Statistics in Medicine

The batched stepped wedge design: a design robust to delays in cluster recruitment

Journal:	*Statistics in Medicine*
Manuscript ID	SIM-21-0977.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Kasza, Jessica; Monash University School of Public Health and Preventive Medicine, Department of Epidemiology and Preventive Medicine Bowden, Rhys; Monash University School of Public Health and Preventive Medicine, Department of Epidemiology and Preventive Medicine Hooper, Richard; Queen Mary University of London, Centre for Primary Care & Public Health Forbes, Andrew; Monash University School of Public Health and Preventive Medicine, Department of Epidemiology and Preventive Medicine
Keywords:	cluster randomised trial, intracluster correlation, sample size calculation, within-cluster correlation structure

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

The batched stepped wedge design: a design robust to delays in cluster recruitment

Jessica Kasza[1] | Rhys Bowden[1] | Richard Hooper[2] | Andrew B. Forbes[1]

[1]School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

[2]Centre for Primary Care and Public Health, Queen Mary University of London, United Kingdom

Correspondence

*Jessica Kasza, Monash University, 553 St Kilda Road, Melbourne, Victoria 3004, Australia. Email: jessica.kasza@monash.edu

Summary

Stepped wedge designs are an increasingly popular variant of longitudinal cluster randomized trial designs, and roll out interventions across clusters in a randomized, but step-wise fashion. In the standard stepped wedge design, assumptions regarding the effect of time on outcomes may require that all clusters start and end trial participation at the same time. This would require ethics approvals and data collection procedures to be in place in all clusters before a stepped wedge trial can start in any cluster. Hence, although stepped wedge designs are useful for testing the impacts of many cluster-based interventions on outcomes, there can be lengthy delays before a trial can commence.

In this paper we introduce "batched" stepped wedge designs. Batched stepped wedge designs allow clusters to commence the study in batches, instead of all at once, allowing for staggered cluster recruitment. Like the stepped wedge, the batched stepped wedge rolls out the intervention to all clusters in a randomized and step-wise fashion: a series of self-contained stepped wedge designs. Provided that separate period effects are included for each batch, software for standard stepped wedge sample size calculations can be used. With this time parameterization, in many situations including when linear models are assumed, sample size calculations reduce to the setting of a single stepped wedge design with multiple clusters per sequence. In these situations sample size calculations will not depend on the delays between the commencement of batches. Hence, the power of batched stepped wedge designs is robust to unexpected delays between batches.

KEYWORDS:

cluster randomised trial; intracluster correlation; sample size calculation; within-cluster correlation structure

1 | INTRODUCTION

The stepped wedge cluster randomised trial design, where clusters are randomised to switch from a control to an intervention condition at different pre-specified time points, has found application in a wide variety of research areas (examples in Mdege et al.[1]). Figure 1 displays an example of a conventional stepped wedge design with four periods and three treatment sequences. The period lengths are typically of equal duration and define the times at which different clusters cross from the control to the intervention condition. Stepped wedge designs are useful when intervention conditions applied at the level of the cluster

cannot be removed once implemented, e.g. educational interventions, or when assessing changes in policy that will be rolled out across systems. A crucial advantage of stepped wedge trials is that they may require fewer clusters and smaller total sample sizes than standard cluster randomised trials, due to the within-cluster comparisons enabled by the stepped wedge design[2]. It is well-recognised that the grouping of participants in clusters must be accounted for in sample size calculations and analysis of data from stepped wedge trials. In addition, due to the dependence between time and treatment in the stepped wedge, it is also essential to account for time in these calculations and analyses[3].

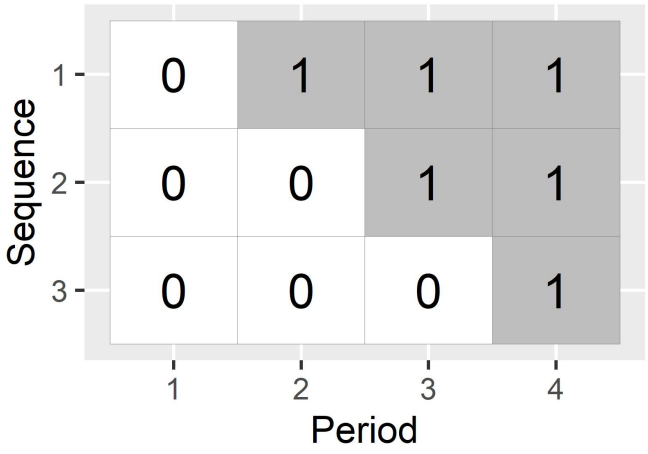

FIGURE 1 An example of a standard stepped wedge design, with 4 periods and 3 sequences (0 indicates periods in which the control condition is implemented; 1 indicates periods in which the intervention is implemented).

In the most commonly-used sample size formulas and statistical models for the design and analysis of stepped wedge trials (e.g[4,5,6]) it is assumed that the effect of "time" on outcomes is identical across clusters, and that time is divided up into distinct trial periods. If clusters commence study participation at different times (i.e. not all on the same date), then a distinction must be drawn between "calendar time" and "time-on-trial" (the amount of time since a cluster commenced trial participation). This distinction is particularly important for stepped wedge designs, where time and treatment are confounded. When clusters commence participation in a trial at the same calendar time, then calendar time and time-on-trial will be aligned: this is the case in Figure 1, where all clusters commence the study at the same time. When clusters are not aligned in calendar time, researchers must consider the distinction between calendar time and time-on trial, and parameterize time to align with their assumptions about the effect of time on outcomes in their statistical models. Three different time parameterizations that could be chosen when clusters are not aligned in calendar time are displayed in Figure 2: calendar time effects could be shared across clusters; time-on-trial effects could be shared across clusters; or separate period effects could be assumed in each batch. If clusters are not aligned in calendar time, but a standard stepped wedge sample size formula is applied, the assumption is that time-on-trial has an identical impact across all clusters, and there is no impact of calendar time (corresponding to time-on-trial effects that are shared across batches as in the middle panel of Figure 2). Further, when the standard sample size formulas are applied, it is required that there are no systematic differences between the batches of clusters that commence trial participation at different time points.

In practice, the great majority of cluster randomised stepped wedge trials have been designed so that all the clusters commence their participation at the same calendar time. This is likely to be for two reasons: firstly because clusters may all have expressed an interest in collaborating from an early stage in the development of the trial, and are all ready to go when the trial begins, but secondly, perhaps, because of concerns about the most appropriate way to model calendar time versus time-on-trial, and the lack of methodological guidance. There may be situations where it is to a triallist's advantage to stagger the commencement of different clusters.

In this paper we formalize the situation where different groups of clusters commence trial participation at different calendar times, defining the "batched stepped wedge design". In the batched stepped wedge design, different groups of clusters commence participation in a stepped wedge trial at different times, in a "batched" structure. The models we consider allow for systematic

