8-6-2018

Robust Inference for the Stepped Wedge Design

James P. Hughes
University of Washington - Seattle Campus, jphughes@uw.edu

Patrick J. Heagerty
University of Washington, heagerty@uw.edu

Fan Xia
University of Washington, fanxia@uw.edu

Yuqi Ren
University of Washington, yuqir8@uw.edu

Robust Inference for the Stepped Wedge Design

James P. Hughes, Patrick J. Heagerty, Fan Xia, Yuqi Ren

Department of Biostatistics

University of Washington

Seattle, WA 98195, U.S.A.

August 6, 2018

Abstract

Based on a permutation argument, we derive a closed form expression for an estimate of the treatment effect, along with its standard error, in a stepped wedge design. We show that these estimates are robust to mis-specification of both the mean and covariance structure of the underlying data-generating mechanism, thereby providing a robust approach to inference for the treatment effect in stepped wedge designs. We use simulations to evaluate the type I error and power of the proposed estimate and to compare the performance of the proposed estimate to the optimal estimate when the correct model specification is known. The limitations, possible extensions, and open problems regarding the method are discussed.

1 Introduction

Stepped wedge designed trials (e.g. figure 1) are a type of cluster-randomized study in which all clusters (clinics, communities, etc.) receive the intervention but the time when the intervention is introduced to each cluster is randomized. Stepped wedge designs are often used to assess the effect of a new treatment or intervention as it is rolled out across a series of clinics or communities (Hussey and Hughes, 2007; Mdege et al., 2011; Hemming et al., 2015). Estimation of the intervention effect in a stepped wedge design is more difficult than in a simple parallel cluster-randomized trial since the stepped wedge design induces a conlinearity between time and the intervention. Mixed effects regression analyses are often used to disentangle these effects (e.g. Hemming et al. (2015); Hooper et al. (2016)) but this approach depends heavily on modelling assumptions, including the functional form chosen for time, the assumption of similar time trends across clusters. and the covariance structure within and between cluster-periods. Misspecification of any of these factors may result in incorrect inference (Thompson et al., 2017). Generalized estimating equations (GEE) provide an alternative analysis approach that is robust to misspecification of the covariance structure; however, GEE still requires correct modelling of the time trend and gives inflated type I error rates when the number of clusters is small (Scott et al., 2017).

Since the cluster is the unit of randomization in a stepped wedge trial, an alternative approach to evaluating the intervention may be based on a permutation test that permutes the treatment sequences among the clusters. Ji et al. (2017) considered properties of permutation tests for stepped wedge designs when the underlying mean (fixed effect) structure of the data generating model is correctly specified, although they do consider situations in which the variance structure is misspecified. Wang and DeGruttola (2017) also investigated the behavior of permutation tests compared to mixed effects models when the mixed effect model fixed effects and variance structure are correctly specified but the error distribution may be misspecified. Most recently Thompson et al. (2018) derive an estimator based on com-

bining weighted within-period comparisons (so-called "vertical" comparisons (Davey et al., 2015)) of cluster-level summaries, similar in spirit to the estimate we define below. They develop both a nonparametric test using a permutation procedure and a parametric procedure in which the variance-covariance components of the proposed estimator are derived using generalized estimating equations. In the following we consider the characteristics of a design-based estimate of the treatment effect when both the mean and variance structure of the data-generating model may be misspecified. We show that even with such a highly misspecified model the proposed estimate is unbiased for the intervention effect and provides valid hypothesis tests and confidence intervals. Further, the estimate and test statistic can be computed from closed-form expressions i.e. no computer intensive permutation procedure is necessary. The result is a robust procedure for inference in stepped wedge randomized trials.

		Time period				
		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: Schematic representation of stepped wedge designs with 4 intervention sequences, one cluster per sequence, and 5 time periods. 0 indicates control condition and 1 represents treatment.

2 Methods

Consider a stepped wedge design with N clusters and T time periods (e.g., in figure 1 $N = 4$, $T = 5$). Often, N is an integer multiple of $T - 1$. Let y_{ijk} be the observation on individual k in cluster i at time j. Assume that clusters are independent and that the number of individuals measured in each cluster-period is constant i.e. $n_{ij} = n$ (we return

to this second assumption later). Let $\mathbf{y}$ denote the NTn-vector $(y_{111}, y_{112} \ldots y_{NTn})$. Let x_{ij} indicate whether the intervention is provided ($x_{ij} = 1$) or not ($x_{ij} = 0$) in cluster i at time j and let $\mathbf{x}$ denote the corresponding NTn individual-level vector where each x_{ij} is replicated n times.

Assume that $\mathbf{y}$ has been generated with mean and variance

$$\begin{aligned} E_Y(\mathbf{y}) &= \mu + \mathbf{x}\delta + \mathbf{z}\beta \\ V_Y(\mathbf{y}) &= \Sigma \end{aligned} \tag{1}$$

where $\mathbf{z}$ is the design matrix for the temporal trend, Σ is a (block diagonal) variance-covariance matrix, and μ, δ and β are the parameters for the baseline mean, intervention effect and time effect, respectively. We explicitly do not make any distributional assumptions in (1).

