Available online at www.sciencedirect.com
ScienceDirect
Contemporary Clinical Trials 28 (2007) 182–191
Contemporary Clinical Trials
www.elsevier.com/locate/conclintrial

# Design and analysis of stepped wedge cluster randomized trials

Michael A. Hussey [a], James P. Hughes [b,*]

[a] *Fred Hutchinson Cancer Research Center, Seattle, WA, United States*
[b] *Department of Biostatistics 357232, University of Washington, Seattle, WA 98195, United States*

Received 29 November 2005; accepted 25 May 2006

**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 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.


*Keywords:* 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 preventing human immunodeficiency virus (HIV) and other sexually transmitted diseases (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. If the cluster sizes are all equal, 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. When cluster sizes vary, individual level analyses using generalized estimating equations [17] or random effects models [16] may be 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 (three examples are [6,11,12]). A crossover CRT requires fewer clusters than a parallel design but may take twice as long (or longer) to complete (since each cluster

* Corresponding author.
*E-mail address:* jphughes@u.washington.edu (J.P. Hughes).

1551-7144/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2006.05.007

Parallel

| | | Time 1 |
|---|---|---|
| Cluster | 1 | 1 |
| | 2 | 1 |
| | 3 | 0 |
| | 4 | 0 |

Crossover

| | | Time 1 | Time 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 |

Fig. 1. Treatment schedules for parallel, crossover, and stepped wedge designs. "0" represents control or existing treatment; "1" represents an intervention.

receives both the treatment and control interventions). If the intervention requires a lengthy follow up 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 included 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 initiate the intervention of interest and the response to the intervention is measured. More than one cluster may start the intervention at a time point, but the time at which a cluster begins the intervention is randomized. Fig. 1 illustrates the differences between the parallel, traditional crossover and stepped wedge designs.

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 unidirectional. 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 are provided below.

In Section 2 we describe a trial being conducted in Washington state that uses a stepped wedge design. This motivating example provides a context for the theoretical and simulation results shown in Section 3 where we describe statistical aspects of the design and analysis of stepped wedge CRTs. In Section 4 we summarize our findings and discuss future areas of research.

## 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 US involves contact of partners by public health authorities. However, the high costs associated with this practice have 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 and/or gonnorrhea infection. The primary outcome was the presence of persistent or recurrent infection in the original index patient 3–19 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 CRT to evaluate the population-level effect of the intervention has been received from the National Institutes of Health. Using a stepped wedge design, twenty-four county health districts in Washington state will be randomized to EPT at one of four possible times (six counties at a time). Cross-sectional surveys will be conducted in each county in each time interval (and at baseline) to measure the prevalence of gonorrhea and chlamydia (with different people in each time interval). The randomization times will be separated by 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. This design will allow the evaluation of the population-level effectiveness of the EPT intervention.

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, where CV is defined to be the ratio of the between-county standard deviation over the mean prevalence. 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 points, and $N$ individuals sampled per cluster per time interval, let $Y_{ijk}$ be the response corresponding to individual k at time $j$ from cluster $i$ ($i$ in 1, ..., $I$; $j$ in 1, ..., $T$; $k$ in 1, ..., $N$) 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 $\alpha_i$ is a random effect for cluster $i$ such that $\alpha_i \sim N(0, \tau^2)$, $\beta_j$ is a fixed effect corresponding to time interval $j$ ($j$ in 1, ..., $T-1$, $\beta_T=0$ for identifiability), $X_{ij}$ is an indicator of the treatment mode in cluster $i$ at time $j$ (1 = intervention; 0 = control), and $\theta$ is the treatment effect.

Individual level responses may be modelled as

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

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

$$Y_{ij} = \mu_{ij} + e_{ij} \tag{3}$$

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

The variance of an individual-level response is

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

and the variance of the cluster-level response is

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

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 on the variance in terms of the coefficient of variation, $\tau/\mu$ [15].

If the individual level responses are binary then the cluster level response $Y_{ij}$ is a proportion and it is reasonable to assume that $\sigma_e^2=\mu*(1-\mu)$. The model (3) is easily adapted to handle different numbers of individuals sampled per cluster per time interval 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. Initially, we focus on equal cluster sizes and the analysis of cluster-level means. We then extend the discussion to the unequal cluster size situation and individual-level analyses.