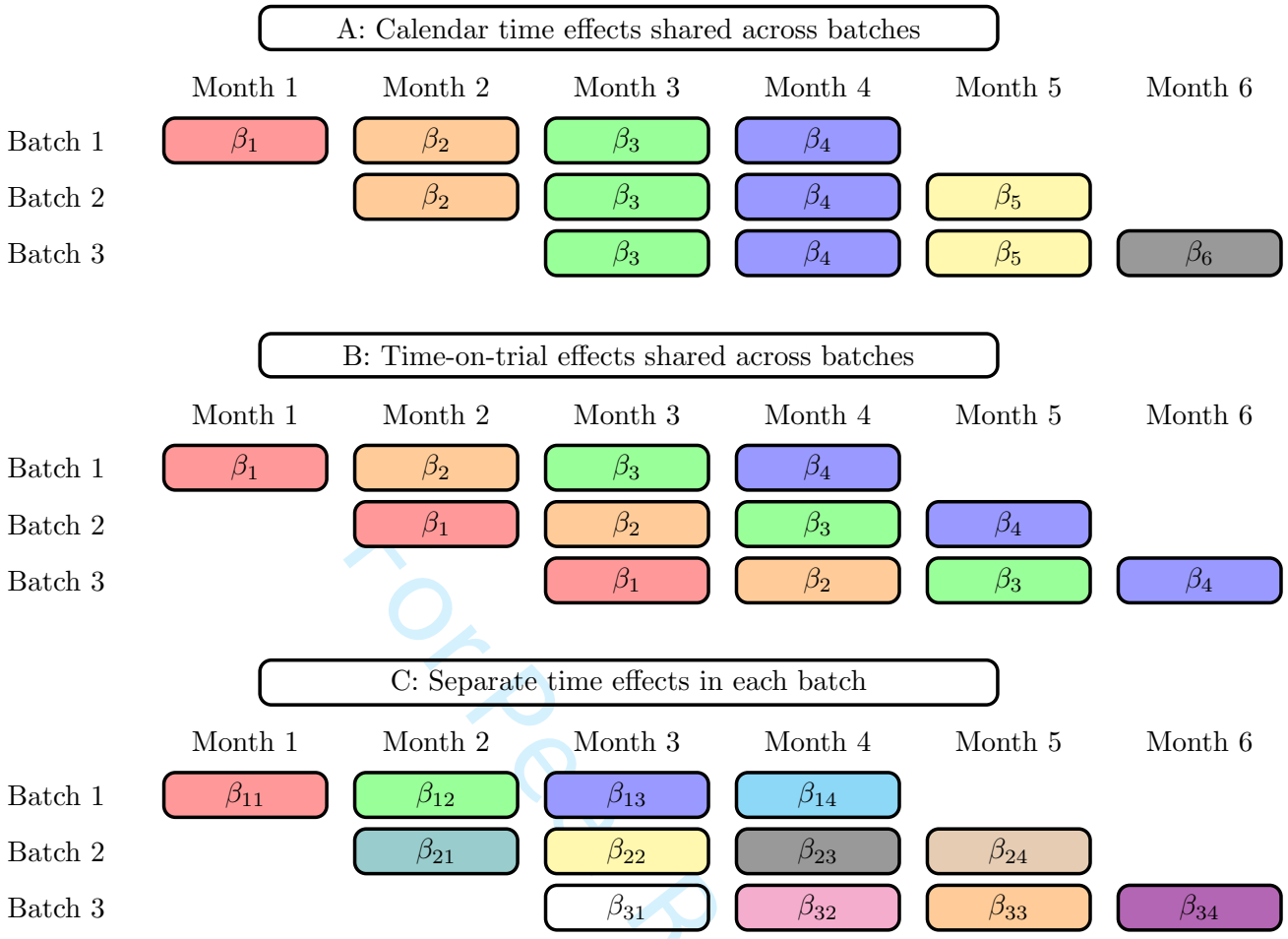

FIGURE 2 Three different ways in which the effect of time can be parameterised in a design where clusters commence study participation in three batches: the βs parameterise the time effects. For example, in the top panel, β_1 parametrises the effect of month 1 on outcomes. The top panel indicates how effects of time are shared across batches when the effects of calendar time are assumed to be constant across batches; the middle panel indicates how the effects of time are shared across batches when time-on-trial effects are assumed to be constant across batches; the bottom panel indicates that no time effects are shared across batches when separate time effects are estimated in each batch.

differences between clusters that commence study participation at different time points, and for differences in the effects of calendar and time-on-trial across these batches. This batched stepped wedge design is an alternative to the standard stepped wedge design: the batched design shares some of the benefits of the stepped wedge but allows for randomization of clusters to stepped wedge trial sequences in batches or blocks. Like the standard stepped wedge, the batched stepped wedge design ensures that all clusters eventually receive the intervention; treatment switches are unidirectional (i.e. the intervention is never removed once implemented); and the intervention is rolled out to each cluster in a randomized order. Examples when all batches are identical are shown in Figure 3; examples when batches differ are shown in Figure 4. This batched stepped wedge design can thus be conducted similarly to standard cluster randomized trials, where clusters may be randomised to the control or the intervention condition in groups as clusters are recruited to the trial, rather than all at once.

Although guidance for researchers and statisticians in sample size and power calculations for batched stepped wedge and related designs is lacking, researchers have already sought to implement designs similar to batched stepped wedge designs. For example, in a study assessing the impact of a mobility program for patients aged 60+ years across 8 veterans affairs hospitals in the USA on discharge destination of patients, Hastings et al.[7] sought to implement a batched stepped wedge design, with two batches of clusters. In reference to this batched structure, Hastings et al.[7] stated, "The full implications of a blocked randomization from a statistical perspective require further study". Similarly, the EAGLE study[8], investigating the impact of a quality

improvement intervention on the reduction of anastomotic leak following right colectomy, is randomizing hospitals in batches to a series of dog-leg designs (the dog-leg can be considered as an incomplete three-sequence stepped wedge design)[9].

The batched stepped wedge design may be appealing to researchers for three reasons: (1) depending on assumptions made about the outcome regression model, the power of a batched stepped wedge design will be unaffected by delays to the commencement of subsequent batches; (2) it can allow trials to get started sooner, by allowing clusters to come on-line in batches (i.e. ethics approvals and data collection procedures can be rolled out across clusters after the study has commenced); and (3) it does not require data to be collected by all clusters in all periods (i.e. this design can be thought of as an "incomplete" stepped wedge design. Expanding on the first reason: in this paper we show that for linear mixed models, if separate period effects are included for each batch of a batched stepped wedge (as in the bottom panel of Figure 2), then the power of the batched stepped wedge design is equivalent to the power of a standard stepped wedge design with multiple clusters assigned to each sequence. We also show that this will also hold when further assumptions about the effect of time are made when binary outcomes are analysed using non-linear link functions and generalised estimating equations. That is, under these assumptions, the designs in Figure 3 would have equivalent power to detect a difference as the design in Figure 1 with 3 clusters per sequence - although we would encourage trialists to include more than 9 clusters in any stepped wedge trial[10].

In this paper we provide researchers with guidance regarding the statistical aspects of batched stepped wedge designs, making recommendations regarding the inclusion of batch and period effects in the outcome regression model. In Section 2 we describe the batched stepped wedge design; in Section 3 we consider sample size calculations for the batched stepped wedge assuming linear mixed models for outcomes; in Section 4 we consider binary outcomes modeled with generalized linear models fit via generalized estimating equations. In Sections 3 and 4 we discuss what assumptions must be made in the specification of the outcome regression model to ensure the robustness property of the batched stepped wedge (where study power is robust to delays in the recruitment and/or commencement of the next batch) will hold. In Section 5 we discuss under what conditions standard stepped wedge sample size software can be applied to batched stepped wedge designs and demonstrate this calculation via an example. In Section 6 we present the results of a simulation study, and conclude with a discussion of our results in Section 7.

2 | WHAT IS A BATCHED STEPPED WEDGE DESIGN?

Simply put, a batched stepped wedge design is a series of stepped wedge cluster randomised trials. There may be some overlap in time between the successive stepped wedge components of the batched stepped wedge design, i.e. some trial periods during which data is being collected from more than one batch of the design. The component stepped wedge trials may be identical (as in the batched stepped wedge designs in Figure 3), or they may differ (as in Figure 4). Different sets of clusters contribute data in different batches of the study, and within each batch, clusters are randomised to the different sequences of the component stepped wedge design. A batched stepped wedge design allows for the recruitment of clusters throughout the duration of a study: once enough clusters for one of the component stepped wedges have been recruited, these clusters can be randomised to the sequences of the next batch, and the next stepped wedge component can commence. The models that we propose account for systematic differences between the clusters in different batches.

In a standard stepped wedge design, the implicit assumption is that all clusters commence participation in the trial at the same time[11] (or that there are assumptions made regarding the effect of time-on-trial as discussed in the Introduction). This is in contrast to the way in which parallel, or standard, cluster randomised trials are conducted. When parallel cluster randomised trials are conducted, clusters are often recruited throughout the duration of the trial. As is well known, in the (unstratified) parallel cluster randomised trial, so long as equal numbers of clusters (with equal numbers of participants) are assigned to the control and intervention arms at each randomisation point, this successive recruitment has no impact on the power of the study. This also holds for cluster randomised crossover designs provided equal numbers of clusters with equal numbers of participants implement the control and the intervention arm at each time point (again, this observation is limited to unstratified designs). The reason for this is that for these parallel and cluster randomised trial designs, treatment condition and time are independent: at each time point of the study, half of the clusters and participants will be in the control condition, and the other half will be in the intervention. In the stepped wedge design, treatment and time are not independent: the proportion of clusters in the intervention condition increases as time passes[2]. Depending on how time and randomisation batch are accounted for in the outcome regression model used to inform sample size calculations, the batched randomisation could have an impact on study power for batched stepped wedge designs, due to the confounding of time and treatment.

It is now well-recognised that time/period effects need to be accounted for in sample size and power calculations for stepped wedge designs[3]. Time/period effects must similarly be accounted for in sample size calculations for batched stepped wedge designs; researchers must provide adequate justification if they do not account for period effects in this calculation. Further, clusters that are included in different batches of the design may differ from each other, and thus it is recommended that batch effects be included in the outcome regression model. When there is an overlap between batches (e.g. the middle and bottom panels of Figure 3; the top panel of Figure 4), we also recommend that separate fixed period effects be included for each batch (equivalent to fixed batch-by-period interaction terms being included in the model). There are three key reasons for this recommendation: the first is that it requires making the fewest assumptions about the effects of time on outcomes and whether these effects are shared across batches; the second is that under this assumption, the variance of the treatment effect estimator for the batched stepped wedge is a combination of the variance for each component stepped wedge; the third (and most important) is that under this assumption, in many situations, study power will be robust to delays in the commencement of batches

We now consider the variance of the treatment effect estimator for batched stepped wedge designs. We first consider linear mixed models in Section 3, discussing batched stepped wedge designs with identical and non-identical components separately, and then discuss generalized linear models fit via GEE in Section 4.

