

# Stepped Wedge Cluster Randomized Trials

STAT 548 Qualifying Paper

Kenny Chiu

November 11, 2021

**Abstract.** TODO

## 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 standard 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 view and critique of the paper; Section 4 describes modern common extensions to the standard SW-CRT model; and Section 5 concludes this report with a discussion. Appendix A includes added derivations from Section 2 to avoid disrupting the flow of the summary.

## 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 that we feel are missing from the paper.

### 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 not convenient 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 withholding the new intervention from certain

clusters would be problematic. Therefore, the main objective of Hussey and Hughes’s 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 possibly different treatments and switches 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 from the paper clearly illustrates the differences between the discussed CRT designs.

Parallel		Time	Crossover		Time	Stepped wedge		Time				
		1			1 2			1	2	3	4	5
Cluster	1	1	Cluster	1	1 0	Cluster	1	0	1	1	1	1
	2	1		2	1 0		2	0	0	1	1	1
	3	0		3	0 1		3	0	0	0	1	1
	4	0		4	0 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 the model for a SW-CRT in detail under a specific setting. How the model changes to varying study 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 4.

## 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 2012 *Washington State Community Expedited Partner Treatment (EPT) Trial*. 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, the individual LHJs are the clusters, and the individuals in a cluster are the women who were tested in a participating clinic during a particular timeframe. It is important to note that the women within a single cluster differ at different time points over the trial (i.e., the study uses a cross-sectional 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.

Another aspect to consider is the form of the outcome. In the EPT trial, the measured outcomes are prevalence of chlamydia and incidence of gonorrhea, both of which are continuous measures. Thus, based on the EPT trial, Hussey and Hughes focus on the properties of a continuous outcome SW-CRT model that assumes a cross-sectional design and fixed cluster sizes.

### 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, we can define the mean for cluster  $i$  at time  $j$  as

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

where

- $\mu$  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,
- $\beta_j$  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
- $\theta$  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 4.

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 fill in the missing details of key ideas 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  $\theta$ . When the variance components  $\tau^2$  and  $\sigma^2$  are known, a cluster-level estimation of  $\theta$  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  $\tau^2$  and  $\sigma^2$  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. 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  $\theta$  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  $\theta$  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 (unless  $\tau^2 = 0$ ). When there are time effects, the within-cluster estimator is more efficient (but likely biased). We provide further details for these statements in Appendix A.3.

### 2.4.2 Power and relevant factors

To test the hypothesis  $H_0 : \theta = 0$  against a simple alternative  $H_a : \theta = \theta_a$  in the general case, Hussey and Hughes (2007) prescribe using a Wald test with the test statistic

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

The power for a two-tailed test of size  $\alpha$  is then

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

where  $\Phi$  is the cumulative distribution function of the standard normal and  $Z_{1-\frac{\alpha}{2}}$  is the  $(1 - \frac{\alpha}{2})$ -quantile of the standard normal. 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.

## 3 Critical appraisal

We comment on the paper by Hussey and Hughes (2007) and critique its 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 such as analysis and power that are of practical relevance. 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.

In our opinion, a weakness of the paper is in its presentation and 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 in the paper with the exception of the power calculation procedure, the factors affecting power, and the simulation study are synthesized overviews of

existing ideas. 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).

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 addresses the breadth of extensions to the standard SW-CRT model that are now commonly seen in modern SW-CRT literature.

## 4 Extensions of the standard model

### TODO

- cluster sizes
- delayed treatment effect
- cluster-time random effects
- non-normal response
- cross-sectional vs cohort
- bayesian

## 5 Discussion

## References

- 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.
- Davis-Plourde, K., Taljaard, M., & Li, F. (2021). Sample size considerations for stepped wedge designs with subclusters. *Biometrics*, n/a(n/a). <https://doi.org/https://doi.org/10.1111/biom.13596>
- 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>
- Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary clinical trials*, 28(2), 182–191.
- 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.

