11/18/05

Cluster Randomized Crossover Trials: Design and Analysis of the Stepped Wedge Design

(Running title: Stepped Wedge Designs)

Michael A. Hussey[1], James P. Hughes[1]

[1]Department of Biostatistics, University of Washington, Seattle, WA

Corresponding author: James P. Hughes, Department of Biostatistics 357232, University of Washington, Seattle, WA 98195 email: jphughes@u.washington.edu

Acknowlegements: This research was supported by NIH grants AI29168, AI46702.

Abstract

Cluster randomized trials (CRT) are often used to evaluate therapies or interventions in situations where individual randomization is not possible or not desirable for logistic, financial or ethical reasons. While a significant and rapidly growing body of literature exists on CRTs utilizing a "parallel" design (i.e. I clusters randomized to each treatment), only a few examples of CRTs using crossover designs have been described. In addition, important statistical aspects of such designs have not been developed. In this article we discuss the design and analysis of a particular type of crossover CRT - the stepped wedge - and provide an example of its use.

KEY WORDS: Cluster randomized trial; Stepped wedge design; Prevention trials

1 Introduction

Cluster (or community, or group) randomized trials (CRT) are distinguished by the fact that individuals are randomized in groups rather than individually. CRTs have been used to evaluate antismoking interventions ([1],[2]), methods of controlling HIV and STDs ([3], [4]), and in a number of other contexts ([5],[6]). Cluster designs may be chosen because the intervention can only be administered on a community-wide scale (e.g. [7]), or to minimize contamination ([8]), or for other logistic, financial or ethical reasons. From a statistical viewpoint, the key characteristic of CRTs is that the individual units within a cluster are correlated and this feature must be incorporated into power calculations and the trial analysis.

CRTs often employ a parallel design: for a two-arm study with $2I$ independent clusters, I clusters are randomly assigned to each intervention at a single time point. A two-sample t-test may be used to compare cluster-level mean responses between the intervention groups. If there are more than 2 treatment arms, a one-way analysis of variance may be used. Sometimes the communities are matched and randomization is done within the matched sets. In that case, a paired analysis (e.g. paired t-test) is used. Statistical aspects of the design and analysis of parallel CRTs have been widely discussed (e.g. [9],[10]).

In contrast, crossover designs are less commonly used in CRTs (two examples are [11], [12]). A crossover CRT requires fewer clusters than a parallel design but may take twice as long (or longer) to complete (since each cluster receives both the treatment and control interventions). If the intervention requires a lengthy followup period, then this fact alone might make a crossover design impractical. In a standard crossover design the order of the interventions is randomized for each cluster and a time period (called the "washout" period) is often allowed between the two interventions so that the first intervention does not affect the second. Analysis of a standard crossover design focuses on within-cluster comparisons using a paired t-test.

A stepped wedge design ([13]) is a type of crossover design in which different clusters cross over (switch treatments) at different time points. In addition, the clusters cross over in one direction only - typically, from control to intervention. The first time point usually corresponds to a baseline measurement where none of the clusters receive the intervention of interest. At subsequent time points, clusters cross over to the intervention of interest and the response to the intervention is measured. More than one cluster may begin the intervention at a time point, but the time at which a cluster begins the intervention is randomized. Figure 1 illustrates the differences between the parallel, traditional crossover and stepped wedge 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: Treatment Schedules for Parallel, Crossover, and Stepped Wedge designs. A "0" represents control or existing treatment; a "1" represents an intervention.

Although the stepped wedge design extends the length of a randomized trial due to the presence of multiple time intervals, the nature of the design may be beneficial in certain settings. In a parallel or traditional crossover design, the intervention must be implemented in half of the total clusters simultaneously. However, limited resources or geographical constraints may make this logistically impossible (e.g. [13]). The stepped wedge design allows the researcher to implement the intervention in a smaller fraction of the clusters at each time point. Another unique feature of the stepped wedge design is that the crossover is unidirec-

tional. All clusters eventually receive the intervention and, in particular, the intervention is never removed once it has been implemented (at least over the course of the trial) which may alleviate ethical and/or community concerns. This makes the stepped wedge design particularly useful for evaluating the population-level impact of an intervention that has been shown to be effective in an individually randomized trial. The unidirectional aspect of the crossover does, however, complicate the analysis since the treatment effect can no longer be estimated exclusively from within-cluster comparisons. More details on the analysis of such trials is provided below.