3 | BATCHED STEPPED WEDGE DESIGNS AND LINEAR MIXED MODELS

3.1 | Batched stepped wedges with identical components

We first suppose that the batched stepped wedge design being considered is composed of B batches of identical T-period and K-sequence designs (for the standard stepped wedge $K = T - 1$), and initially suppose that one cluster is assigned to each sequence of each batch. We consider the following linear mixed model for the outcome Y_{bkti} from participant $i = 1, \dots, m$ in period $t = 1, \dots, T$ from cluster $k = 1, \dots, K$ in batch $b = 1, \dots, B$:

$$Y_{bkti} = \beta_{bt} + \theta X_{bkt} + \alpha_{bkt} + \epsilon_{bkti}, \quad \epsilon_{bkti} \sim N(0, \sigma^2_\epsilon). \tag{1}$$

In this model we have numbered the periods within each batch separately and thus period is identical to time-on-trial: Y_{11Ti} represents the outcome for the ith participant in cluster 1 in the final period (period T) of the first batch. If there is one period of overlap between successive batches, period T of batch 1 would correspond to period 1 of batch 2. Our model set-up automatically allows for separate period effects in each batch through the inclusion of the β_{bt} period terms (i.e. the scenario in the bottom panel of Figure 2): there are $B \times T$ period terms in total. These fixed period effects could alternatively be parameterised as period effects (where the effect for each period is shared by all clusters contributing data in that period, no matter their batch), batch effects, and terms for the period-by-batch interaction. This would require either constraining some of the β_{bt} to be identical, or re-numbering period from 1 to the total number of periods in the entire study (e.g. in Design 1 of Figure 3, the period subscript would range from 1 to 12; in Design 2 of Figure 3, the period subscript would range from 1 to 10). Given that later in this paper we recommend that separate period effects be included for each batch, throughout this paper we will number period within each batch (conceiving of time as time-on-trial; although this distinction from calendar time is immaterial when including an interaction with the batch term).

The treatment effect of interest is θ, assumed to be constant across batches, and the treatment group of cluster k in batch b at time period t is indicated by the binary variable X_{bkt}. The T-length vector of random effects $\alpha_{bk} = \left(\alpha_{bk1}, \dots, \alpha_{bkT}\right)^T$ for cluster k in batch b is assumed to have a multivariate normal distribution, centered around zero, with a variance matrix such that $var(\alpha_{bkt}) = \sigma^2_\alpha$ and $cov(\alpha_{bkt}, \alpha_{bks}) = r_{ts}\sigma^2_\alpha$, with $0 \leq r_{ts} \leq 1$. If $r_{ts} = r^{|t-s|}$ for some $0 < r < 1$, the discrete-time decay model of[12] is returned; if $r_{ts} = r$ for some $0 < r \leq 1$, the nested exchangeable model is returned, with $r = 1$ corresponding to the Hussey and Hughes model[4].

It is mathematically convenient to collapse Model 1 to cluster-period means when investigating the statistical power of designs[13]:

$$Y_{bkt} = \frac{1}{m}\sum_{i=1}^{m} Y_{bkti} = \beta_{bt} + \theta X_{bkt} + \alpha_{bkt} + \epsilon_{bkt}, \quad \epsilon_{bkt} \sim N\left(0, \frac{\sigma^2_\epsilon}{m}\right). \tag{2}$$

In the following result we consider the variance of the treatment effect estimator for models of the form given in Equations 1 and 2.

Result 1. Suppose that each batch of the batched stepped wedge design is identical, and models of the form in Equations 1 and 2 are considered, so that the cluster-period means from each cluster share a common variance matrix, denoted by V. V is a $T \times T$ matrix, with the (t, s) element given by $cov(Y_{bkt}, Y_{bks})$. If X_{bk} is the $T \times 1$ vector containing the treatment indicators of cluster k in batch b for all T periods, then $X_{bk} = X_{b'k} = X_k$ for all pairs of batches b and b', and the variance of the treatment effect estimator $\hat{\theta}$ is given by:

$$var(\hat{\theta}) = \frac{1}{B}\left[\sum_{k=1}^{K} X_k^T V^{-1} X_k - \frac{1}{K}\left(\sum_{k=1}^{K} X_k^T V^{-1} \sum_{k=1}^{K} X_k\right)\right]^{-1} = \frac{1}{B} var_0(\hat{\theta}), \tag{3}$$

where $var_0(\hat{\theta})$ is the variance of the treatment effect estimator for one of the components of the batched design with one cluster per sequence. This result can be generalised to the situation where C_b clusters are assigned to each sequence of batch b. When this is the case,

$$var(\hat{\theta}) = \frac{1}{\sum_{b=1}^{B} C_b} var_0(\hat{\theta}). \tag{4}$$

Result 1 indicates that when batch-by-time interaction terms are included in the model for the outcome, the treatment effect estimator is simply a weighted combination of treatment effect estimators obtained from each batch separately. Specifically, the estimator from each batch is weighted by its variance.

Result 1 is a consequence of the more general result discussed in Section 3.2, with the proof provided in Section 1 of the Supplementary Material available online. Equation 3 indicates that the variance of the treatment effect estimator from the batched stepped wedge design with B batches of identical T-period stepped wedge designs on K clusters is equivalent to the variance of the treatment effect estimator for a single T-period stepped wedge design with $B \times K$ clusters. When C_b clusters are assigned to each sequence of batch b, the variance of the treatment effect estimator for the batched design is equivalent to that of the single component design with $\sum_{b=1}^{B} C_b$ clusters per sequence. When all batches are identical and a model such as that in Equation 1 is assumed, sample size calculations for batched stepped wedge designs are thus straightforward. We demonstrate such calculations in Section 5.

Model 1 can be extended to allow for closed or open cohort schemes (as described in Kasza et al.[14], for example), to incorporate treatment effect heterogeneity (as described in Kasza et al.[15], for example), and to allow for differing numbers of subjects in each cluster in each period (as described in Kasza at al.[16], for example). When treatment effect heterogeneity is included in the model, the variance matrix V will not be identical across the clusters within a batch. However, the variance of the treatment effect estimator for a batch (denoted by $var_0(\hat{\theta})$ in Equation 3) will be common across batches. Hence when treatment effect heterogeneity is included in the model, the variance of the treatment effect estimator from the batched stepped wedge design with B batches of identical component designs with C_b clusters per sequence in batch b is again equivalent to the variance of the treatment effect estimator for a single component with $\sum_{b=1}^{B} C_b$ clusters per sequence.

When different clusters have different numbers of participants in each cluster-period, there may not be a common joint variance matrix $var_0(\hat{\theta})$ across batches. When cluster sizes differ, but each cluster is expected to collect the same number of observations in each of their data collection periods (i.e. cluster k collects m_k observations in each period), researchers could calculate the mean and coefficient of variation of cluster sizes and use the approximation presented in[17] to obtain a common $var_0(\hat{\theta})$ for each batch of the design.

3.2 | Batched stepped wedges with non-identical components

We now consider batched stepped wedge designs with non-identical components (examples in Figure 4): we suppose that there are B batches of stepped wedge designs, where batch b is a T_b-period stepped wedge design, with K_b clusters. Result 2 provides the variance of the treatment effect estimator when batches are no longer identical.

Result 2. If Y_{bkti} is the outcome for participant $i = 1, \ldots, m_{bkt}$ in period $t = 1, \ldots, T_b$ in cluster $k = 1, \ldots, K_b$ in batch $b = 1, \ldots, B$, and Y_b is the $M_b = \sum_{k=1}^{K_b} \sum_{t=1}^{T_b} m_{bkt}$-length vector of outcomes from all clusters in batch b, we suppose that

$$Y_b \sim N(Z_b\gamma_b + \theta X_b, \Sigma_b)$$

where γ_b is the T_b-length vector of period effects for batch b, Z_b is the design matrix associated with these period effects for cluster b (of dimension $M_b \times T_b$), θ is the treatment effect of interest (assumed to be shared across all batches), X_b is the M_b-length vector indicating if a participant is in a cluster-period in the control condition ($X_{bkti} = 0$) or the intervention condition

($X_{bkti} = 1$), and Σ_b is the $M_b \times M_b$ covariance matrix of the outcomes from all clusters in batch b. Then if $\hat{\theta}$ is the generalised least squares estimator of θ,

$$var(\hat{\theta}) = \left(\sum_{b=1}^{B} \frac{1}{var_b(\hat{\theta})} \right)^{-1}, \tag{5}$$

where $var_b(\hat{\theta})$ is the variance of the generalised least squares estimator of θ obtained by considering batch b only. Further, if $var_b(\hat{\theta}) = var_0(\hat{\theta})$ then $var(\hat{\theta}) = \frac{1}{B} var_0(\hat{\theta})$.

The proof of Result 2 is provided in Section 1 of the Supplementary Material. This result assumes normally distributed outcomes where only the treatment effect is shared across batches; no assumptions are made regarding the equality of within-cluster correlation structures within or between batches. However, if $var_b(\hat{\theta}) = var_0(\hat{\theta})$ for some $var_0(\hat{\theta})$ for all $b = 1, \ldots, B$ then Equation 5 collapses to the result given in Result 1, i.e. the situation where all batches are identical. Once again, the treatment effect estimator is the weighted sum of the estimators obtained for each batch, with each batch's estimator weighted by its variance.