## A Appendix

TODO

### A.1 ICC and CV

### A.2 Within-cluster estimator for treatment effect

Let  $t_i \in \{1, \dots, T-1\}$  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

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

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

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

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 the variances, i.e.,

$$\text{efficacy}(\hat{\theta}, \tilde{\theta}) = \frac{\text{Var}(\tilde{\theta})}{\text{Var}(\hat{\theta})}.$$

Hussey and Hughes (2007) state that when there are no time effects, the ratio is

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

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

#### A.3.1 Variance of within-cluster estimator

TODO

$$\text{Var}(\tilde{\theta}) = \frac{\sigma^2}{NI^2} \sum_i \left( \frac{1}{t_i} + \frac{1}{T-t_i} \right)$$

#### A.3.2 Variance of WLS estimator

Let  $\hat{\theta}$  denote the WLS estimator of  $\theta$  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  $\mathbf{V} = \text{Var}(\mathbf{y})$  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

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

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

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

We then have

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

where

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

where 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,

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

By the formula for the inverse of a  $2 \times 2$  matrix, we then have

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

TODO

#### A.4 Wald test and power

TODO

Consider testing the hypothesis  $H_0 : \theta = 0$  versus  $H_a : \theta = \theta_a$ . A Wald test may be based on  $Z = \frac{\hat{\theta}}{\sqrt{\text{Var}(\hat{\theta})}}$  where  $\hat{\theta}$  is from  $\hat{\eta}$ . The approximate power for a two-tailed test of size  $\alpha$  is

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

where  $\Phi$  is the cumulative standard normal distribution and  $Z_{1-\alpha/2}$  is the  $(1 - \alpha/2)$ -th quantile of the standard normal distribution.  $\text{Var}(\hat{\theta})$  is an element of  $(\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1}$  but may be possible to express in the closed form

$$\text{Var}(\hat{\theta}) = \frac{I\sigma^2 \left( \frac{\sigma^2}{N} + T\tau^2 \right)}{(IU - W)\sigma^2 + N(U^2 + ITU - TW - IV)\tau^2}$$

where

$$\begin{aligned}
U &= \sum_{ij} X_{ij} \\
W &= \sum_j \left( \sum_i X_{ij} \right)^2 \\
V &= \sum_i \left( \sum_j X_{ij} \right)^2
\end{aligned}$$

Note: under  $H_a$ ,

$$Z - \frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}} \sim N(0, 1) \rightarrow Z \sim N \left( \frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}}, 1 \right)$$

and so

$$\begin{aligned}
\text{power} &= P(Z > Z_{1-\alpha/2}) \\
&= P\left(Z - \frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}} > Z_{1-\alpha/2} - \frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}}\right) \\
&= 1 - \Phi\left(Z_{1-\alpha/2} - \frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}}\right) \\
&= \Phi\left(\frac{\theta_a}{\sqrt{\text{Var}(\hat{\theta})}} - Z_{1-\alpha/2}\right)
\end{aligned}$$

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

### A.6 Introduction

- In cluster randomized trials (CRTs), randomization of interventions are done at the level of groups rather than individuals. This is most useful when intervention can only be administered at the community-level, when contamination may be an issue, or for other reasons. The individual units within a cluster are correlated.
- In parallel CRTs, clusters are all assigned an intervention at a single time point. If cluster sizes are equal, t-tests/ANOVA may be used to compare cluster-level mean responses. Clusters may also be matched for a paired setup. When cluster sizes vary, individual level analyses such as generalized estimating equations (GEE) or random effects models may be used.
- In crossover CRTs, each interventions is applied to a cluster at different time points (with a possible “washout” period in between). The order of the interventions is randomized for each cluster. Crossover designs are less commonly used in CRTs due to extending the period of the study. Crossover CRTs usually use paired t-tests to make within-cluster comparisons.
- In a stepped wedge CRT, clusters cross from one intervention to another at different time points (typically all starting from the control intervention and all ending with the treatment intervention). The time at which each cluster crosses over is randomized. The stepped wedge CRT is useful when there are limited resources for applying the intervention. The key features of the stepped wedge CRT is that the crossover is unidirectional and that the intervention is never removed once implemented. The unidirectionality does complicate the analysis as the treatment effect cannot be estimated from only within-cluster comparisons.

### A.7 Model

- For a design with  $I$  clusters,  $T$  time points,  $N$  individuals sampled per cluster per time interval, define the cluster means

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

where

- $\alpha_i \sim N(0, \tau^2)$  is a random effect for cluster  $i \in \{1, \dots, I\}$
- $\beta_j$  is a fixed effect for time interval  $j \in \{1, \dots, T-1\}$  (with  $\beta_T = 0$  for identifiability)
- $X_{ij}$  is the intervention indicator in cluster  $i$  at time  $j$  (with 1 denoting intervention)
- $\theta$  is the treatment effect.

- Individual level responses are modelled as

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

where  $\epsilon_{ijk} \stackrel{iid}{\sim} N(0, \sigma^2)$ . Cluster means are modelled as

$$\bar{Y}_{ij.} = \mu_{ij} + \bar{\epsilon}_{ij.}$$

where  $\bar{\epsilon}_{ij.} = \frac{\sum_k \epsilon_{ijk}}{N}$ . Assume that  $\epsilon_{ijk}$  are independent of  $\alpha_i$ .