In section 2 we describe a trial being conducted in Washington state that uses a stepped wedge design. In section 3 we describe statistical aspects of the design and analysis of stepped wedge CRTs and in section 4 we summarize our findings.

2 Example - Partner Notification

Partner notification is the process by which sex partners of patients with sexually transmitted infections (STIs) are notified of potential exposure to infection and encouraged to seek treatment. Standard practice for partner notification in most states in the U.S. involves contact of partners by public health authorities. However, the high costs associated with this practice has influenced investigators to seek alternative partner treatment methods. One alternative strategy is patient delivered partner therapy (PDPT) in which infected persons are given drugs or drug vouchers to give to their sex partners. In the case of vouchers, these can be redeemed for appropriate drugs at local pharmacies.

An individually randomized trial conducted by Golden *et al.* ([14]) in King County, Washington between 1998 and 2003 evaluated the effectiveness of a PDPT-based partner notification strategy dubbed EPT (expedited partner therapy) versus standard partner notification for the treatment of chlamydia or gonnorrhea infection. The primary outcome was the presence of persistent or recurrent infection in the original index patient 3 to 19 weeks

after treatment. Overall, the trial showed a significantly increased proportion of partners treated (per participant report) and a decreased risk of recurrent or persistent infection among participants in the EPT group compared to the control.

Based on the success of this individually randomized trial, the county health commissioners of Washington state have agreed to implement EPT in all the counties in Washington. Support for a stepped wedge cluster randomized trial to evaluate the population-level effect of the intervention has been received from NIH. Twenty four health districts in Washington state will be randomized to EPT at one of four possible times. The randomization times are separated by a period of 6 months to allow implementation and assessment of the intervention within each time period. The primary outcomes are the prevalence of chlamydial infection among women tested in family planning clinics and the number of reported gonorrhea infections in women in each county.

Preliminary data suggest that overall baseline prevalence of chlamydial infection will be 0.05 and the coefficient of variation (CV) for county to county variation ([15]) is 0.30. Gonorrhea infection is much rarer and incidence rates in the 10 - 44 year old female population average 79 per 100,000 person years. However, there is substantial variation from county to county and the estimated CV is 0.90.

3 Statistical Issues

In this section we examine a number of issues related to the design and analysis of stepped wedge CRTs.

3.1 Model

Random effects are commonly used to model the correlation between individuals within the same cluster in CRT's. For a design with I clusters, T time intervals, and N individuals per

cluster, let Y_{ijk} be the response corresponding to individual k at time j from cluster i (i in $1 \ldots I$, j in $1 \ldots T$, k in $1 \ldots N$) and let $Y_{ij.}$ be the mean for cluster i at time j. Define

$$\mu_{ij} = \mu + \alpha_i + \beta_j + X_{ij}\theta \tag{1}$$

where α_i is a random effect for cluster i such that $\alpha_i \sim N(0, \tau^2)$, β_j is a fixed effect corresponding to time interval j (j in $1 \ldots T-1$, $\beta_T = 0$ for identifiability), X_{ij} is an indicator of the treatment mode in cluster i at time j (1 = intervention; 0 = control), and θ is the treatment effect.

Individual level responses may be modelled as

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

where $e_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$ (individual level covariates may be added to this model by defining μ_{ijk} in an analogous manner). A model for the cluster means is obtained by summing over the individuals in a cluster to obtain:

$$Y_{ij.} = \mu_{ij} + e_{ij} \tag{2}$$

where $e_{ij} = \sum_k e_{ijk}/N \overset{\text{iid}}{\sim} N(0, \sigma^2)$ and $\sigma^2 = \sigma_e^2/N$. We also assume that the e_{ijk} (and, hence, e_{ij}) are independent of the α_i.

The variance of an individual-level response is

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

and the variance of the cluster-level response is

$$Var(Y_{ij.}) = \tau^2 + \sigma^2 = \frac{\tau^2 + \sigma_e^2}{N}[1 + (N-1)\rho]$$

where $\rho = \tau^2/(\tau^2 + \sigma_e^2)$ is referred to as the intraclass correlation and characterizes the correlation between individuals from the same cluster. The increase in the variance of $Y_{ij.}$ due to the clustering (relative to independent data) is given by the "variance inflation factor"

$1 + (N-1)\rho$. Alternatively, some authors characterize the cluster effect in terms of the coefficient of variation, τ/μ.