4 | BINARY AND COUNT OUTCOMES AND BATCHED STEPPED WEDGES

Several modelling options are available when researchers are interested in binary, rather than continuous outcomes. One option, discussed in Hussey and Hughes[4], and taken in Hemming et al.[5], is to apply Equation 1 to binary outcomes, setting σ^2_ϵ equal to $p(1-p)$, where $p = P(Y_{bkti} = 1)$. The generalized least squares estimator of θ is then considered, and results in Section 3 apply. However, it has been pointed out that the variance of the treatment effect estimator may not be adequately approximated when this approach is applied[18]. Zhou et al.[18] developed an alternative approach, assuming a truncated normal distribution for cluster random effects, to ensure that estimated probabilities lie between 0 and 1. When batch and period are parameterised as in Equation 1 (i.e. separate period effects are included for each batch in the outcome regression model), the variance of the treatment effect estimator can be written as the sum of variances for each stepped wedge component as for the continuous outcome.

When binary or count outcomes are of interest, researchers are frequently interested in estimating a marginal treatment effect instead of the conditional treatment effect. The use of generalized estimating equations (GEE) for the analysis of longitudinal cluster randomized trials allows for estimation of such marginal effects, and implications of this analysis approach for sample size calculations have previously been discussed[6,19]. When the GEE approach is used, a working correlation matrix structure must be assumed. This working correlation structure describes the pattern of within-cluster correlations; an exchangeable correlation structure would imply equal correlations between all observations in a cluster, for example. As discussed in Li et al.[6], when GEE is the intended analysis approach, power calculations can proceed via generalized least squares. We now state the main result for this scenario.

Result 3. If Y_{bkti} is the outcome for participant $i = 1, \ldots, m_{bkt}$ in period $t = 1, \ldots, T_b$ in cluster $k = 1, \ldots, K_b$ in batch $b = 1, \ldots, B$, with $\mu_{bkti} = E[Y_{bkti}]$, we assume

$$g(\mu_{bkti}) = \beta_{bt} + \theta X_{bkt}. \tag{6}$$

where g is the link function, β_{bt} is the fixed effect for period t in batch b, θ is the treatment effect of interest, and X_{bkt} is the indicator for whether cluster k in batch b and period t is in the intervention or control condition. Let μ_b be the vector of means for batch b. If $\beta_b = (\beta_{b1}, \ldots, \beta_{bT_b})^T$ is the set of time effects for batch b, $\hat{\beta}_b$ is the generalised least squares estimator of β_b, and $\hat{\theta}$ is the generalised least squares estimator of θ then

$$\begin{aligned} var(\hat{\theta}) &= \left(\sum_{b=1}^{B} \frac{\partial \mu_b}{\partial \hat{\theta}}^T W_b^{-1} \frac{\partial \mu_b}{\partial \hat{\theta}} - \frac{\partial \mu_b}{\partial \hat{\theta}}^T W_b^{-1} \frac{\partial \mu_b}{\partial \hat{\beta}_b} \left[\frac{\partial \mu_b}{\partial \hat{\beta}_b}^T W_b^{-1} \frac{\partial \mu_b}{\partial \hat{\beta}_b} \right]^{-1} \frac{\partial \mu_b}{\partial \hat{\beta}_b}^T W_b^{-1} \frac{\partial \mu_b}{\partial \hat{\theta}} \right)^{-1} \\ &= \left(\sum_{b=1}^{B} \frac{1}{var_b(\hat{\theta})} \right)^{-1} \end{aligned} \tag{7}$$

where $var_b(\hat{\theta})$ is the variance of the treatment effect estimator obtained via GEE when batch b is considered separately, and W_b is the covariance matrix of the observations from batch b. W_b has the form $A_b^{1/2} R_b A_b^{1/2}$. A_b is a diagonal matrix with elements

given by $var(Y_{bkti})$ and R_b is the assumed correlation matrix of the observations from batch b. If a binomial distribution for outcomes is assumed $\mu_{bkti} = P(Y_{bkti} = 1)$ and diagonal elements of A_b will be given by $var(Y_{bkti}) = \mu_{bkti}(1 - \mu_{bkti})$.

The proof of this result is shown in the Appendix. As was the case in the linear mixed model scenario, the treatment effect estimator is again a weighted sum of each batch's treatment effect estimator. However, in contrast to the linear model discussed in Section 3, the variance of the treatment effect estimator depends on the assumed period effects through μ_{bkti} in the model in Equation 6. For example, when calculating sample sizes for a batched stepped wedge trial where a marginal model will be used to analyse binary outcomes, researchers must include predicted prevalences of the outcome in each period of the trial in sample size calculations. For binary outcomes that will be analysed in this way, the variance of the batched stepped wedge design will collapse to the simplified form given in Equation 4 only if no period effects are included in the model. This is a strict assumption, requiring the prevalence of the outcome in the control arm to be the same for the entire trial duration and across batches (i.e. there are no secular time effects and no batch effects).

Result 3 applies not only to binary outcomes analysed with a logit link function; the proof given in the Appendix does not rely on the choice of link function or the outcome type. Thus, Result 3 holds for binary outcomes with a linear link function, or for count outcomes with a log link function, to name just two alternatives.

5 | DEMONSTRATION OF POWER CALCULATIONS FOR BATCHED STEPPED WEDGE DESIGNS

We demonstrate how to calculate power for a batched stepped wedge design with two identical batches. The example we consider is from Unni et al.[20]: in that paper, various different stepped wedge-like designs were considered for the Patient-Centered Care Transitions in Heart Failure Trial (PACT-HF), including a batched design with "early" and "late" blocks, shown in Figure 5. Each of these blocks was a 5-sequence, 6-period stepped wedge design, with one cluster assigned to each sequence, with 54 patients in each cluster in each of these periods. The primary outcome considered was a binary outcome, that was a composite of a number of clinical outcomes, with a prevalence of 28% under the control condition, and an intracluster correlation of 0.01. We assume that this intracluster correlation was conditional on the inclusion of a "batch" term in the model (a point we return to in the Discussion), but in practice we would recommend assessing the impact of varying this correlation on study power. The aim was to detect a 25% reduction in the prevalence of the outcome: that is, a reduction from 28% to 21%. The effect on the logit link scale is -0.38.

We consider two methods for calculating the power of this design: first, we assume a linear model for the binary outcome (applying the results of Section 3.1); second, we assume that a generalised estimating equations approach will be taken to fitting a model with a logit link (applying the results of Section 4).

To perform the power calculation for the first approach, one need only calculate the power of a standard 6-period stepped wedge design with 2 clusters assigned to each sequence. The Shiny CRT calculator[5] accommodates this by allowing users to set the number of clusters assigned to each sequence; the Stata `steppedwedge` program[21] accommodates this through the "k" option. When the linear model is assumed, this study has a power of 77% to detect the difference. There are two additional ways to use the Shiny CRT calculator to calculate the power of the batched design. The user could get the precision of each of the component designs separately using the the "Precision" tab on the Shiny CRT calculator, and then combine these according to Result 1. Alternatively, the user could upload the design matrix for the batched design (ensuring that there is no overlap between successive batches) and obtain the power of the design directly. Were a design matrix uploaded with an overlap between successive batches, the Shiny CRT calculator would assume that batches with overlapping periods share period effects (that is, calendar time effects would be shared across batches as in the top panel of Figure 2).

For the second method we use the R `swdpwr` package[22] to calculate the power of the batched stepped wedge design. Since the GEE approach depends on the baseline prevalence of the outcome, we consider two different scenarios:

1. The prevalence of the outcome under the control condition remains at 28% for the entire duration of the trial.
2. The prevalence of the outcome under the control condition is initially 30%, but decreases to 28% by the final period of the trial, in a linear fashion. That is, at the time that the second batch starts data collection, the prevalence of the outcome under the control condition is 29%.

Since there is no change in the underlying prevalence of the condition over time in the first scenario, the power of the batched stepped wedge using the GEE approach is equivalent to the power of the 6-period stepped wedge with 2 clusters assigned to

each sequence. The power of this design can be obtained directly by using the `swdpower` command in the `swdpwr` package, and is 98.8%. Power is high due to the omission of period effects in this calculation.

For the second scenario, the variance of the treatment effect estimator must be obtained for each of the two component designs separately. We assume that the treatment effect is -0.38 on the logit link scale for both batches, with the aim to detect a reduction from 29% to 21.75% for the first batch, and from 28% to 21% in the second batch. The `swdpower` command cannot provide the power of the batched design directly. However, the variance of the treatment effect estimator can be obtained for each batch separately from the power calculated by the command. These variances are then combined using Result 3. The power of the batched design to detect a change from 28% to 21% is 80.8%. Commands to replicate this calculation are provided in Section 2 of the Supplementary Material.

6 | SIMULATION STUDY

We conducted a simulation study to verify our theoretical results, inspired by the PACT-HF design discussed in Section 5. As in the design schematic in Figure 5, we consider a design consisting of two batches, each a 6-period stepped wedge design. However, we vary the number of overlapping periods between the two batches from 0 (indicating no overlap between the two batches, as in Figure 5) to 5 (batches that overlap completely); the key aim of this simulation study is to assess whether inclusion of separate period effects for each batch has an impact on empirical power. Does power decrease as the number of batch-by-time terms in the model increases? In the simulation we increase the total number of clusters to 40 (4 clusters assigned to each of the 10 sequences). We simulate both binary and continuous outcomes for a range of correlation parameter values. Code to replicate this simulation study and the nested loop plots is available at `https://github.com/jkasza/BatchSW`.