Suppose we completely ignore the underlying time trend, $\mathbf{z}\beta$, and the true covariance structure, Σ and fit the following model

$$\mathbf{y} \sim (\mu^\star + \mathbf{x}\delta^\star, \sigma^2\mathbf{I}). \tag{2}$$

In this model, provided $n_{ij} = n$, identical estimates are obtained regardless of whether the model is fit based on individual-level data or cluster-period means. Therefore, let $\mathbf{Y}$ be the vector of cluster-period level means, $(Y_{11}, Y_{12}, \ldots Y_{NT})$, where $Y_{ij} = \sum_k y_{ijk}/n$ and, similarly, let $\mathbf{X}$ denote the cluster-period level vector $(x_{11}, x_{12}, \ldots x_{NT})$. The least squares estimate of $(\mu^\star, \delta^\star)$ is

$$\left(\hat{\mu}^\star, \hat{\delta}^\star\right) = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Y} \tag{3}$$

where $\mathbf{W} = [\mathbf{1}, \mathbf{X}]$ is a $NT \times 2$ matrix with the first column all ones and the second column equal to $\mathbf{X}$. Letting f denote the proportion of the cluster-periods that are assigned to the

intervention condition (e.g., in figure 1, $f = 1/2$), it is straightforward to show that

$$(\mathbf{W}^T\mathbf{W}) = NT \begin{pmatrix} 1 & f \\ f & f \end{pmatrix}$$

so

$$(\mathbf{W}^T\mathbf{W})^{-1} = \frac{1}{f(1-f)NT} \begin{pmatrix} f & -f \\ -f & 1 \end{pmatrix}.$$

Then, based on (3),

$$\hat{\delta}^\star = \frac{1}{f(1-f)NT} \sum_{ij} Y_{ij}(x_{ij} - f). \tag{4}$$

$\hat{\delta}^\star$ is, of course, a biased estimate of δ (Rao, 1971). However, consider the distribution of $\hat{\delta}^\star$ with respect to the permutation distribution of the stepped wedge design. As noted above, the permutation distribution is obtained by permuting the rows of the stepped wedge design matrix in figure 1. Importantly, $(\mathbf{W}^T\mathbf{W})^{-1}$ is the same for any permutation of the rows in figure 1.

Let $E_\mathcal{P}$ and $V_\mathcal{P}$ denote expectation and variance, respectively, under the permutation distribution. Then

$$\begin{aligned} E_\mathcal{P}(x_{ij}) &= \bar{x}_j & (5) \\ V_\mathcal{P}(x_{ij}) &= \bar{x}_j(1-\bar{x}_j) & (6) \\ Cov_\mathcal{P}(x_{ij}, x_{i'j'}) &= \bar{x}_j(1-\bar{x}_{j'}) \qquad i = i', j < j' \\ &= -\frac{1}{N-1}\bar{x}_j\left(1-\bar{x}_{j'}\right) \qquad i \neq i', j \leq j' & (7) \end{aligned}$$

where $\bar{x}_j = \sum_i x_{ij}/N$. These results make use of the stepped wedge design feature that an intervention is never removed once introduced (i.e. $x_{ij} \leq x_{ij'}$ for $j < j'$).

Since $\mathbf{Y}$ is constant with respect tot he permutation distribution then, based on (5),

$$E_{\mathcal{P}}(\hat{\delta}^{\star}) = \frac{1}{f(1-f)NT}\sum_{ij} Y_{ij}(\bar{x}_j - f) \tag{8}$$

and combining (4) and (8) gives

$$\Delta = \hat{\delta}^{\star} - E_{\mathcal{P}}(\hat{\delta}^{\star}) = \frac{1}{f(1-f)NT}\sum_{ij} Y_{ij}(x_{ij} - \bar{x}_j). \tag{9}$$

Now consider the expectation of Δ under the (true) distribution of Y. From (1)

$$E_Y(Y_{ij}) = \mu + x_{ij}\delta + z_{ij}\beta \tag{10}$$

where z_{ij} is the row (vector) of $\mathbf{z}$ corresponding to the i,j'th observation. Most stepped wedge models assume that the temporal component of the model is constant across all clusters. This implies that $z_{ij}\beta$ does not depend on i. Then, since $\sum_i (x_{ij} - \bar{x}_j) = 0$ and $x_{ij}^2 = x_{ij}$,

$$E_Y(\Delta) = \delta\frac{1}{f(1-f)NT}\sum_{ij} x_{ij}(1-\bar{x}_j) = \delta\frac{1}{f(1-f)T}\sum_{j}\bar{x}_j(1-\bar{x}_j) \tag{11}$$

Importantly, this implies that, using the permutation distribution, the treatment effect, δ, can be estimated unbiasedly even if the temporal portion of the model is ignored. Specifically,

$$\hat{\delta} = \frac{\Delta}{\frac{1}{f(1-f)T}\sum_j \bar{x}_j(1-\bar{x}_j)} = \frac{\sum_{ij} Y_{ij}(x_{ij}-\bar{x}_j)}{N\sum_j \bar{x}_j(1-\bar{x}_j)}. \tag{12}$$

If the assumption of temporal constancy across clusters is violated then permutations could be done within strata for which the assumption is met, and the argument carries through (see the appendix for formulas).