If the individual level responses are binary then the cluster level response $Y_{ij.}$ is a proportion and it is reasonable to assume that $\sigma_e^2 = \mu_{ij} * (1 - \mu_{ij})$. The model (2) is easily adapted to handle different numbers of individuals per cluster by substituting N_{ij} for N.

3.2 Approaches to Data Analysis

In the following we discuss approaches to analysis of data from a study employing the stepped wedge design. We focus on analysis of the cluster-level means as these are typically the primary units of analysis in a CRT.

3.2.1 τ^2 and σ^2 known

Model (2) is an example of a linear mixed model (LMM). If the values of the variance components τ^2 and σ^2 are known, then estimates of the fixed effects can be obtained using weighted least squares (WLS). Specifically, let $\mathbf{Z}$ be the $IT \times (T+1)$ design matrix corresponding to the parameter vector $\eta = (\mu, \beta_1, \beta_2, \ldots, \beta_{T-1}, \theta)$ for a stepped wedge design. Then $\hat{\eta} = (\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Y})$ and the covariance matrix of $\hat{\eta}$ is $(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}$, where $\mathbf{V}$ is an $IT \times IT$ block diagonal matrix. Each $T \times T$ block within $\mathbf{V}$ describes the correlation structure between the repeated (in time) cluster means and has the structure

$$\begin{bmatrix} \sigma^2+\tau^2 & \tau^2 & \ldots & \tau^2 \\ \tau^2 & \ddots & & \vdots \\ \vdots & & \ddots & \tau^2 \\ \tau^2 & \ldots & \tau^2 & \sigma^2+\tau^2 \end{bmatrix}.$$

Since τ^2 and σ^2 are seldom known this approach is generally not applicable for data analysis, but provides a useful approach to power analyses.

3.2.2 τ^2 and σ^2 unknown

When the variance components are unknown, Laird and Ware ([16]) describe an empirical Bayes approach to estimating the fixed effect parameters and variance components of LMM when the response is continuous and normally distributed. Variants (e.g. GLMM, for generalized LMM ([17])) have been developed to handle binary and other non-normally distributed endpoints but implementations are less common. LMM and GLMM approaches provide estimates of the variance components based on the assumed model structure. However, use of the wrong model structure can lead to invalid inference.

Alternatively, generalized estimating equations (GEE) ([18]), which can flexibly handle normal or non-normal endpoints, are sometimes used to analyze CRT data. GEE tends to be more robust to misspecification of the variance structure than LMM or GLMM since "sandwich" type variance estimates are used. GEE is more natural than LMM for individual-level binary outcomes since a logit link can be used to analyze the individual-level data. Also, an individual-level GEE analysis automatically accounts for varying cluster sizes when necessary. In contrast, LMM analyses are typically done at the cluster mean level and so weights must be assigned to the cluster means if the cluster sizes vary.

Both LMM and GEE should be used with care if the number of clusters and time points is small since theoretical results for these methods are based on asymptotics. Feng *et al.* ([19]) contrast these two approaches for parallel design CRTs. Section 3.7 uses simulations to compare these two approaches in the context of the stepped wedge design.

3.2.3 Within-Cluster Analysis

The methods discussed above use both within-cluster and between-cluster information to estimate the treatment effect. This approach is necessary to avoid confounding the treatment effect with changes over time. However, if there are no temporal effects on the outcome (i.e. $\beta_j = 0$ for all j in model (1)), then a within-cluster analysis can be used to estimate the

treatment effect. This type of analysis was used in the Gambia Hepatitis trial ([13]).

Consider a design with I clusters and T time points. Let w_i be the number of time points in cluster i that receive the control. Consequently, $T - w_i$ is the number of time points in cluster i that receive the intervention. Furthermore, let $\mathcal{C}_i$ and $\mathcal{T}_i$ be the sets of time points receiving control and intervention in cluster i, respectively. Then, a within-cluster estimate of θ is given by

$$\tilde{\theta} = \frac{1}{I}\sum_i [\frac{\sum_{j\in\mathcal{T}_i} Y_{ij\cdot}}{T - w_i} - \frac{\sum_{j\in\mathcal{C}_i} Y_{ij\cdot}}{w_i}] \tag{3}$$

and under model (2) (assuming all $\beta_j = 0$), the variance is given by

$$Var(\tilde{\theta}) = \frac{\sigma^2}{I^2}\sum_i (\frac{1}{w_i} + \frac{1}{T - w_i}) \tag{4}$$

Notice that this variance formula does not contain τ^2 since the between-cluster variance is eliminated in the paired analysis.