Table 1 lists the parameters considered for the simulation study for the continuous outcomes. Along with varying the number of periods of overlap between successive batches, the intracluster correlation and the cluster autocorrelation, datasets were simulated with an effect size of 0 (to allow an examination of significance level) and 0.15. For each combination of parameters in Table 1, 1000 datasets were simulated, with separate time effects in each batch (as in the bottom panel of Figure 2). The period effects for each batch in each period were simulated from a normal distribution with mean 0 and variance 1. With 1000 simulated datasets, the Monte Carlo standard error associated with a power of 80% is expected to be around $\pm 1.3\%$[23]. Each simulated dataset was analysed using a linear mixed-effects model with random effects for cluster and cluster-period, and separate categorical fixed period effects for each batch (i.e. period effects, batch effects, and period by batch interaction terms). Our focus here is on the comparison of theoretical and simulated power, so for each set of parameters, we calculated the percentage of hypothesis tests $H_0 : \theta = 0$ rejected at the two-sided 5% significance level.

TABLE 1 The continuous outcome simulation settings. 1000 datasets were simulated for each combination of parameters (108 combinations).

Parameter	Meaning	Values
T	Number of periods in each stepped wedge design	6
B	Number of batches	2
K	Number of clusters assigned to each sequence	4
m	Number of observations in each cluster in each sequence	10
N_O	Number of periods of overlap between successive batches	5, 4, 3, 2, 1, 0
ρ	Intra-cluster correlation	0.01, 0.05, 0.1
r	Cluster autocorrelation	1, 0.95, 0.75
θ	Effect size	0, 0.15

Figure 6 displays the empirical type I error rates and power for each set of parameters using nested loop plots[24]. This figure indicates that the number of periods of overlap does not have an impact on empirical type I error rates and power: as the number of periods of overlap changes, there is no pattern to the variation in empirical type I error rates or power. This provides support for our theoretical result, which indicates that if period, batch, and batch by period interaction terms are included in the model,

the degree of overlap has no impact on study power. As expected, Figure 6 does indicate that the intracluster correlation and cluster autocorrelation do have an impact on empirical power levels. It is interesting to note that for some combinations of the cluster autocorrelation and intracluster correlation (e.g. when the cluster autocorrelation is 0.75, and the intracluster correlation is equal to 0.05 or 0.1), the empirical power is slightly inflated. However, this does not change as the overlap between batches decreases. That is, empirical power does not decrease as the number of time effects included in the model increases.

Table 2 lists the parameters considered for the simulation study for the binary outcomes. As was the case for the simulation study for continuous outcomes, 1000 datasets were simulated for each combination of parameters. Binary data was simulated using the method of Qaqish[25], as coded by Li et al[19]. The range of intracluster correlations permitted by the simulation method of Qaqish is limited, so we only consider intracluster correlations of 0.01 and 0.05 for the binary outcomes. Each simulated dataset was analysed via GEE with a logit link with an exchangeable working correlation structure, with separate coefficients for period (treated as a continuous covariate) in each batch (this choice is to match the sample size calculation in the R `swdpwr` package[22]). Again, our focus is on the comparison of theoretical and simulated power, so for each set of parameters and each analysis choice, we calculated the percentage of hypothesis tests $H_0 : \theta = 0$ rejected at the two-sided 5% significance level. Theoretical power for each combination of parameters was also calculated, using the `swdpwr` package. Figure 7 displays the empirical type I error rates and power for each set of parameters for the binary outcomes analysed via GEE with an exchangeable working correlation. As for the simulation study for continuous outcomes, the simulated power and type I error rates do not depend on the degree of overlap between successive batches, aligning with our theoretical results.

TABLE 2 The binary outcome simulation settings. 1000 datasets were simulated for each combination of parameters (24 combinations).

Parameter	Meaning	Values
T	Number of periods in each stepped wedge design	6
B	Number of batches	2
K	Number of clusters assigned to each sequence	4
m	Number of observations in each cluster in each sequence	10
$P(Y_{bkt} = 1 \mid X_{bkt} = 0)$	Probability of the outcome in a non-treatment period	$0.4 + t \times 0.01$
N_O	Number of periods of overlap between successive batches	5, 4, 3, 2, 1, 0
ρ	Intra-cluster correlation	0.01, 0.05
r	Cluster autocorrelation	1
$P(Y_{bkt} = 1 \mid X_{bkt} = 1)$ $-P(Y_{kt} = 1 \mid X_{kt} = 0)$	Change in probability of the outcome caused by the intervention	0, 0.025

7 | DISCUSSION

The batched stepped wedge design is a promising alternative to the standard stepped wedge design. By allowing clusters to come on-line to the study in batches, the batched stepped wedge design has the potential to get started sooner than a standard stepped wedge, which typically requires all clusters to commence at the same point in time. If separate period effects are included for each batch (as in the bottom panel of Figure 2), the power of the batched stepped wedge design will, depending on the assumed outcome regression model, be robust to delays in the commencement of batches. This holds when linear models for the outcome are assumed, or when the prevalence of the outcome in the control condition is not expected to change over time. Hence, in these settings, study power will be unaffected if there is an unanticipated delay before the next batch commences study involvement when separate period effects are assumed for each batch. Under this assumed model, standard stepped wedge software can be used to calculate the required sample size and study power for batched stepped wedge designs.

Our key result indicates that in certain situations a batched stepped wedge design consisting of B identical stepped wedge designs provides the same power to detect an effect as one of the stepped wedge components with B clusters assigned to each sequence. However, the choice of variance components will have an impact on sample size and power calculations for all batched

stepped wedge designs. Inclusion of the batch term implies that variance components must now be treated as "within-batch" variance components, and will likely be smaller than if a model without batch effects was considered. When batch effects are included in the outcome regression model, variance components will be conditional on the inclusion of these batch effects in the model. Hence, researchers must consider the impact of batches on variance components and intracluster correlations when considering sample size and power.

Our key results have broad applicability. They generalise to batches of any other type of longitudinal cluster randomised trial design. and do not rely on the design type. For example, our results apply to a "batched dog-leg" design. Provided that separate period effects are included for each batch, the variance of such a design would have the form given in Sections 3 and 4: summing over the variances of treatment effect obtained for each of the individual component designs. Further, our key results do not depend on the precise form of the variance of the treatment effect estimator for each batch. The models considered in the Results could be extended to allow for closed or open cohorts, treatment effect heterogeneity, etc. The key result only requires that separate period effects are included in the outcome model for each batch: if this is the case, then the variances of the treatment effect estimators for each batch can be combined according to Results 1, 2, or 3 as appropriate.

We recommend that separate period effects are included for each stepped wedge batch (i.e. the time parameterization as in the bottom panel of Figure 2) to provide robustness to the sample size calculation in case of unexpected recruitment and set-up delays. In addition to allowing for robustness to delays and making the fewest assumptions about the effect of time on outcomes, assuming separate period effects across batches would be appropriate for trials where clusters in different batches are from geographically distinct areas, or where clusters in different batches are otherwise distinct. Additionally, if batch effects are not included in the outcome regression model and batches are assumed to have shared period effects for overlapping periods, then study power will depend on the separation between successive batches. If unexpected delays between batches occur, the power of the study will not be robust to this change, in that it will differ from that calculated a priori. Future work will investigate the impact of increasing degrees of overlap between successive batches when period effects are shared across batches.

Adaptive variants of the batched stepped wedge design are a logical next step of this work. Such adaptations may include sample size re-estimation, or more formal stopping rules for efficacy or futility of the intervention based on assessments at suitable time points, for example after participants in each batch have completed their followup. While adaptive variants of the stepped wedge design have been discussed in the literature, these do require that all clusters commence data collection at the same time. These adaptive variants will be explored in future work.

ACKNOWLEDGMENTS

This work was supported by the Australian Research Council Discovery Project DP210101398 and National Health and Medical Research Council of Australia Project Grant ID 1108283.

Conflict of interest

The authors declare no potential conflict of interests.

Data sharing

Results of simulations and the code to replicate the simulation study in Section 6 is available at `https://github.com/jkasza/BatchSW`.

References

1. Mdege ND, Man MS, Taylor CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *Journal of Clinical Epidemiology* 2011; 64(9): 936-948.

2. Matthews JNS, Forbes AB. Stepped wedge designs: insights from a design of experiments perspective. *Statistics in Medicine* 2017; 36(24): 3772-3790.

3. Hemming K, Taljaard M, McKenzie JE, et al. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. *BMJ* 2020; 363: k1614.

4. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials* 2007; 28: 182-191.

5. Hemming K, Kasza J, Hooper R, Forbes A, Taljaard M. A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator. *International Journal of Epidemiology* 2020; 49: 979-995.

6. Li F, Turner EL, Preisser JS. Sample size determination for GEE analyses of stepped wedge cluster randomized trials. *Biometrics* 2018; 74: 1450-1458.

7. Hastings SN, Stechuchak KM, Choate A, et al. Implementation of a stepped wedge cluster randomized trial to evaluate a hospital mobility program. *Trials* 2020; 21: 863.

8. ESCP EAGLE Safe Anastomosis Collective . ESCP Safe Anastomosis ProGramme in CoLorectal SurgEry (EAGLE): Study protocol for an international cluster randomised trial of a quality improvement intervention to reduce anastomotic leak following right colectomy. *Colorectal Disease* 2021: doi:10.1111/codi.15806.