Now consider the variance of $\hat{\delta}$. Assuming independence between clusters, straightforward

calculations based on (12) give

$$V_Y(\hat{\delta}) = \frac{\sum_i \left(\sum_j \text{Var}(Y_{ij})(x_{ij} - \bar{x}_j)^2 + 2 \sum_{j<j'} \text{Cov}(Y_{ij}, Y_{ij'})(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}) \right)}{(N \sum_j \bar{x}_j(1 - \bar{x}_j))^2} \quad (13)$$

In the special case where the covariance matrix of Y does not depend on cluster (i.e. $\text{Var}(Y_{ij}) = \sigma_j^2$ and $\text{Cov}(Y_{ij}, Y_{ij'}) = \sigma_{j,j'}$) (13) reduces to

$$V_Y(\hat{\delta}) = \frac{\left(\sum_j \sigma_j^2 \bar{x}_j(1 - \bar{x}_j) + 2 \sum_{j<j'} \sigma_{j,j'} \bar{x}_j(1 - \bar{x}_{j'}) \right)}{N(\sum_j \bar{x}_j(1 - \bar{x}_j))^2} \quad (14)$$

Expressions (13) and (14) depend on the true variance-covariance matrix and are, therefore, of limited utility in practice. Instead, we seek a variance estimator that does not depend on knowledge of the true variance of the data-generating process. We accomplish this by considering the variance of $\hat{\delta}$ across the permutation distribution and derive two unbiased variance estimates that can be used for inference (see appendix for derivation).

The first is suitable for any stepped wedge design and is given by

$$\begin{aligned} V_\delta^1(\hat{\delta}) \;=\; & \left\{ \sum_i \left[\sum_j (Y_{ij} - x_{ij}\delta)^2 \bar{x}_j(1 - \bar{x}_j) + 2 \sum_{j<j'} (Y_{ij} - x_{ij}\delta)(Y_{ij'} - x_{ij'}\delta) \bar{x}_j(1 - \bar{x}_{j'}) \right] \right. \\ & \left. - \frac{2}{N-1} \sum_{i<i'} \sum_{j,j'} (Y_{ij} - x_{ij}\delta)(Y_{i'j'} - x_{i'j'}\delta) \bar{x}_{\min(j,j')} \left(1 - \bar{x}_{\max(j,j')}\right) \right\} \Big/ \\ & (N \sum_j \bar{x}_j(1 - \bar{x}_j))^2. \end{aligned} \quad (15)$$

$V_\delta^1(\hat{\delta})$ has expectation (with respect to the distribution of Y) equal to $V_Y(\hat{\delta})$ when the covariance matrix of Y_i does not depend on cluster (i.e. (14)). If $\hat{\delta}$ is used in place of δ in (15) then $V_{\hat{\delta}}^1(\hat{\delta})$ is a biased estimate of $V_Y(\hat{\delta})$. The bias is a complex expression that depends in part on the true variance of Y. Nonetheless, in simulations we have run so far, the simple adjustment of multiplying $V_{\hat{\delta}}^1(\hat{\delta})$ by $N/(N-1)$ provides an approximately unbiased estimate,

especially for large N. We investigate the behavior of both $V^1_{\delta}(\hat{\delta})$ and $V^1_{\hat{\delta}}(\hat{\delta})$ in simulations in section 3.

The restriction that (15) is unbiased only when the variance of Y_i does not depend on cluster is non-trivial. Two examples where this assumption is violated are i) there is a cluster $\times$ intervention interaction (random intervention effect); in the presence of a random intervention effect the covariance of Y_i depends on the intervention sequence and hence on cluster; ii) the sample size varies by cluster. The second proposed variance estimate does not depend on the assumption that the variance of Y_i is independent of i. However, this alternative variance estimate does require that each intervention sequence (each row of figure 1) is replicated at least once (i.e. there are two or more clusters with each sequence). Specifically, suppose there are $m(h)$ clusters with intervention sequence h ($m(h) > 1$ for all h). Let Y_{hij} denote the cluster-period mean for cluster i in sequence h ($i = 1 \ldots m(h)$) at time j and similarly for the intervention indicators x_{hij}. Then

$$\begin{aligned} V^2(\hat{\delta}) \;=\; & \sum_h \Bigg\{ \sum_{i=1}^{m(h)} \left[\sum_j Y_{hij}^2 (x_{hij} - \bar{x}_j)^2 + 2 \sum_{j<j'} Y_{hij} Y_{hij'} (x_{hij} - \bar{x}_j)(x_{hij'} - \bar{x}_{j'}) \right] \\ & - \frac{2}{n(h) - 1} \sum_{i<i'} \sum_{j,j'} Y_{hij} Y_{hi'j'} (x_{hij} - \bar{x}_j)(x_{hi'j'} - \bar{x}_{j'}) \Bigg\} \Bigg/ (N \sum_j \bar{x}_j (1 - \bar{x}_j))^2 \end{aligned} \tag{16}$$

has expectation equal to (13) for any covariance matrix structure. In addition, $V^2(\hat{\delta})$ does not depend on δ. We evaluate the relative performance of $V^1_{\delta}(\hat{\delta})$ and $V^2(\hat{\delta})$ using simulations in section 3.