The drawback of a within-cluster analysis is the potential for bias. If the time effects, $\beta_1, \ldots, \beta_T$ are not all 0, then the estimated treatment effect (3) will, in general, be biased. The bias of the treatment effect estimate when the analysis ignores time effects is a linear combination of $\beta_1 \ldots \beta_{T-1}$ ([20]):

$$b(\tilde{\theta}, \theta) = \sum_{k=1}^{T-1} \beta_k \left(\frac{I(\sigma^2 + T\tau^2)(T\sum_i X_{ik} - U)}{IT[(\sigma^2 + T\tau^2)U - V\tau^2] - U^2\sigma^2} \right) \tag{5}$$

where $U = \sum_{ij} X_{ij}$, $V = \sum_i (\sum_j X_{ij})^2$, and $X_{ij} = 1$ if cluster i receives the intervention at time j and 0 otherwise. Inspection of equation (5) shows that the coefficent for β_k will be zero only when $U = T(\sum_i X_{ik})$. This occurs when the number of clusters randomized to treatment at time k is equal to the average number of clusters randomized to treatment over all times. Although this may be true for a single time interval in the stepped wedge design, it will not be true for all time points. Thus, failure to model time effects during analysis will bias the treatment effect if time effects exist. Note, however, that the bias in $\tilde{\theta}$ is independent of the true value of θ. Furthermore, the coefficients of the β's can be calculated once the

treatment schedule is determined. Thus, understanding of each β's contribution to the bias can occur during the design phase of the trial.

3.3 Power calculations

Suppose the goal is to test the hypothesis $H_o : \theta = 0$ versus $H_a : \theta = \theta_A$ in model (2) using a stepped wedge design with I sites and T time points. A Wald test may be based on $Z = \frac{\hat{\theta}}{\sqrt{Var(\hat{\theta})}}$. The approximate power for conducting a two-tailed test of size α is given as

$$power = \Phi\left(\frac{\theta_A}{\sqrt{Var(\hat{\theta})}} - Z_{1-\alpha/2}\right) \tag{6}$$

where Φ is the cumulative standard Normal distribution function and $Z_{1-\alpha/2}$ is the $(1-\alpha/2)^{th}$ quantile of the standard Normal distribution function. In general, $Var(\hat{\theta})$ is the appropriate element of $(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}$ from a weighted least squares analysis (section 3.2.1). However, for models of the form (2) (which includes parallel and crossover as well as stepped wedge designs) and assuming X_{ij} is coded 0 or 1, it is possible to express $Var(\hat{\theta})$ in closed form. As before let $X_{ij} = 0$ if cluster i receives the control at time j and $X_{ij} = 1$ if cluster i receives the intervention at time j. Assuming equal N per cluster it can be shown that

$$Var(\hat{\theta}) = \frac{I\sigma^2(\sigma^2 + T\tau^2)}{(IU - W)\sigma^2 + (U^2 + ITU - TW - IV)\tau^2} \tag{7}$$

where $U = \sum_{ij} X_{ij}$, $W = \sum_j (\sum_i X_{ij})^2$, and $V = \sum_i (\sum_j X_{ij})^2$ [21].

In the Washington EPT trial, the baseline prevalence of Chlamydia is approximately 0.05 and we plan to test 100 individuals per cluster per time period. For the power calculations, therefore, we use $\sigma^2 = \frac{(.05)(.95)}{100} = 0.000475$. The 24 counties will be randomized 6 at a time, so that $T = 5$. Figure 2 shows the power of the trial as a function of effect size (expressed as a relative risk) for a coefficient of variation of 0.3 and 0.5. Because the stepped wedge design uses both within-cluster and between-cluster information, power is relatively insensitive to variations in the CV. For a CV of 0.3 the plot shows that the trial has about 80% power to detect a decrease in prevalence of roughly 36 percent (from 0.05 to 0.032).