#### 3.2.1. $\tau^2$ and $\sigma^2$ known

Model (3) is an example of a linear mixed model (LMM). If the values of the variance components $\tau^2$ and $\sigma^2$ 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 $\hat{\eta}=(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Y})$ (so $\hat{\theta}$ is the $T+1$st element of $\hat{\eta}$) 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

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

Since $\tau^2$ and $\sigma^2$ are seldom known this approach is generally not applicable for data analysis, but provides a useful approach to pre-trial power analyses.

#### 3.2.2. $\tau^2$ and $\sigma^2$ 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. In addition, this approach can be used even with non-normal individual-level data (e.g. binary responses) if the cluster sizes are approximately equal, since the analysis can then be done at the cluster mean level. However, if the responses are non-normal and the cluster sizes vary then an efficient analysis at the cluster mean level requires weights that depend on the unknown parameters, $\tau^2$ and $\sigma^2$. In this case an analysis at the individual level using generalized linear mixed models (GLMM) or generalized estimating equations (GEE) is preferred.

GLMM is an extension to the LMM procedure for non-normal data [24]. The expected value of the outcome, which may be binary, a count or a continuous response, is linked to the linear predictor (1) via a (possibly) nonlinear transformation. The underlying distribution of the outcome can follow any distribution in the exponential family. Use of a GLMM facilitates modeling of individual level binary responses since a logit link can be used to analyze individual-level data. Also, an individual-level GLMM-based analysis automatically provides proper weighting when cluster sizes vary. Software to fit such models has recently been incorporated into many general statistical packages.

Alternatively, generalized estimating equations (GEE) [17], 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 [18]. As with GLMM, GEE is more natural than LMM for individual-level binary outcomes since a logit link can be used to analyze the individual-level data and the individual-level analysis automatically accounts for variable cluster sizes. However, Sharples and Breslow [23] show that the GEE procedure tends to give inflated type I error rates when the number of clusters is small.

The above methods (LMM, GEE, GLMM) 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 GEE and LMM approaches for parallel design CRTs. Section 3.7 uses simulations to compare these three 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 (1)), then a within-cluster analysis can be used to estimate the treatment effect.

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 $C_i$ and $T_i$ be the sets of time points receiving control and intervention in cluster $i$, respectively. Then, a within-cluster estimate of $\theta$ is given by

$$\widetilde{\theta}=\frac{1}{I}\sum_i\left[\frac{\sum_{j\in\tau_i}Y_{ij\cdot}}{T-w_i}-\frac{\sum_{j\in C_i}Y_{ij\cdot}}{w_i}\right] \tag{4}$$

and under model (3) (assuming all $\beta_j=0$), the variance is given by

$$\mathrm{Var}(\widetilde{\theta})=\frac{\sigma^2}{I^2}\sum_i\left(\frac{1}{w_i}+\frac{1}{T-w_i}\right) \tag{5}$$

Notice that this variance formula does not depend on $\tau^2$ since the cluster effect, $\alpha_i$, cancels out in the computation of $\tilde{\theta}$. In this scenario, the paired $t$-test is an appropriate method for testing the hypothesis of no treatment effect.

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 (4) will, in general, be biased. The bias is a linear combination of $\beta_1, \ldots, \beta_{T-1}$:

$$b(\widetilde{\theta},\theta)=\frac{1}{I}\sum_i\left[\frac{\sum_{j\in\tau_i}\beta_j}{T-w_i}-\frac{\sum_{j\in C_i}\beta_j}{w_i}\right]=\sum_j\beta_j\sum_i\frac{w_i-T(1-X_{ij})}{Iw_i(T-w_i)} \tag{6}$$