No distributional assumptions have been necessary for the development thus far. For inference, we assume that either individual observations are normally distributed or the central limit theorem holds, which is a reasonable assumption in most cases since the analysis is based on (sums of) cluster-period level means. In that case, the estimates and variances

derived above can be used to test the hypothesis Ho: $\delta = \delta_o$ using a Z statistic such as

$$Z_1 = \frac{\hat{\delta} - \delta_o}{\sqrt{V^1_{\delta_o}(\hat{\delta})}} \quad \text{or} \quad Z_2 = \frac{\hat{\delta} - \delta_o}{\sqrt{V^2(\hat{\delta})}} \tag{17}$$

Further, a $100 * (1 - \alpha)\%$ confidence interval for $\hat{\delta}$ may be defined as

$$\{\delta : Z_{\alpha/2} \leq (\hat{\delta} - \delta)/\sqrt{V^1_{\delta}(\hat{\delta})} \leq Z_{1-\alpha/2}\} \tag{18}$$

or by the interval

$$\hat{\delta} \pm Z_{1-\alpha/2} * \sqrt{V^2(\hat{\delta})} \tag{19}$$

where Z_α is the α'th percentile of the standard normal distribution.

3 Simulation Results

We simulate datasets for a stepped wedge design with $T = 5$ time periods and varying numbers of clusters. Data were simulated from the mixed model

$$Y_{ijk} = \mu + \beta_j + x_{ij}\delta + a_i + b_{ij} + c_i x_{ij} + e_{ijk} \tag{20}$$

where $\beta = (0, -0.1, -0.2, -0.3, -0.4)$ in all simulations and

$$\begin{aligned} a_i &\sim N(0, \tau^2) \\ b_{ij} &\sim N(0, \psi^2) \\ c_i &\sim N(0, \eta^2) \\ e_{ijk} &\sim N(0, \sigma^2) \end{aligned}$$

Table 1 gives the values of the variance components used in five specific scenarios.

Table 1: Scenarios for stepped wedge simulations

	Scenario				
	1	2	3	4	5
Random Effects	None	Cluster	Cluster Intervention	Cluster Time	Cluster Intervention Time
σ^2	1	1	1	1	1
τ^2	0	0.2	0.2	0.2	0.2
η^2	0	0	0.1	0	0.1
ψ^2	0	0	0	0.04	0.04

Table 2 shows confidence interval coverage of the proposed estimator across the five scenarios shown in table 1 for N = 12, 24 and 36 clusters and where the number of observations per cluster per time period(n) is either constant (n = 10) or varies between clusters (with average = 10) according to a lognormal (rounded to the nearest integer) with variance 0.2 (low var) or variance 1.0 (high var). As expected, the estimator is unbiased for all cluster sizes, sample sizes and scenarios (data not shown). Coverage using the variance estimator $V_\delta^1(\hat{\delta})$ is close to the nominal 95% across all scenarios, even in the scenarios where $\text{Var}(Y_i)$ varies across clusters (scenarios 3 and 5, and the scenarios with nonconstant n). $V_{\hat{\delta}}^1(\hat{\delta})$ (multiplied by the correction factor $N/(N-1)$) also generally gives good coverage although we note some undercoverage when $N = 12$. In contrast, use of the variance estimator $V^2(\hat{\delta})$ generally results in confidence intervals with greater undercoverage, although this also improves as N increases.

Tables 3 and 4 give type I error rates and power, respectively, from 10,000 simulations for tests of the null hypothesis Ho: $\delta = 0$ using the variance estimators in equations (15) and (16). When $\delta = \hat{\delta}$ is used in (15) a correction factor of $N/(N-1)$ is applied to the variance estimate. Interestingly, even though the assumptions for using $V_\delta^1(\hat{\delta})$ are only met when n is constant and $\eta^2 = 0$, the type I error rate using this variance estimate is quite close to nominal levels under all scenarios when $\delta = 0$. When $V_\delta^1(\hat{\delta})$ is used with $\delta = \hat{\delta}$ some slight type I error inflation is observed with $N = 12$ but nominal levels are achieved as N increases

Table 2: Confidence interval coverage, based on 1000 simulations, of the proposed estimator using three difference variance formulas for the five simulation scenarios described in table 1. Number of clusters (N) is 12, 24 or 36. Number of time periods (T) is 5 with a linearly decreasing time effect ($\beta = 0, -0.1, -0.2, -0.3, -0.4$). Number of individuals per cluster per time period (n) is either constant (n = 10) or varies between clusters (average = 10) according to a lognormal with variance 0.2 (low var) or lognormal with variance 1.0 (hi var). Nominal coverage is 95%.