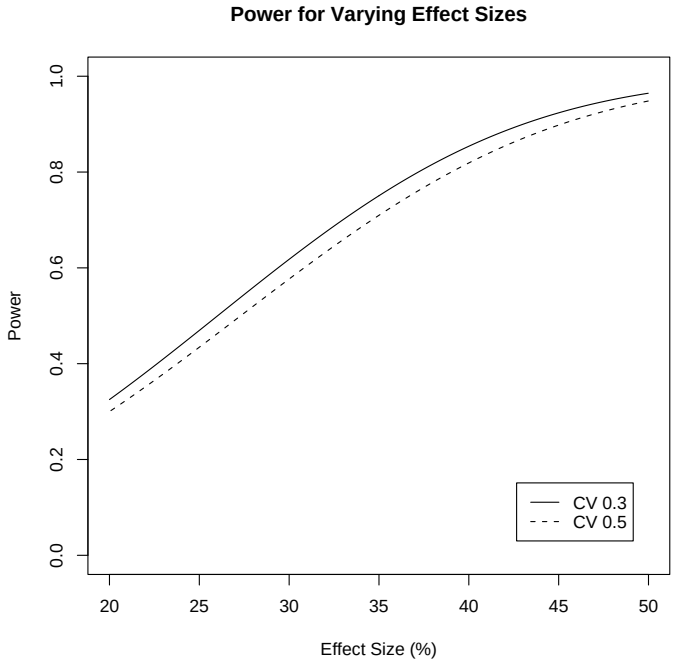

Figure 2: Theoretical power for the Washington EPT trial. The overall prevalence is assumed to be 5 percent, with 100 individuals sampled per cluster per time point. Power is displayed versus effect size for two coefficients of variation.

3.4 Effect of Number of Steps

An important choice in the stepped wedge design is the number of clusters randomized at each time step. Figure 3 illustrates the effect of i) varying the number of clusters randomized at each time step (but holding the total number of time steps (=measurement times) constant) and ii) varying the number of time steps, for the Washington State EPT trial assuming a relative risk of 0.7 (other alternatives give similar results).

The optimal power is achieved when each cluster is randomized to the intervention at its own randomization step. However, this may be infeasible for logistic reasons, especially if the design calls for the steps to be separated by a period of months. From figure (3a) we see that relatively little power is lost by randomizing multiple clusters at some time steps and zero at others provided the total number of measurement times is held constant. Again, however, this may make the trial unacceptably long. In figure (3b) we illustrate the effect of randomizing multiple clusters at each time point and reducing the overall number of measurement points. In this case, power is significantly adversely effected. Note that the lines stay approximately "parallel" across a wide range of the CV's indicating that the loss in power is relatively independent of the coefficient of variation.

3.5 Efficacy of WLS relative to a within-cluster analysis

The relative efficiency of the WLS estimator, $\hat{\theta}$, versus the within-cluster estimate, $\tilde{\theta}$, can be determined by taking the (inverse of the) ratio of the respective variances. This ratio is

$$\text{effic}(\hat{\theta}, \tilde{\theta}) = \frac{\sum_i (\frac{1}{w_i} + \frac{1}{T-w_i})[(ITU - U^2)\sigma^2 + IT(TU - V)\tau^2]}{I^3(\sigma^2 + T\tau^2)} \tag{8}$$

(note: the WLS variance here is different from eq (7) since this comparison is developed under the assumption that there are no time effects). It can be shown that the WLS estimator always exceeds the t-test in efficiency unless $\tau^2 = 0$ ([20]). However, if time effects are included in the WLS model (so that the variance (7) is appropriate) then $\hat{\theta}$ is less efficient