Thus, failure to model time effects will result in a biased estimate of the treatment effect unless $\beta_1, \ldots, \beta_T=0$. Even a WLS analysis that utilizes both within and between cluster information in estimating $\theta$ will be biased if time effects are not included in (1) [20]. Note, however, that the bias in $\tilde{\theta}$ is independent of the true value of $\theta$. Furthermore, the coefficients of the $\beta$'s in (6) can be calculated once the treatment schedule is determined. Thus, understanding of each $\beta$'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 Ha: $\theta=\theta_A$ in model (3) using a stepped wedge design with $I$ sites and $T$ time points. A Wald test may be based on $Z=\theta/\sqrt{\mathrm{Var}(\theta)}$, where $\hat{\theta}$ is the estimated treatment effect from a weighted least squares analysis (Section 3.2.1). The approximate power for conducting a two-tailed test of size $\alpha$ is given as

$$\text{power}=\Phi\left(\frac{\theta_A}{\sqrt{\mathrm{Var}(\hat{\theta})}}-Z_{1-\alpha/2}\right) \tag{7}$$

where $\Phi$ 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, $\mathrm{Var}(\hat{\theta})$ is the appropriate element of $(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}$ from the weighted least squares analysis. However, for models of the form (3) (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 $\mathrm{Var}(\hat{\theta})$ 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 per time interval it can be shown that

$$\mathrm{Var}(\hat{\theta}) = \frac{I\sigma^2(\sigma^2 + T\tau^2)}{(IU - W)\sigma^2 + (U^2 + ITU - TW - IV)\tau^2} \tag{8}$$

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 interval. For the power calculations, therefore, we use $\sigma^2 \frac{(0.05)(0.95)}{100} = 0.000475$. The 24 counties will be randomized 6 at a time, so that $T=5$. Fig. 2 shows the power of the trial as a function of effect size (expressed as a relative risk) for a coefficient of variation ($\frac{\tau}{\mu}$) 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% (from 0.05 to 0.032).

### 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. Fig. 3 illustrates the effect of varying the number of clusters randomized at each time step (so that there are fewer time steps and fewer measurement times) for the Washington State EPT trial, assuming a relative risk of 0.7 (other alternatives give similar results).

Not surprisingly, 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 Fig. 3 we see that randomizing multiple clusters at each time point and thereby reducing the overall number of measurement times significantly reduces power. Separate analyses (not shown) indicate that the loss of power is primarily due to the loss of measurement times rather than the loss of randomization times (in other words, if groups of clusters are randomized to begin the intervention simultaneously but the number of measurement times is not decreased, there is little loss of power; it is not clear why one would design a trial in this manner, however, since the trial would not be shortened). Note that the lines in Fig. 3 stay approximately "parallel" across a wide range of the CVs indicating that the loss in power is relatively independent of the coefficient of variation.

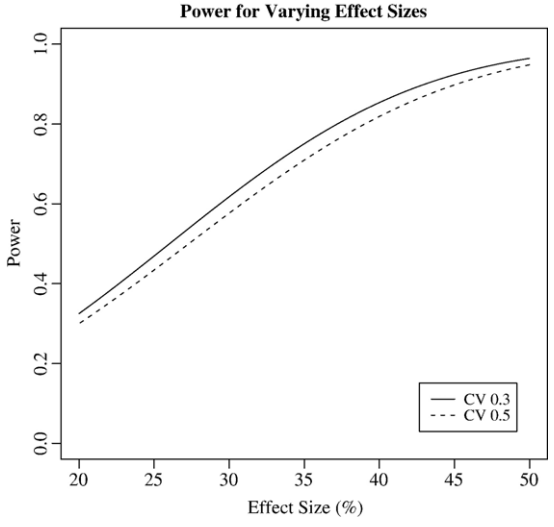


Fig. 2. Theoretical power for the Washington EPT trial. The overall prevalence is assumed to be 5%, with 100 individuals sampled per cluster per time point. Power is displayed versus effect size for two coefficients of variation.