9. Hooper R, Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *BMJ* 2015: 350.

10. Taljaard M, Teerenstra S, Ivers NM, Fergusson DA. Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. *Clinical Trials* 2016; 13: 459-463.

11. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford R. The stepped-wedge cluster randomised trial: rationale, design, analysis and reporting. *BMJ* 2015; 350: h391.

12. Kasza J, Hemming K, Hooper R, Matthews JNS, Forbes AB. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Statistical Methods in Medical Research* 2019; 28: 703-716.

13. Grantham KL, Kasza J, Heritier S, Hemming K, Forbes AB. Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. *Statistics in Medicine* 2019; 38: 1918-1934.

14. Kasza J, Hooper R, Copas A, Forbes AB. Sample size and power calculations for open cohort longitudinal cluster randomized trials. *Statistics in Medicine* 2020; 39: 1871-1883.

15. Kasza J, Taljaard M, Forbes AB. Information content of stepped-wedge designs when treatment effect heterogeneity and/or implementation periods are present. *Statistics in Medicine* 2019; 38: 4686-4701.

16. Kasza J, Bowden R, Forbes AB. Information content of stepped wedge designs with unequal cluster-period sizes in linear mixed models: Informing incomplete designs. *Statistics in Medicine* 2021: DOI: 10.1002/sim.8867.

17. Harrison LJ, Chen T, Wang R. Power calculation for cross-sectional stepped wedge cluster randomized trials with variable cluster sizes. *Biometrics* 2020; 76: 951-962.

18. Zhou X, Liao X, Kunz LM, Normand SLT, Wang M, Spiegelman D. A maximum likelihood approach to power calculations for stepped wedge designs of binary outcomes. *Biostatistics* 2020; 21: 102-121.

19. Li F, Forbes A, Turner EL, Preisser JS. Power and sample size requirements for GEE analyses of cluster randomized crossover trials. *Statistics in Medicine* 2019; 38: 636-649.

20. Unni RR, Lee SF, Thabane L, Connolly S, Van Spall HG. Variations in stepped-wedge cluster randomized trial design: Insights from the Patient-Centered Care Transitions in Heart Failure trial. *American Heart Journal* 2020; 220: 116-126.

21. Hemming K, Girling A. A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. *Stata Journal* 2014; 14: 363-380.

22. Chen J, Zhou X, Li F, Spiegelman D. swdpwr: A SAS macro and an R package for power calculation in stepped wedge cluster randomized trials. *ArXiV* 2020: arxiv:2011.06031v1.

23. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019; 38: 2074-2102.

24. Kammer M. *looplot: Create nested loop plots*. 2022. R package version 0.5.0.9002.

25. Qaqish BF. A family of multivariate binary distributions for simulating correlated binary variables. *Biometrika* 2003; 90: 455-463.

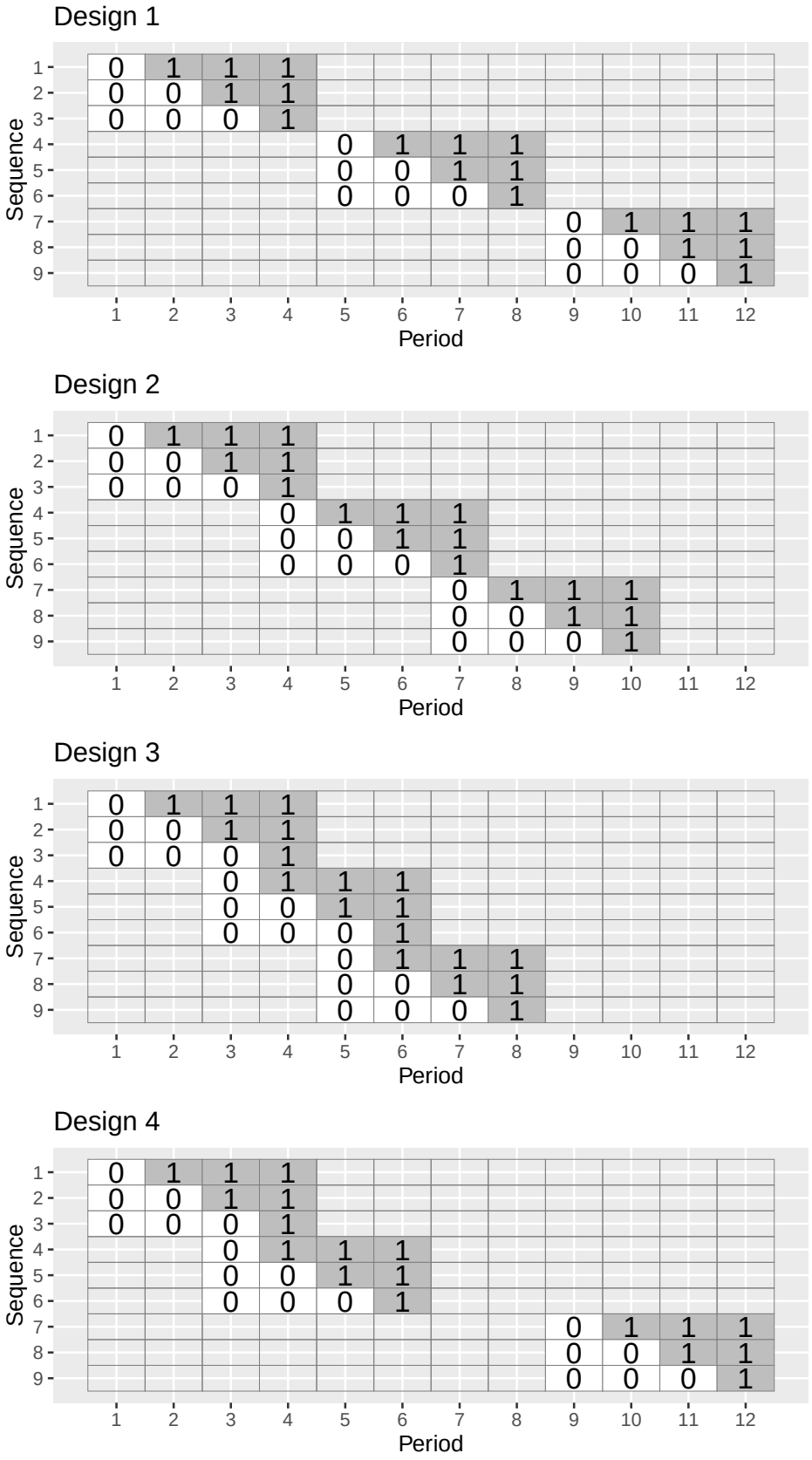

FIGURE 3 Four examples of batched stepped wedge designs with identical component designs (0 indicates control periods; 1 indicates intervention periods). Each of these designs has three batches of three-period stepped wedge designs, with differing degrees of overlap between successive batches. Design 1 (top row): no overlap between successive batches; Design 2 (second from top): overlap of one period between successive batches; Design 3 (second from bottom): overlap of two periods between successive batches; Design 4 (bottom row): variable overlap between successive batches.

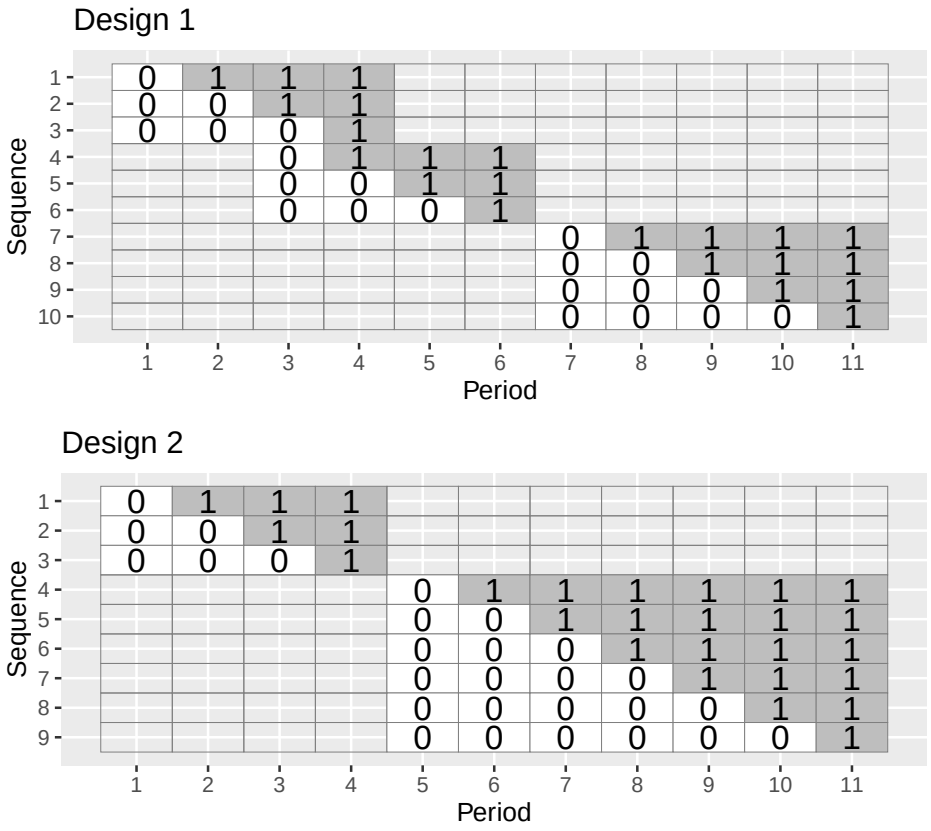

FIGURE 4 Two examples of batched stepped wedge designs without identical component designs (0 indicates control periods; 1 indicates intervention periods).