		$V_\delta^1(\hat{\delta})$					$\frac{N}{N-1}V_\delta^1(\hat{\delta})$					$V^2(\hat{\delta})$				
		Simulation scenario														
N	n	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
12	constant	0.95	0.94	0.94	0.96	0.96	0.93	0.92	0.91	0.94	0.93	0.90	0.88	0.89	0.90	0.91
	low var	0.96	0.96	0.96	0.97	0.95	0.94	0.94	0.94	0.95	0.92	0.92	0.92	0.91	0.91	0.90
	hi var	0.97	0.95	0.95	0.96	0.96	0.96	0.94	0.93	0.94	0.93	0.91	0.90	0.90	0.90	0.90
24	constant	0.96	0.94	0.94	0.94	0.95	0.96	0.93	0.94	0.93	0.94	0.94	0.92	0.92	0.92	0.94
	low var	0.95	0.95	0.94	0.96	0.94	0.95	0.93	0.93	0.95	0.94	0.93	0.92	0.92	0.94	0.93
	hi var	0.94	0.96	0.94	0.95	0.94	0.93	0.94	0.94	0.94	0.94	0.93	0.93	0.94	0.93	0.93
36	constant	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.94	0.94	0.93	0.93	0.94	0.93	0.94
	low var	0.94	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.94	0.94	0.93	0.94	0.94	0.93	0.93
	hi var	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94

(we speculate that use of a t-distribution as a reference may produce type I error rates close to nominal levels over the entire range of N; however, the correct degrees of freedom calculation is unclear). Use of the variance estimate $V^2(\hat{\delta})$ results in greater type I error inflation for small N, although the type I error rate again approaches nominal levels as N increases. For tests based on $V_\delta^1(\hat{\delta})$ the type I error rates appear to more sensitive to the variance of the random treatment effect compared to the variation in the number of individuals per cluster-time period, at least across the ranges investigated in these simulations.

Figure 2 shows the power for various effect sizes (based on 1000 simulations) of a test based on the proposed estimator (using variance equation (15)) for testing Ho: $\delta = 0$ versus a test based on the correct model that includes an appropriate time effect and within-cluster correlation structure (using the R function lmer() (Bates et al., 2015)) for five scenarios - i) independence, ii) random cluster effect iii) random cluster and treatment effects iv) random cluster and time effects v) random cluster, treatment and time effects. In each case the variance components were chosen so that the power curve for the robust test is similar

Table 3: Type I error rates for testing the null hypothesis Ho: $\delta = 0$ based on 10000 simulations. Data are simulated from equation (1) with a random cluster effect ($\tau^2 = 0.2$), random treatment effect ($\eta^2 = 0, 0.1, 0.4$) and random error ($\sigma^2 = 1$). Number of individuals per cluster per time period (n) is either constant (n = 10) or varies between clusters (average = 10) according to a lognormal with variance 0.2 (low var) or lognormal with variance 1.0 (hi var). Number of clusters (N) is 12, 24 or 36. Number of time periods (T) is 5 with a linearly decreasing time effect (-0.1 per time period).

		$V^1_{\delta=0}(\hat{\delta})$			$\frac{N}{N-1}V^1_{\delta=\hat{\delta}}(\hat{\delta})$			$V^2(\hat{\delta})$		
		Treatment variance (η^2)								
N	n	0	0.1	0.4	0	0.1	0.4	0	0.1	0.4
12	constant	0.05	0.05	0.05	0.06	0.07	0.07	0.09	0.09	0.10
	low var	0.05	0.05	0.05	0.06	0.07	0.07	0.10	0.09	0.09
	hi var	0.05	0.05	0.05	0.07	0.07	0.07	0.09	0.08	0.08
24	constant	0.05	0.05	0.06	0.06	0.06	0.06	0.07	0.07	0.07
	low var	0.05	0.05	0.06	0.06	0.06	0.06	0.07	0.07	0.07
	hi var	0.04	0.04	0.05	0.05	0.05	0.05	0.07	0.06	0.06
36	constant	0.05	0.05	0.06	0.05	0.06	0.06	0.06	0.06	0.06
	low var	0.05	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.06
	hi var	0.05	0.06	0.06	0.06	0.06	0.07	0.06	0.06	0.06

(specifically, σ^2 was varied and the ICC's for the cluster, treatment and time random effects were set at 0.17, 0.091 and 0.038, respectively). The proposed estimator is less efficient than the maximum likelihood estimate based on the correct model since the latter uses both within-cluster and between-cluster information to estimate the treatment effect. However, this gain in efficiency must be balanced against the potential for inflation of the type I error rate when the covariance structure used for the analysis does not match the data generating mechanism. In general, the type I error will be inflated if the model used for analysis does not include all the random effects from the data-generating mechanism. The proposed estimator is robust to such model misspecification, as well as misspecification of the time trend.

Table 4: Power for testing the null hypothesis Ho: $\delta = 0$ versus Ha: $\delta = 1$, based on 10000 simulations. Data are simulated from equation (1) with a random cluster effect ($\tau^2 = 0.2$), random treatment effect ($\eta^2 = 0, 0.1, 0.4$) and random error ($\sigma^2 = 1$). Number of individuals per cluster per time period (n) is either constant (n = 10) or varies between clusters (average = 10) according to a lognormal with variance 0.2 (low var) or lognormal with variance 1.0 (hi var). Number of clusters (N) is 12, 24 or 36. Number of time periods (T) is 5 with a linearly decreasing time effect (-0.1 per time period).