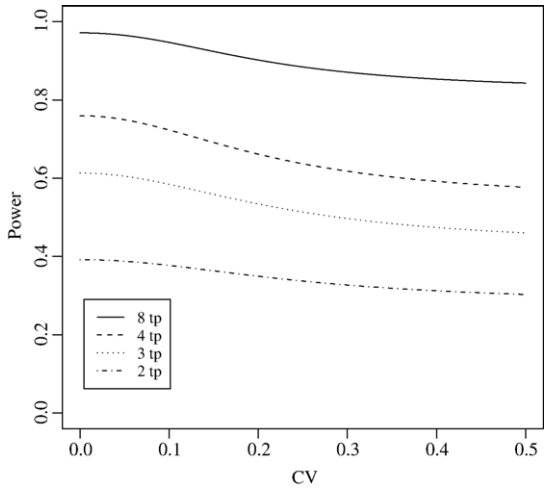


Fig. 3. Power curves when 24 clusters are randomized and number of randomization steps is varied. The number of measurement times (tp) varies from 8 (3 clusters randomized at each time) to 2 (12 clusters randomized at each time). The baseline event prevalence is 0.05 and the intervention effect corresponds to an risk ratio of 0.7.

### 3.5. Efficacy of WLS relative to a within-cluster analysis

The relative efficiency of the WLS estimator, $\hat{\theta}$ (Section 3.2.1), versus the within-cluster estimate, $\tilde{\theta}$ (Section 3.2.3), can be determined by taking the (inverse of the) ratio of the respective variances. If there are no time effects, this ratio is

$$\text{effic}\left(\hat{\theta},\tilde{\theta}\right)=\frac{\sum_i\left(\frac{1}{w_i}+\frac{1}{T-w_i}\right)\left[\left(ITU-U^2\right)\sigma^2+IT(TU-V)\tau^2\right]}{I^3(\sigma^2+T\tau^2)} \tag{9}$$

(note that the WLS variance here is different from (8) 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 within-cluster estimate in efficiency unless $\tau^2=0$ [20]. However, if time effects are included in the WLS model (so that the variance (8) is used) then $\hat{\theta}$ is less efficient than $\tilde{\theta}$ but, as described in Section 3.2.3, $\tilde{\theta}$ is likely 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, $\theta$, which can be interpreted as the maximum or full treatment effect. The delay may be modelled by allowing the $X_{ij}$ in (1) to be fractional. Power may then be calculated as outlined in Section 3.3 although the closed form expression (8) is not valid when the $X_{ij}$ are fractional.

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. Fig. 4 shows the effect of a minor delay (80%, 90%, and 100% at 1, 2, and 3 time units postintervention, 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

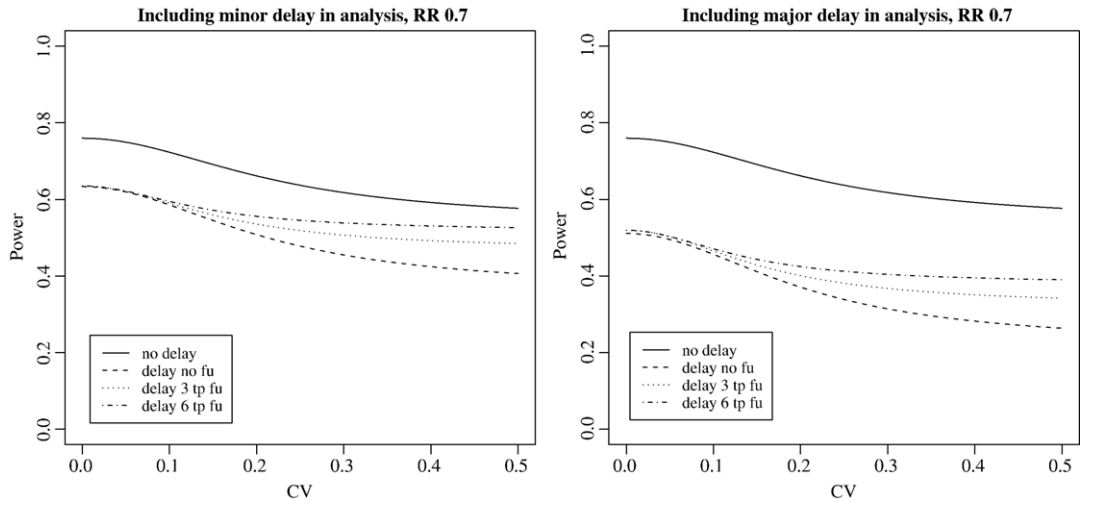


Fig. 4. Theoretical power vs. CV comparing situations in the Washington EPT trial where a minor treatment effect delay is assumed and when a major delay is assumed. Figures are shown for a risk ratio of 0.7. Plots have lines corresponding to situations with no delay, delay and no additional monitoring,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

increases power, it is difficult to recover the full power that was seen with no delay. 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. A simulation comparing analysis methods