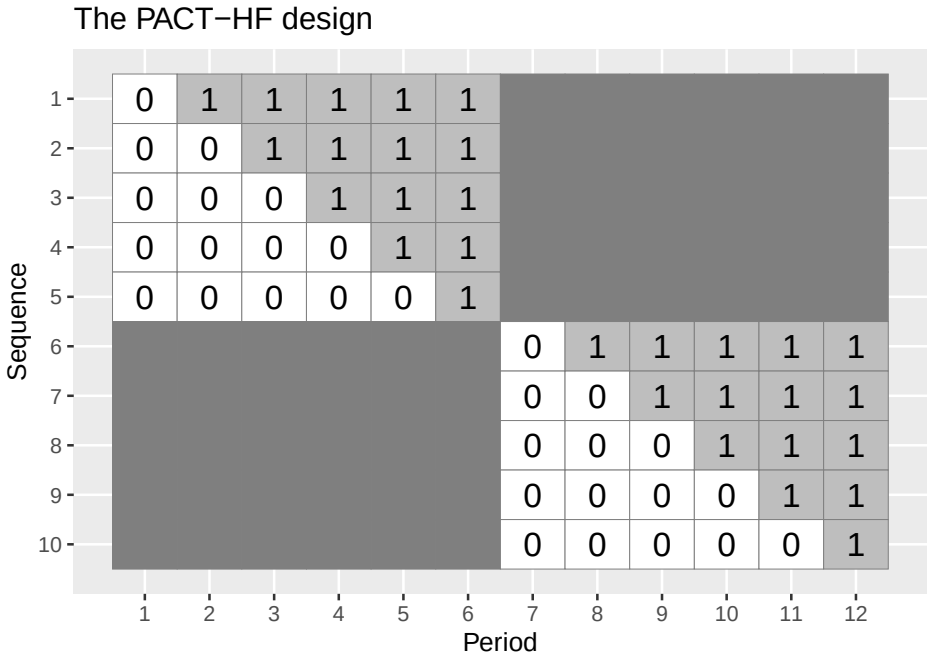

FIGURE 5 The design schematic for the PACT-HF trial: two batches of a 5-sequence stepped wedge design.

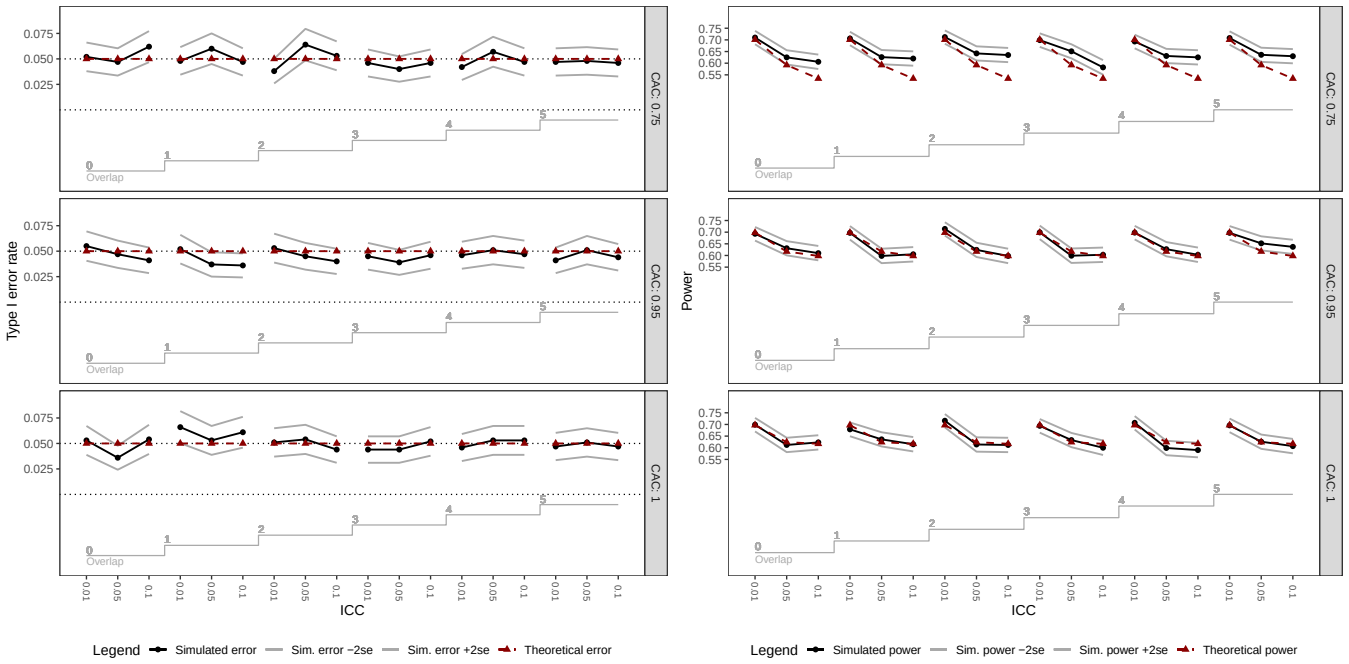

FIGURE 6 Empirical and theoretical type I error rates (left panel) and power (right panel) for the simulated continuous outcomes. ICC=intracluster correlation; CAC = cluster autocorrelation. Within each panel, sub-panels correspond to a different value of the CAC. The theoretical and empirical Type I error rate or power is displayed for each combination of number of periods of overlap, ICC, and CAC, with the empirical result plus and minus 2 standard errors also displayed.

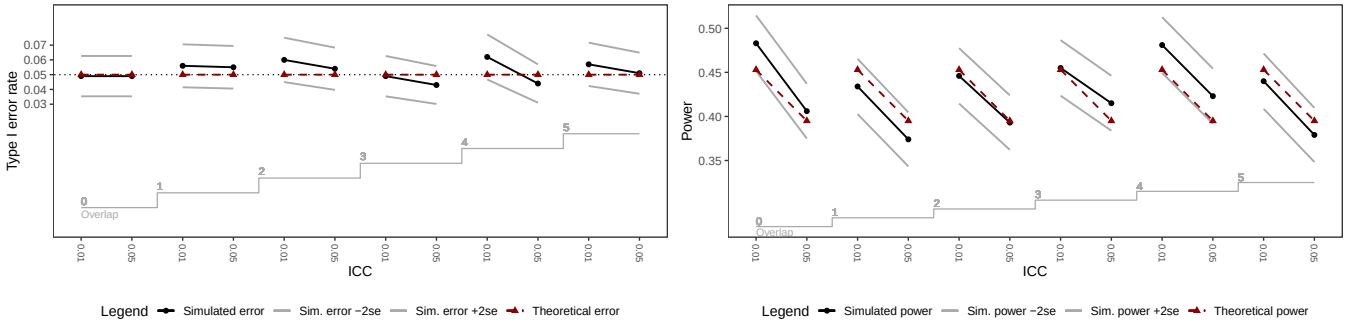

FIGURE 7 Empirical and theoretical type I error rates (left panel) and power (right panel) for the simulated binary outcomes analysed via GEE. ICC=intracluster correlation. The theoretical and empirical Type I error rate or power is displayed for each combination of number of periods of overlap and ICC, with the empirical result plus and minus 2 standard errors also displayed.

Appendix to "The batched stepped wedge design: a design robust to delays in cluster recruitment"

Jessica Kasza, Rhys Bowden, Richard Hooper, Andrew Forbes

jessica.kasza@monash.edu

School of Public Health and Preventive Medicine,

Monash University,

553 St Kilda Road, Melbourne 3004, Victoria, Australia

1 Proof of results

Result 1. *We consider the following linear mixed model for the outcome Y_{bkti} from participant $i = 1, \ldots, m$ in period $t = 1, \ldots, T$ from cluster $k = 1, \ldots, K$ in batch $b = 1, \ldots, B$:*

$$Y_{bkti} = \beta_{bt} + \theta X_{bkt} + \alpha_{bkt} + \epsilon_{bkti}, \quad \epsilon_{bkti} \sim N(0, \sigma^2_\epsilon). \tag{1}$$

The treatment effect of interest is θ, assumed to be constant across batches, and the treatment group of cluster k in batch b at time period t is indicated by the binary variable X_{bkt}. β_{bt} is the average outcome under the control condition in period t of batch b. The T-length vector of random effects $\alpha_{bk} = (\alpha_{bk1}, \ldots, \alpha_{bkT})^T$ for cluster k in batch b is assumed to have a multivariate normal distribution, centered around zero. We suppose that the vector of cluster-period means is a sufficient statistic for the treatment effect, and that the cluster-period means from each cluster share a common variance matrix, denoted by V. V is a $T \times T$ matrix, with the (t, s) element given by $cov(Y_{bkt}, Y_{bks})$.