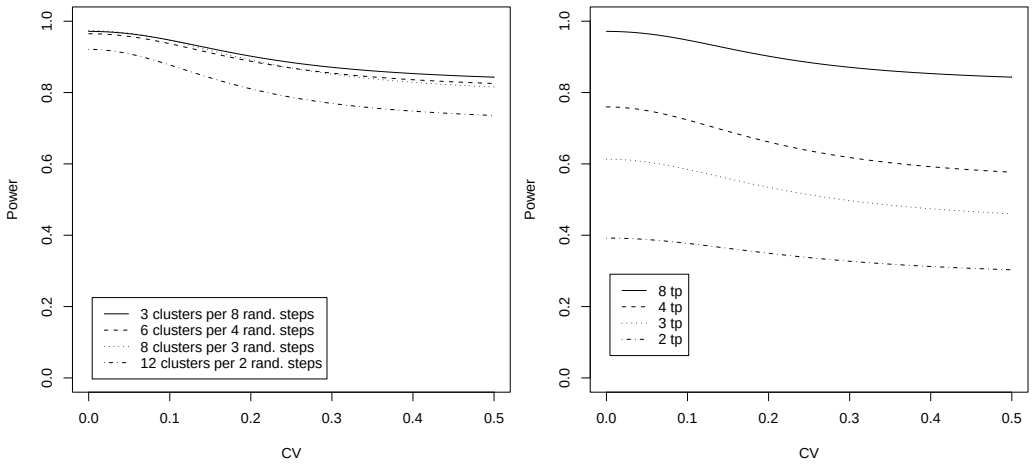

Figure 3: Power curves when 24 clusters are randomized and number of randomization steps is varied. In the left plot all designs include 9 measurement times and the number of randomization times is varied; in the right plot the number of measurement times (tp) equals the number of randomization times. In both plots the baseline event prevalence is 0.05 and the intervention effect corresponds to an risk ratio of 0.7

than $\tilde{\theta}$ but, as described in section 3.2.3, $\tilde{\theta}$ may be biased.

3.6 Delayed treatment effect

The results presented in the previous sections assume that the full effect of the intervention is realized in the same time interval that the intervention is introduced. In some situations, however, the full effect of the intervention may not be realized until several time intervals following implementation. This section explores changes in power due to such a delay.

Suppose we expect that the intervention will be 50% effective after one time interval, 80% effective after two time intervals and 100% effective after three time intervals. We may continue to parameterize the treatment effect in terms of a single parameter, θ, which can be interpreted as the maximum or full treatment effect. The delay may be modelled by allowing X_{ij} in equation (1) to be fractional. Power may then be calculated as outlined in section 3.3 although the closed form expression (7) is not valid when the X_{ij} are fractional.

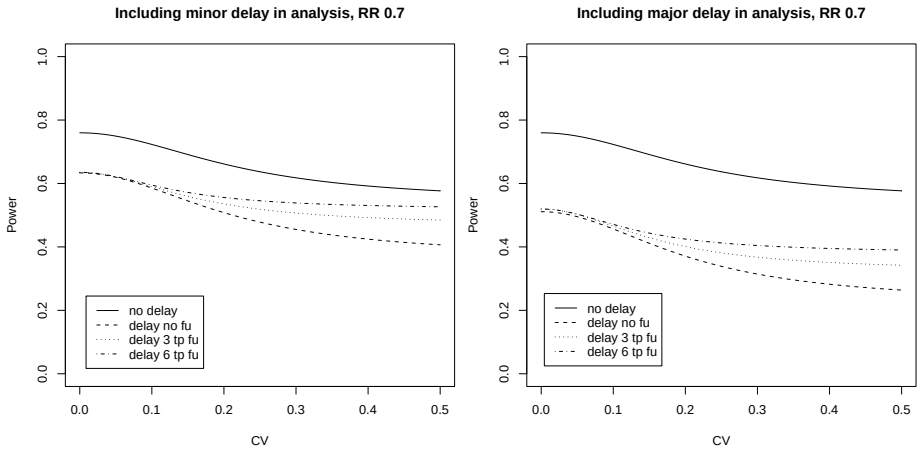

Figure 4: Theoretical power vs. CV comparing situations in the Washington EPT trial where a minor treatment effect delay is assumed and when a major delay is assumed. Figures are shown for a risk ratio of 0.7. Plots have lines corresponding to situations where no delay, delay and no additional monitoring, delay and 3 additional monitoring time points, and delay and 6 additional monitoring time points exist.

The overall effect of such a delay is to reduce power. Power can be partly, but not completely, recovered by adding additional measurement periods onto the end of the trial. The greater the delay in the intervention effect, the greater is the effect on power. Figure 4 shows the effect of a minor delay (80%, 90%, and 100% at 1, 2, and 3 time units post-intervention, respectively) and major delay (50%, 80%, and 100% at 1, 2, and 3 time units post-intervention, respectively) on power in the Washington state EPT trial as well as the potential for recovery of power through the addition of extra measurement periods. Although inclusion of additional monitoring periods at the end of the study increases power, it is difficult to recover full power. It is important, therefore, to make the time intervals sufficiently long so that the full intervention effect is realized in a single interval.