We did a small simulation experiment to compare the size and power of the hypothesis test, $H_o$: $\theta=0$ vs. $H_a$: $\theta\neq 0$, using LMM, GEE, and GLMM methods in the context of the stepped wedge design. Individual level data were simulated using (2). LMM analyses were conducted at the cluster mean level, while GEE and GLMM analyses were done at the individual level. Analyses were done in *R*. LME was implemented using the *R* function lme() [22], GEE was implemented using gee() and GLMM was implemented using glmmPQL(). Algorithms for fitting LMM and GEE models have been fairly standardized; however, algorithms for fitting GLMM models are more variable between software packages so our results may not reflect other implementations. We evaluated two situations: where equal sample sizes are available for each cluster and where variable sample sizes are available for each cluster. These two situations correspond to 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 $\tau^2$ was assumed to be 0.000225, which corresponds to a 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 a total of 2400 individuals to 24 clusters using a multinomial distribution with parameters selected from a flat prior Dirichlet distribution (parameters (1,1,1)). Using this distribution, the interquartile range for the number of individuals per cluster was (32,168).

Table 1
Estimated power to test the hypothesis $H_o$: $\theta=0$ for designs with clusters that have the same sample size ($N=100$) and clusters with different sample sizes (24 clusters, 5 time points, $\tau^2=0.000225$, $\mu=0.05$, 1000 simulations)

| 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) |

For all methods the power using both the standard variance and a jackknife estimate of variance (in parentheses) is given.

The estimated power based on the simulations is given in Table 1. For both equal and unequal cluster sizes a jackknife estimate of the variance is needed to maintain the size of the test for both GEE and GLMM. For equal cluster sizes LMM has slightly higher power than GEE which, in turn, has greater power than GLMM (assuming jackknife variance estimates are used). The differences are not great, however. When cluster sizes vary, power is much better for GEE and GLMM compared to LMM. This is because the LMM approach analyzes the results at the cluster level and weights must be used to account for the different cluster sizes. However, the correct weights depend on the 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. In contrast, GEE and GLMM analyze these binary data at the individual level and thereby provide the correct weighting for each cluster. For this reason, we recommend using individual level analyses when cluster sizes vary significantly. A jackknife estimate of the variance is recommended to maintain the size of the test in GEE and GLMM analyses.

## 4. Discussion

Using theoretical calculations and simulation we have investigated statistical characteristics of the stepped wedge design for cluster randomized trials. In particular, we have outlined a procedure for computing power in such trials and investigated the effect of varying intercluster correlation, number of randomization steps and treatment delay on trial power. The design is relatively insensitive to variations in the intercluster correlation. We also found that, for a fixed number of clusters, power decreases as the number of randomization steps decreases. Most of the power loss is due to a reduction in the number of measurement times rather than the reduction in randomization steps, per se. 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.

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.

Analyses that rely on within-cluster information only (e.g. paired $t$-test) provide a valid analysis of the stepped wedge design only if there are no time effects. Otherwise, a within-cluster analysis provides a biased estimate of the treatment effect. A formula for the bias was derived based on the treatment schedule and the true values of time effect parameters $\beta_1, \ldots, \beta_{T-1}$. Within-cluster analyses should only be used if no significant temporal trends or fluctuations are expected over the course of the trial. However, if external or a priori information suggests that there are no time effects then an analysis based on model (3) without parameters for time still provides a more efficient analysis than the paired $t$-test.