		$V^1_{\delta=0}(\hat{\delta})$			$\frac{N}{N-1}V^1_{\delta=\hat{\delta}}(\hat{\delta})$			$V^2(\hat{\delta})$		
		Treatment variance (η^2)								
N	n	0	0.1	0.4	0	0.1	0.4	0	0.1	0.4
12	constant	0.59	0.57	0.51	0.64	0.61	0.55	0.65	0.62	0.55
	low var	0.59	0.56	0.49	0.63	0.60	0.53	0.64	0.60	0.54
	hi var	0.42	0.41	0.38	0.46	0.45	0.42	0.48	0.47	0.42
24	constant	0.90	0.87	0.81	0.90	0.88	0.82	0.90	0.88	0.80
	low var	0.89	0.87	0.81	0.90	0.88	0.82	0.90	0.88	0.81
	hi var	0.78	0.76	0.71	0.79	0.78	0.72	0.82	0.79	0.73
36	constant	0.97	0.97	0.94	0.98	0.97	0.94	0.97	0.97	0.93
	low var	0.97	0.96	0.93	0.97	0.96	0.94	0.97	0.96	0.93
	hi var	0.84	0.83	0.79	0.85	0.84	0.80	0.84	0.82	0.78

4 Example

The Washington state EPT trial was a stepped wedge trial of expedited partner treatment (EPT - the practice of treating the sex partners of persons with sexually transmitted infections without prior medical evaluation of the partner) for the prevention of chlamydia and gonorrhea infection. The trial was conducted between July, 2007 and August, 2010. The primary outcome for the trial was chlamydia positivity, measured in sentinel sites throughout Washington state during the course of the trial, and incidence of reported gonorrhea, both in women. Twenty two local health jurisdictions (LHJs - equal to counties or clusters of counties in the state) were randomized to one of four different intervention sequences. Additional details are provided in Golden et al. (2015).

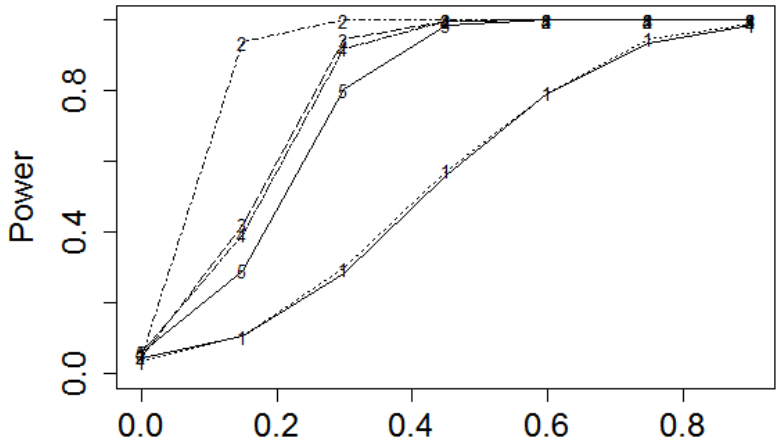

Figure 2: Power curves for testing Ho: $\delta = 0$ computed across 1000 simulations for the robust test proposed here (solid line) versus an asymptotic test based on the correctly specified model for the five scenarios described in table 1. In each scenario the variance components where chosen so that the power curve for the robust test was similar; specifically, σ^2 was varied and the ICC's for the cluster, intervention and time random effects (when included) were set at 0.17, 0.091 and 0.038, respectively.

Table 5 shows the trial design and chlamydia positivity by LHJ and time. The median sample size per cluster-period was 171 (IQR: 78 - 396). The risk difference due to the intervention is estimated as -0.015 (95%CI: -0.033 - 0.003; p = 0.10). For comparison, Golden et al. (2015) report a relative risk of 0.89 (p = 0.15) from a baseline positivity rate of 0.083, equivalent to a risk difference of -0.009.

5 Discussion

We have developed a design-based approach for obtaining unbiased estimates of the intervention effect and doing robust inference (confidence intervals and hypothesis tests) in a stepped wedge study design. Although the methods are motivated by permutation arguments, closed form expressions for the intervention effect estimate and its variance are derived, so the ap-

proach is computationally easy. The proposed methods do not depend on detailed knowledge of the temporal mean structure, covariance structure or distribution of the data-generating mechanism. Similar to Thompson et al. (2018), the intervention effect estimate derived here is a "vertical" estimate (Davey et al., 2015) i.e. it relies only on between-cluster information on the intervention effect. This explains the robustness to misspecification of the time-trend - a comparison of intervention and control clusters at a point in time (between-cluster comparison) does not depend on the underlying time trend whereas any within-cluster comparison of intervention and control periods must first correct for time trends. While the reliance on between-cluster comparisons helps explain the robustness of the proposed intervention effect estimate, this also explains the loss of efficiency relative to the intervention effect estimate from a correctly specified model that uses both between-cluster and within-cluster information.

We have developed three variance estimates that can be used for inference, namely, $V^1_\delta(\hat{\delta})$, $V^1_{\hat{\delta}}(\hat{\delta})$ and $V^2(\hat{\delta})$. The first two (collectively, V^1) assume that $Var(Y_i)$ does not depend on i while the last (V^2) does not depend on this assumption. However, V^1 appears to be relatively insensitive to violations of this assumption and in simulations $V^1_\delta(\hat{\delta})$ performs well across a range of scenarios. $V^1_{\hat{\delta}}(\hat{\delta})$ is a biased estimate of $V_Y(\hat{\delta})$ and the bias depends on the true variance. Although multiplying $V^1_{\hat{\delta}}(\hat{\delta})$ by a correction factor of $N/(N-1)$ performed reasonably well in our simulations, further research is needed before this approach can be broadly recommended.