- The variance of the individual-level response is

$$\text{Var}(Y_{ijk}) = \tau^2 + \sigma^2$$

The variance of the cluster-level response is

$$\text{Var}(\bar{Y}_{ij.}) = \tau^2 + \frac{\sigma^2}{N} = \left( \frac{\tau^2 + \sigma^2}{N} \right) (1 + (N-1)\rho)$$

where  $(1 + (N-1)\rho)$  is the “variance inflation factor” and  $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$  is the intraclass correlation. Note:

$$\begin{aligned} \left( \frac{\tau^2 + \sigma^2}{N} \right) (1 + (N-1)\rho) &= \left( \frac{\tau^2 + \sigma^2}{N} \right) \left( 1 + \frac{(N-1)\tau^2}{\tau^2 + \sigma^2} \right) \\ &= \frac{\tau^2 + \sigma^2}{N} + \frac{N\tau^2(\tau^2 + \sigma^2)}{N(\tau^2 + \sigma^2)} - \frac{(\tau^2 + \sigma^2)\tau^2}{N(\tau^2 + \sigma^2)} \\ &= \frac{\tau^2 + \sigma^2}{N} + \tau^2 - \frac{\tau^2}{N} \\ &= \tau^2 + \frac{\sigma^2}{N} \end{aligned}$$

- Some characterize the cluster effect on the variance using the coefficient of variation (CV)  $\frac{\tau}{\mu}$ .
- If the individual level responses are binary, then the cluster-level response  $\bar{Y}_{ij.}$  is a proportion and it is assumed  $\sigma^2 = \mu(1 - \mu)$ .
- For varying-sized clusters, replace  $N$  with  $N_{ij.}$ .

## A.8 Data analysis

### A.8.1 $\tau^2$ and $\sigma^2$ known

If the variances  $\tau^2$  and  $\sigma^2$  are known, then estimates of the fixed effects can be estimated using weighted least squares (WLS) at the cluster-level. Let  $\mathbf{X}$  be the  $IT \times (T+1)$  design matrix of cluster-time means corresponding to parameter vector  $\eta = (\mu, \beta_1, \dots, \beta_{T-1}, \theta)$ . Let  $\mathbf{W}$  be a  $IT \times IT$  block diagonal matrix where each  $T \times T$  block of  $\mathbf{W}$  describes the correlation structure between the repeated cluster means over time and has the structure

$$\begin{bmatrix} \tau^2 + \frac{\sigma^2}{N} & \tau^2 & \dots & \tau^2 \\ \tau^2 & \ddots & & \vdots \\ \tau^2 & & \ddots & \tau^2 \\ \tau^2 & \dots & \tau^2 & \tau^2 + \frac{\sigma^2}{N} \end{bmatrix}$$

Then the fixed effect estimates are given by

$$\hat{\eta} = (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{-1} \mathbf{y}$$

**A.8.2  $\tau^2$  and  $\sigma^2$  unknown**

- When the response is continuous and normally distributed, an empirical Bayes approach at the cluster-level is possible to estimate the fixed effects in the LMM. This approach also works for non-normal individual-level data when the cluster sizes are approximately equal.
- If the responses are non-normal and the cluster sizes vary, then analysis at the individual-level using GLMM or GEE is preferred.

**A.9 Effect of number of steps**

Optimal power is achieved when a single cluster crosses over to the intervention at each time. The loss of power is primarily due to the loss of measurement times rather than due to the loss of randomization times. The loss of power is also relatively independent of the CV.

**A.10 Delayed treatment effect**

If effect of intervention only reaches full effect  $\theta$  after some time, then power is reduced. The delay may be modelled by allowing  $X_{ij}$  to be fractional (although the given  $\text{Var}(\hat{\theta})$  is not valid in this case). Power is increased by adding additional measurement periods at the end of the trial or by increasing the time intervals.

**A.11 Simulation results**

Compared power of test for LMM, GEE and GLMM for varying levels of relative risk. In equal cluster size case,  $\text{LMM} > \text{GEE} > \text{GLMM}$ . In unequal cluster size case,  $\text{GEE} \approx \text{GLMM} > \text{LMM}$ . When cluster sizes vary significantly, it is suggested to do individual-level analyses. A jackknife estimate of the variance is suggested to maintain the size of the test in GEE/GLMM analyses.