An anonymous reviewer suggested modifying (1) by including time as a random effect. We felt that this approach did not reflect our interest in controlling for temporal trends and fluctuations in disease prevalence over the course of a particular trial (and a relatively complex model for the time effect might be required since – for infectious disease studies – adjacent time periods are unlikely to be independent). Nonetheless, we found this idea interesting and potentially applicable in some circumstances. Such an approach might be particularly appropriate if temporal variations in the outcome were thought to be due to factors unrelated to changes in the underlying disease prevalence (e.g. changes in personnel doing outcome surveys). Further development of this idea is warranted.

Using simulations, we compared LMM, GLMM, and GEE with respect to size and power for a trial with 24 clusters and 5 time intervals (to mimic the Washington state EPT trial). The simulation results agreed well with predictions based on asymptotics—LMM maintained the nominal test size and had power close to that predicted by Eq. (7) for the case of equal cluster sizes. GEE and GLMM showed evidence of inflated size that could be resolved using a jackknife variance estimate. This phenomenon may be due to the limited number of clusters [23]. Although LMM had a slight power advantage when cluster sizes were equal, GEE and GLMM were substantially more efficient than LMM when cluster sizes varied.

Model (3) 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 stratum by time interaction could be included as a factor in the model.

The stepped wedge design provides 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 that have been proven effective in individual level trials and so-called "phase IV" effectiveness trials.

## Acknowledgements

This research was supported by NIH grants AI29168, AI46702.

## References

[1] Gail MH, Byar DP, Pechacek TF, et al. Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT). Control Clin Trials 1992;13:6–21.
[2] Peterson Jr AV, Kealey KA, Mann SL, et al. Hutchinson Smoking Prevention Project: long-term randomized trial in school-based tobacco use prevention—results on smoking. J Natl Cancer Inst 2000;92:1979–91.
[3] Grosskurth H, Mosha F, Todd J, et al. Impact of improved treatment of sexually transmitted diseases on hiv infection in rural tanzania: randomised controlled trial. Lancet 1995;346:530–6.
[4] Wawer MJ, Sewankambo NK, Serwadda D, et al. Control of sexually transmitted diseases for AIDS prevention in Uganda: a randomised community trial. Lancet 1999;353:525–35.
[5] Martiniuk A, O'Connor K, King W. A cluster randomized trial of a sex education programme in Belize, Central America. Int J Epidemiol 2003;32:131–6.
[6] Sjögren T, Nissinen K, 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 crossover trial. J Int Assoc Stud Pain 2005;116:119–28.
[7] Gail MH. On design considerations and randomization-based inference for community intervention trials. Stat Med 1996;15:1069–92.
[8] Torgerson DJ. Contamination in trials: is cluster randomisation the answer? BMJ 2001;322:355–7.
[9] Donner A, 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, Hsu LN, et al. A randomized controlled trial of quality assurance in sixteen ambulatory care practices. Med Care 1985;23:751–70.
[12] Menzies R, Tamblyn R, Farant JP, et al. The effect of varying levels of outdoor air supply on the symptoms of sick building syndrome. N Engl J Med 1993;328:821–7.
[13] Gambia Hepatitis Study Group. The Gambia Hepatitis Intervention Study. Cancer Res 1987;47:5782–7.
[14] Golden M, Whittington W, Handsfield H, et al. Effect of expedited treatment of sex partners on recurrent or persistent gonorrhea or chlamydial infection. N Engl J Med 2005;352(7):676–85.
[15] Hayes R, Bennett S. Simple sample size calculation for cluster-randomized trials. Int J Epidemiol 1999;28(2):319–26.
[16] Laird N, Ware J. Random-effects models for longitudinal data. Biometrics 1982;38:963–74.
[17] Liang K, Zeger S. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
[18] Diggle P, Heagerty P, Liang K, Zeger S. Analysis of longitudinal data. 2nd ed. Oxford University Press; 2002.
[19] Feng Z, Diehr P, Peterson A, et al. Selected statistical issues in group randomized trials. Annu Rev Public Health 2001;22:167–87.
[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, 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, Bates D. Mixed-effects models in S and S-PLUS. Springer Publishing; 2000.
[23] Sharples K, Breslow N. Regression analysis of correlated binary data: some small sample results for the estimating equation approach. J Stat Comput Simul 1992;42(1):1–20.
[24] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc 1993;88:9–25.