^{T} \frac{X_{ij}}{T - t_i} < \sum_{i=1}^{I} \sum_{j=t_i+1}^{T} \frac{1}{T - t_i} = \sum_{i=1}^{I} 1 = I.$$

Thus,

$$\mathrm{Var}(\tilde{\theta}) = \frac{\sigma^2}{N} \left(\sum_{i=1}^{I} \sum_{j=1}^{T} \frac{X_{ij}}{T - t_i} \right)^{-2} \sum_{i=1}^{I} \left(\frac{1}{T - t_i} + \frac{1}{t_i} \right) > \frac{\sigma^2}{NI^2} \sum_{i=1}^{I} \left(\frac{1}{T - t_i} + \frac{1}{t_i} \right)$$

where the right-hand side of the inequality is the variance of the within-cluster estimator in the case of no delays in treatment effects. The increase in variance due to the delay in treatment effect leads to a decrease in power.

B Example generated datasets

This appendix section visualizes example datasets generated through the (equal cluster size) simulation procedure described in Section 4.1.

Figure 2 shows two example cluster-level datasets that are obtained from aggregating the individual-level data. A faint stepped wedge pattern may be observed in the plot on the right (simulated with RR = 0.5) in contrast to the plot on the left (simulated with RR = 1), but overall, the variation across cluster-times (each containing 100 individual units) overwhelm any noticeable patterns.

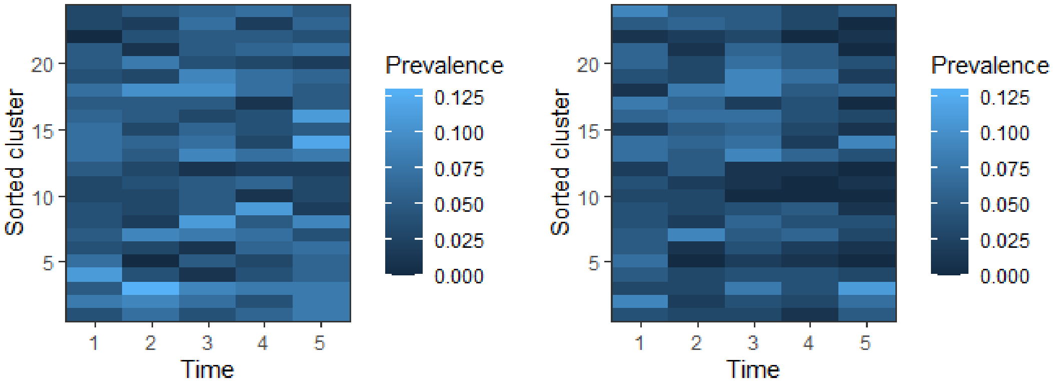

Figure 2: Cluster-level data simulated for RR = 1 (left) and RR = 0.5 (right). The clusters are sorted such that clusters 1–6 cross over at time 2, clusters 7–12 cross over at time 3, and so on. The prevalence at each cluster-time was computed from 100 individual-level units.

For comparison, Figure 3 shows the same plots but with the prevalence computed over 1000 units at each cluster-time. The stepped wedge pattern is now clearly noticeable when RR = 0.5.

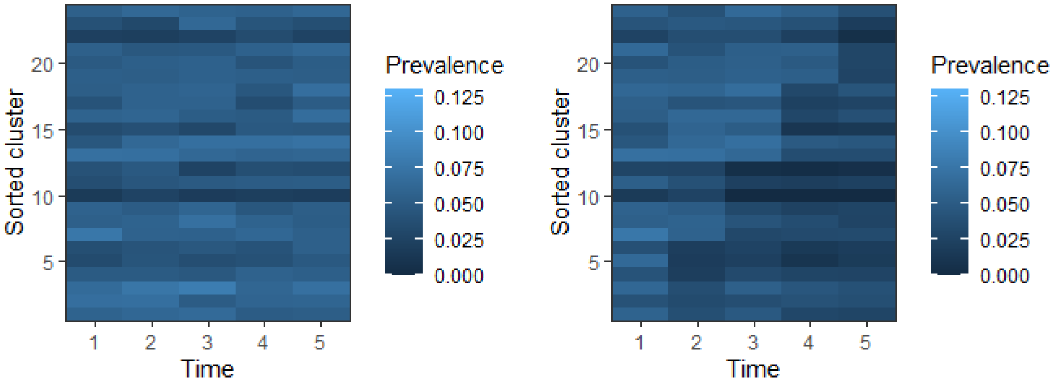

Figure 3: Cluster-level data simulated for RR = 1 (left) and RR = 0.5 (right). The clusters are sorted such that clusters 1–6 cross over at time 2, clusters 7–12 cross over at time 3, and so on. The prevalence at each cluster-time was computed from 1000 individual-level units.