3.7 LMM vs GEE

We did a small simulation experiment to compare the size and power of LMM and GEE in the context of the stepped wedge design. We evaluated two situations: where equal samples sizes are available for each cluster and where variable samples sizes are available for each cluster. These two situations correspond to the sampling plans for comparing chlamydial and gonorrheal rates (respectively) in the Washington state EPT trial described in section 2. A trial with 24 clusters and 4 randomization steps was considered. The baseline prevalence of disease was 0.05 and the between-cluster variance τ^2 was assumed to be 0.000225, which corresponds to coefficient of variation of 0.3. We used 100 individuals per cluster per time interval for the simulations with equal sample sizes per cluster. For the simulations with different cluster sizes we randomly assigned the total 2400 individuals in each time interval to 24 clusters using a multinomial distribution with a flat prior Dirichlet distribution (parameters (1,1,1)). Using this distribution, the interquartile range for the number of individuals per cluster was (28, 136).

Table 1: Estimated power comparing clusters that have the same sample size ($N = 100$) and clusters with different sample sizes (24 clusters, 5 time points, $\tau^2 = 0.000225$, $\mu = 0.05$, 500 iterations)

	Same cluster sizes		Different cluster sizes	
Odds Ratio	LMM	GEE	LMM	GEE
1.0	0.054	0.056	0.040	0.044
0.7	0.688	0.706	0.298	0.694
0.6	0.912	0.914	0.510	0.896
0.5	0.976	0.976	0.704	0.984

The estimated power based on the simulations is given in table (1). We see little difference between the two approaches when the cluster sizes are equal. However, power was much

better for GEE when cluster sizes varied. This is likely due to the ability of GEE to analyze (binary) data at the individual level and thereby provide the correct weighting for each cluster. Since the LMM approach (implemented using the R function lme(), developed by Pinheiro and Bates ([22])), is based on a normal model, the data were analyzed at the cluster level and weights were imposed to account for the different cluster sizes. However, the correct weights depend on the unknown variance components. Since these are unknown prior to an analysis we tried using weights proportional to the cluster size (results shown in the table) and equal weights (not shown but results similar to those given in the table). Both approaches are inefficient relative to a correctly weighted analysis and this is manifest as low power in the table. Due to this difficulty, we recommend using individual level analyses when cluster sizes vary significantly and GEE if the individual observations are binary.

4 Discussion

Using theoretical calculations and simulation we have investigated statistical characteristics of the stepped wedge design for cluster randomized trials. When there is no delay in treatment effect and the samples from each cluster are assumed to be of equal size, several important results were obtained. Given the treatment schedule and estimates of the variance components, a closed form for the variance of the treatment effect estimate was derived. This formula can be used to calculate theoretical power during the design stage.

We found that, for a fixed number of clusters, power decreases as the number of steps decreases. Most of the power loss is due to a reduction in the number of measurements rather than the reduction in randomization steps. However, in practice, the optimal situation of having one cluster randomized to the intervention at each time point may be infeasible. A practical strategy is simply to maximize the number of time intervals given constraints on the number of clusters that can logistically be started at one time point and the desired length of the trial.

A paired t-test provides a valid analysis of the stepped wedge design only if there are no time effects. Otherwise, a paired analysis provides a biased estimate of the treatment effect. A formula for the bias was derived based on the treatment schedule, hypothesized true values of time effect parameters $\beta_1 \ldots \beta_{T-1}$, and a hypothesized true value of the between-cluster variation τ^2. However, if external or apriori information suggests that there are no time effects then an analysis based on model (2) without parameters for time still provides a more efficient analysis than the paired t-test.

We found that a delay in the treatment effect (i.e. where the full treatment effect is not realized until one or more time intervals after the intervention is introduced) significantly reduces power. Delays can be incorporated into the power calculations by using fractional values for the treatment covariate in the design matrix $\mathbf{Z}$. Explicit modeling of the delay in this manner recovers a small portion of the power. Adding additional monitoring periods at the end of the trial results in additional power recovery. However, the loss in power due to a delay in the treatment effect generally cannot be fully recovered. Therefore, it is desirable to make each monitoring period long enough so that the effect of the treatment is fully realized before the next period begins.