A key assumption (which is necessary for all approaches to the analysis of stepped wedge trials) is that the underlying time trend is the same for all clusters. If this assumption is violated then $\hat{\delta}$ may be biased and estimates of $V(\hat{\delta})$ may be incorrect as well. However, if clusters can be grouped into strata with similar temporal trends (ideally, these strata would be defined apriori) then it is possible to derive an estimate of the intervention effect and its sampling variance based on a stratified permutation distribution. The resulting estimate is unbiased and has correct sampling variance (under the same constraints/assumptions as

(15) and (16)). Formulas for these stratified estimates are given in the appendix.

The approach outlined here uses cluster-period level summaries and should, therefore, be robust to the underlying distribution of individual data points, provided the cluster-period sample sizes are moderately large. Thus, the proposed methods may be used with continuous, binary or count data, and the intervention effect will be interpretable as a mean difference, risk difference, or rate difference, respectively. Critically, however, the equivalence between an individual-level analysis and an analysis of cluster-period means used to derive $\hat{\delta}$ only holds for the identity link. Specifically, if $\hat{\delta}$ is computed using (nonlinearly) transformed cluster-period level summaries (e.g., $\log(Y_{ij})$ or $\text{logit}(Y_{ij})$), it will be a biased estimate of the intervention effect from an individual-level model with the corresponding nonlinear link. An extension of the proposed methods to other links to allow unbiased estimation of e.g. risk ratios is an area of ongoing research.

Acknowledgements: This research was supported by PCORI contract ME-1507-31750 and NIH grant AI29168.

Table 5: Chlamydia positivity in women by LHJ and time from the Washington state EPT trial. Shaded areas indicate the times when the intervention was provided.

	Time				
LHJ	0	1	2	3	4
1	0.10	0.07	0.07	0.07	0.09
2	0.08	0.06	0.04	0.06	0.06
3	0.23	0.13	0.04	0.04	0.06
4	0.05	0.06	0.05	0.05	0.06
5	0.03	0.07	0.06	0.02	0.08
6	0.04	0.05	0.07	0.04	0.05
7	0.09	0.13	0.10	0.10	0.12
8	0.11	0.10	0.10	0.07	0.08
9	0.04	0.03	0.05	0.04	0.08
10	0.04	0.07	0.03	0.04	0.06
11	0.02	0.07	0.05	0.06	0.02
12	0.09	0.09	0.09	0.11	0.05
13	0.12	0.06	0.07	0.06	0.06
14	0.05	0.09	0.07	0.05	0.04
15	0.12	0.07	0.10	0.07	0.05
16	0.18	0.10	0.08	0.07	0.12
17	0.06	0.14	0.10	0.06	0.07
18	0.08	0.07	0.15	0.08	0.10
19	0.09	0.10	0.07	0.05	0.05
20	0.04	0.05	0.17	0.12	0.02
21	0.11	0.07	0.06	0.06	0.08
22	0.06	0.05	0.04	0.04	0.05

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Davey, C., Hargreaves, J., Thompson, J., Copas, A., Beard, E., Lewis, J., and Fielding, K. (2015). Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials*, 16:358.

Golden, M., Kerani, R., M, S., Hughes, J., Aubin, M., Malinski, C., and Holmes, K. (2015). Uptake and population-level impact of expedited partner therapy (ept) on chlamydia trachomatis and neisseria gonorrhoeae: The washington state community-level randomized trial of ept. *PLoS ONE*, 12:e1001777.

Hemming, K., Lilford, R., and Girling, A. (2015). Stepped-wedge cluster randomised controlled trials: A generic framework including parallel and multiple level designs. *Statistics in Medicine*, 34:181–196.

Hooper, R., Teerenstra, S., de Hoop, E., and Eldridge, S. (2016). Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Statistics in Medicine*, 35:4718 – 4728.

Hussey, M. and Hughes, J. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, 28:182–191.

Ji, X., Fink, G., Robyn, P., and Small, D. (2017). Randomization inference for stepped-wedge cluster-randomized trials: An application to community-based health insurance. *Annals of Applied Statistics*, 1:1–20.

Mdege, N., Man, M., Taylor, C., and Torgerson, D. (2011). Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *Journal of Clinical Epidemiology*, 64:936–948.

Rao, P. (1971). Notes on misspecification in multiple regression. *The American Statistician*, 25:37 – 39.

Scott, J., deCamp, A., M, J., MP, F., and PB, G. (2017). Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. *Statistical Methods in Medical Research*, 26:583–597.

Thompson, J., Davey, C., Fielding, K., Hargreaves, J., and Hayes, R. (2018). Robust analysis of stepped wedge trials using cluster-level summaries within periods. *Statistics in Medicine*, 37:2487–2500.

Thompson, J., Fielding, K., Davey, C., Aiken, A., Hargreaves, J., and Hayes, R. (2017). Bias and inference from misspecified mixed-effect models in stepped wedge trial analysis. *Statistics in Medicine*, 36:3670 – 3682.

Wang, R. and DeGruttola, V. (2017). The use of permutation tests for the analysis of parallel and steppedwedge clusterrandomized trials. *Statistics in Medicine*, 36:2831 – 2843.