To mimic the Washington state EPT trial, we used simulations with 24 clusters and 5 time intervals to compare GEE and LMM. In this case, the simulation results agreed well with predictions based on asymptotics - both GEE and LMM maintained the nominal test size and had similar power for the case of equal cluster sizes. However, GEE was preferable when cluster sizes varied. As an alternative, randomization based procedures could be used to evaluate test results for smaller studies in which asymptotics may be suspect.

Model (2) assumes that there are no cluster by time interactions. Including such interactions would result in an overparameterized model, however. If a cluster by time interaction is expected then one possible strategy is to create strata of clusters with similar expected time trends. Then a strata by time interaction could be included as a factor in the model.

The stepped wedge design is an innovative choice for a cluster randomized crossover trial

that is subject to constraints that limit the use more conventional designs. The stepped wedge seems particularly suited to investigations of community level public health interventions and so-called "phase IV" effectiveness trials.

References

[1] Gail MH, Byar DP, and Pechacek TF *et al.* Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT). *Controlled Clinical Trials*, 13:6–21, 1992.

[2] Peterson AV Jr, Kealey KA, and Mann SL *et al.* Hutchinson Smoking Prevention Project: long-term randomized trial in school-based tobacco use prevention- results on smoking. *Journal of the National Cancer Institute*, 92:1979–91, 2000.

[3] Grosskurth H, Mosha F, and Todd J *et al.* Impact of improved treatment of sexually transmitted diseases on hiv infection in rural tanzania: randomised controlled trial. *The Lancet*, 346:530–6, 1995.

[4] Wawer MJ, Sewankambo NK, and Serwadda D *et al.* Control of sexually transmitted diseases for AIDS prevention in Uganda: a randomised community trial. *The Lancet*, 353:525–35, 1999.

[5] Martiniuk A, O'Connor K, and King W. A cluster randomized trial of a sex education programme in Belize, Central America. *International Journal of Epidemiology*, 32:131–136, 2003.

[6] Sjögren T, Nissinen K, and Järvenpää K *et al.* Effects of a workplace physical exercise intervention on the intensity of headache and neck and shoulder symptoms and upper extremity muscular strength of office workers: A cluster randomized controlled cross-over trial. *Journal of the International Association for the Study of Pain*, 116:119–128, 2005.

[7] Gail MH. On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, 15:1069–1092, 1996.

[8] Torgerson DJ. Contamination in trials: is cluster randomisation the answer? *BMJ*, 322:355–7, 2001.

[9] Donner A and Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research.* Arnold Publishers, 2000.

[10] Murray D. *Design and Analysis of Group-Randomized Trials.* Oxford University Press, 1998.

[11] Palmer RH, Louis TA, and HSU LN *et al.* A randomized controlled trial of quality assurance in sixteen ambulatory care practices. *Medical Care*, 23:751–70, 1985.

[12] Menzies R, Tamblyn R, and Farant JP *et al.* The effect of varying levels of outdoor air supply on the symptoms of sick building syndrome. *New England Journal of Medicine*, 328:821–7, 1993.

[13] Gambia Hepatitis Study Group. The Gambia Hepatitis Intervention Study. *Cancer Research*, 47:5782–87, 1987.

[14] Golden M, Whittington W, and Handsfield H *et al.* Effect of expedited treatment of sex partners on recurrent or persistent gonorrhea or chlamydial infection. *New England Journal of Medicine*, 352(7):676–685, 2005.

[15] Hayes R and Bennett S. Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology*, 28(2):319–326, 1999.

[16] Laird N and Ware J. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

[17] Breslow NE and Clayton DG. Approximate inference in generalized linear mixed models. *J. American Statistical Assoc.*, 88:9–25, 1993.

[18] Liang K and Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

[19] Feng Z, Diehr P, and Peterson A *et al.* Selected statistical issues in group randomized trials. *Annual Review of Public Health*, 22:167–187, 2001.

[20] Hussey M. Cluster randomized crossover trials: Aspects of power, variance, and bias in the stepped wedge design. Master's thesis, University of Washington, 2005.

[21] Hughes JP, Goldenberg RL, and Wilfert CM *et al.* Design of the HIV prevention trials network (HPTN) protocol 054: A cluster randomized crossover trial to evaluate combined access to nevirapine in developing countries. Technical Report 195, University of Washington, Department of Biostatistics, 2003.

[22] Pinheiro J and Bates D. *Mixed-effects models in S and S-PLUS*. Springer Publishing, 2000.