Appendix

Permutation variance of $\hat{\delta}$

The estimated intervention effect is

$$\hat{\delta} = \frac{\sum_{ij} Y_{ij}(x_{ij} - \bar{x}_j)}{N \sum_j \bar{x}_j(1 - \bar{x}_j)}.$$

so, letting $V_{\mathcal{P}}$ denote variance with respect to the permutation distribution,

$$V_{\mathcal{P}}(\hat{\delta}) = \frac{V_{\mathcal{P}}(\sum_{ij} Y_{ij}(x_{ij} - \bar{x}_j))}{(N \sum_j \bar{x}_j(1 - \bar{x}_j))^2}$$

since $\bar{x}_j$ is constant across permutations. Further, since Y_{ij} is also constant with respect to the permutation distribution, one can use (5) - (7) to show that the variance of $\hat{\delta}$ over all possible permutations is

$$\begin{aligned} V_{\mathcal{P}}(\hat{\delta}) &= \Bigg\{ \sum_i \left[\sum_j Y_{ij}^2 \bar{x}_j(1 - \bar{x}_j) + 2 \sum_{j<j'} Y_{ij} Y_{ij'} \bar{x}_j(1 - \bar{x}_{j'}) \right] \\ &\quad - \frac{2}{N-1} \sum_{i<i'} \sum_{j,j'} Y_{ij} Y_{i'j'} \bar{x}_{\min(j,j')} \left(1 - \bar{x}_{\max(j,j')}\right) \Bigg\} / (N \sum_j \bar{x}_j(1 - \bar{x}_j))^2. \end{aligned}$$

Unfortunately, however, $V_{\mathcal{P}}(\hat{\delta})$ is a biased estimator of $V_Y(\hat{\delta})$. To find the expected value of $V_{\mathcal{P}}(\hat{\delta})$ with respect to the distribution of Y we make use of

$$\begin{aligned} E_Y(Y_{ij}^2) &= \mathrm{Var}(Y_{ij}) + E(Y_{ij})^2 \\ E_Y(Y_{ij}, Y_{ij'}) &= \mathrm{Cov}(Y_{ij}, Y_{ij'}) + E(Y_{ij})E(Y_{ij'}) \\ E_Y(Y_{ij}, Y_{i'j'}) &= E(Y_{ij})E(Y_{i'j'}) \end{aligned}$$

as well as $x_{ij}^2 = x_{ij}$ and $x_{ij}x_{ij'} = x_{ij}$ for $j < j'$, to derive

$$\begin{aligned} E_Y(V_{\mathcal{P}}(\hat{\delta})) &= \frac{\sum_i \left[\sum_j \mathrm{Var}(Y_{ij})\bar{x}_j(1-\bar{x}_j) + 2\sum_{j<j'} \mathrm{Cov}(Y_{ij}, Y_{ij'})\bar{x}_j(1-\bar{x}_{j'})\right] + \delta^2 C}{(N\sum_j \bar{x}_j(1-\bar{x}_j))^2} \\ C &= \sum_i \left[\sum_j x_{ij}\bar{x}_j(1-\bar{x}_j) + 2\sum_{j<j'} x_{ij}\bar{x}_j(1-\bar{x}_{j'})\right] \\ &\quad - \frac{2}{N-1}\sum_{i<i'}\sum_{j,j'} x_{ij}x_{i'j'}\bar{x}_{min(j,j')}(1-\bar{x}_{max(j,j')}). \end{aligned}$$

Comparing this to equation (13), we see that the bias depends only on δ and not other parameters of the mean model for Y. In fact, the bias of $V_{\mathcal{P}}(\hat{\delta})$ does not depend on any covariate that is constant within a column of the stepped wedge design matrix (e.g. the time parameters β_j). Using this same approach, one may show that $V_\delta^1(\hat{\delta})$ is unbiased for $V_Y(\hat{\delta})$.

Stratified Estimation

The following estimators should be used for stratified estimation, where Y_{hij} represents the observation on the i'th cluster in stratum h at time j (note that i takes on values from $1 \ldots m_h$ in these formulae):

$$\hat{\delta} = \frac{\sum_{hij} Y_{hij}(x_{hij} - \bar{x}_{hj})}{\sum_h N_h \sum_j \bar{x}_{hj}(1-\bar{x}_{hj})}.$$

$$\begin{aligned} V_\delta(\hat{\delta}) &= \Bigg\{ \sum_{hi}\left[\sum_j (Y_{hij} - x_{hij}\delta)^2 \bar{x}_{hj}(1-\bar{x}_{hj}) + 2\sum_{j<j'}(Y_{hij} - x_{hij}\delta)(Y_{hij'} - x_{hij'}\delta)\bar{x}_{hj}(1-\bar{x}_{hj'})\right] \\ &\quad - \frac{2}{N_h - 1}\sum_h \sum_{i<i'}\sum_{j,j'}(Y_{hij} - x_{hij}\delta)(Y_{hi'j'} - x_{hi'j'}\delta)\bar{x}_{h,\min(j,j')}\left(1 - \bar{x}_{h,\max(j,j')}\right)\Bigg\} \Big/ \\ &\quad (\sum_h N_h \sum_j \bar{x}_{hj}(1-\bar{x}_{hj}))^2. \end{aligned}$$