Supposing that each batch of the batched stepped wedge design is identical, and X_{bk} is the $T \times 1$ vector containing the treatment indicators of cluster k in batch b for all T periods, then $X_{bk} = X_{b'k} = X_k$ for all pairs of batches b and b', the variance of the treatment effect estimator $\hat{\theta}$ is given by:

$$var(\hat{\theta}) = \frac{1}{B}\left[\sum_{k=1}^{K} X_k^T V^{-1} X_k - \frac{1}{K}\left(\sum_{k=1}^{K} X_k^T V^{-1} \sum_{k=1}^{K} X_k\right)\right]^{-1} = \frac{1}{B} var_0(\hat{\theta}), \tag{2}$$

where $var_0(\hat{\theta})$ is the variance of the treatment effect estimator for one of the components of the batched design

with one cluster per sequence. This result can be generalised to the situation where C_b clusters are assigned to each sequence *of batch b. When this is the case,*

$$var(\hat{\theta}) = \frac{1}{\sum_{b=1}^{B} C_b} var_0(\hat{\theta}). \tag{3}$$

Proof. The proof of this result follows directly from the proof of the more general Result 2, and is shown at the end of the proof of that result below. □

Result 2. *If Y_{bkti} is the outcome for participant $i = 1, \ldots, m_{bkt}$ in period $t = 1, \ldots, T_b$ in cluster $k = 1, \ldots, K_b$ in batch $b = 1, \ldots, B$, and Y_b is the $M_b = \sum_{k=1}^{K_b} \sum_{t=1}^{T_b} m_{bkt}$-length vector of outcomes from all clusters in batch b, we suppose that*

$$Y_b \sim N(Z_b\gamma_b + \theta X_b, \Sigma_b)$$

where γ_b is the T_b-length vector of period effects for batch b, Z_b is the design matrix associated with these period effects for cluster b (of dimension $M_b \times T_b$), θ is the treatment effect of interest (assumed to be shared across all batches), X_b is the M_b-length vector indicating if a participant is in a cluster-period in the control condition ($X_{bkti} = 0$) or the intervention condition ($X_{bkti} = 1$), and Σ_b is the $M_b \times M_b$ covariance matrix of the outcomes from all clusters in batch b. Then if $\hat{\theta}$ is the generalised least squares estimator of θ,

$$var(\hat{\theta}) = \left(\sum_{b=1}^{B} \frac{1}{var_b(\hat{\theta})} \right)^{-1}, \tag{4}$$

where $var_b(\hat{\theta})$ is the variance of the generalised least squares estimator of θ obtained by considering batch b only. Further, if $var_b(\hat{\theta}) = var_0(\hat{\theta})$ then $var(\hat{\theta}) = \frac{1}{B} var_0(\hat{\theta})$.

Proof. First write $Y = (Y_1, \ldots, Y_B)'$, the $\sum_{b=1}^{B} M_b$-length vector of all outcomes from the batched stepped wedge trial. Then we can write

$$Y \sim N(G\beta, \Sigma) \tag{5}$$

with

$$G = \begin{pmatrix} Z_1 & 0_{M_1 \times T_2} & \cdots & 0_{M_1 \times T_B} & X_1 \\ 0_{M_2 \times T_1} & Z_2 & \cdots & 0_{M_2 \times T_B} & X_2 \\ \vdots & \vdots & & & \\ 0_{M_B \times T_1} & 0_{M_B \times T_2} & \cdots & Z_B & X_B \end{pmatrix}, \quad \beta = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_B \\ \theta \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_1 & 0_{M_1 \times M_2} & \cdots & 0_{M_1 \times M_B} \\ 0_{M_2 \times M_1} & \Sigma_2 & \cdots & 0_{M_2 \times M_B} \\ \vdots & & \ddots & \\ 0_{M_B \times M_1} & 0_{M_B \times M_2} & \cdots & \Sigma_B \end{pmatrix} \tag{6}$$

where $0_{n\times m}$ is an $n \times m$ matrix of zeros. Then the generalised least squares estimator of β is given by:

$$\hat{\beta} = (G^T\Sigma^{-1}G)^{-1}G^T\Sigma^{-1}Y$$

and

$$var(\hat{\beta}) = (G^T\Sigma^{-1}G)^{-1}.$$

The blocked structure of G and Σ means that

$$G^T\Sigma^{-1}G = \begin{pmatrix} Z_1^T\Sigma_1^{-1}Z_1 & 0_{M_1\times M_2} & \cdots & 0_{M_1\times M_B} & Z_1^T\Sigma_1^{-1}X_1 \\ 0_{M_2\times M_1} & Z_2^T\Sigma_2^{-1}Z_2 & \cdots & 0_{M_2\times M_B} & Z_2^T\Sigma_1^{-1}X_2 \\ \vdots & & \ddots & & \vdots \\ 0_{M_B\times M_1} & 0_{M_B\times M_2} & \cdots & Z_B^T\Sigma_B^{-1}Z_B & Z_B^T\Sigma_1^{-1}X_B \\ X_1^T\Sigma_1^{-1}Z_1 & X_2^T\Sigma_2^{-1}Z_2 & \cdots & X_B^T\Sigma_B^{-1}Z_B & \sum_{b=1}^B X_b^T\Sigma_b X_b \end{pmatrix} \tag{7}$$

and $var(\hat{\theta})$ is the final entry in $(G^T\Sigma^{-1}G)^{-1}$

$$\begin{aligned} var(\hat{\theta}) &= \left\{\sum_{b=1}^B X_b^T\Sigma_b^{-1}X_b - \sum_{b=1}^B X_b^T\Sigma_b^{-1}Z_b\left(Z_b^T\Sigma_b^{-1}Z_b\right)^{-1}Z_b^T\Sigma_b^{-1}X_b\right\}^{-1} \\ &= \left\{\sum_{b=1}^B \left(X_b^T\Sigma_b^{-1}X_b - X_b^T\Sigma_b^{-1}Z_b\left(Z_b^T\Sigma_b^{-1}Z_b\right)^{-1}Z_b^T\Sigma_b^{-1}X_b\right)\right\}^{-1}. \end{aligned} \tag{8}$$

Note that $var_b(\hat{\theta}) = \left(X_b^T\Sigma_b^{-1}X_b - X_b^T\Sigma_b^{-1}Z_b\left(Z_b^T\Sigma_b^{-1}Z_b\right)^{-1}Z_b^T\Sigma_b^{-1}X_b\right)^{-1}$ (i.e. the variance of $\hat{\theta}$ were only batch b used to estimate θ).

Hence,

$$var(\hat{\theta}) = \left\{\sum_{b=1}^B var_b(\hat{\theta})^{-1}\right\}^{-1}. \tag{9}$$

If $var_b(\hat{\theta}) = var_0(\hat{\theta})$ for $b = 1, \ldots, B$, then

$$var(\hat{\theta}) = \left(Bvar_0(\hat{\theta})^{-1}\right)^{-1} = \frac{1}{B}var_0(\hat{\theta}). \tag{10}$$

□

Result 3. *If Y_{bkti} is the outcome for participant $i = 1, \ldots, m_{bkt}$ in period $t = 1, \ldots, T_b$ in cluster $k =$*

$1, \ldots, K_b$ *in batch* $b = 1, \ldots, B$, *with* $\mu_{bkti} = E[Y_{bkti}]$, *we assume*

$$g(\mu_{bkti}) = \beta_{bt} + \theta X_{bkt}. \tag{11}$$

where g *is the link function,* β_{bt} *is the fixed effect for period* t *in batch* b, θ *is the treatment effect of interest, and* X_{bkt} *is the indicator for whether cluster* k *in batch* b *and period* t *is in the intervention or control condition. Let* μ_b *be the vector of means for batch* b. *If* $\beta_b = (\beta_{b1}, \ldots, \beta_{bT_b})^T$ *is the set of time effects for batch* b, $\hat{\beta}_b$ *is the generalised least squares estimator of* β_b, *and* $\hat{\theta}$ *is the generalised least squares estimator of* θ *then*

$$\begin{aligned} var(\hat{\theta}) &= \left(\sum_{b=1}^{B} \frac{\partial \mu_b}{\partial \hat{\theta}}^T W_b^{-1} \frac{\partial \mu_b}{\partial \hat{\theta}} - \frac{\partial \mu_b}{\partial \hat{\theta}}^T W_b^{-1} \frac{\partial \mu_b}{\partial \hat{\beta}_b} \left[\frac{\partial \mu_b}{\partial \hat{\beta}_b}^T W_b^{-1} \frac{\partial \mu_b}{\partial \hat{\beta}_b} \right]^{-1} \frac{\partial \mu_b}{\partial \hat{\beta}_b}^T W_b^{-1} \frac{\partial \mu_b}{\partial \hat{\theta}} \right)^{-1} \\ &= \left(\sum_{b=1}^{B} \frac{1}{var_b(\hat{\theta})} \right)^{-1} \end{aligned} \tag{12}$$

where $var_b(\hat{\theta})$ *is the variance of the treatment effect estimator obtained via GEE when batch* b *is considered separately, and* W_b *is the covariance matrix of the observations from batch* b. W_b *has the form* $A_b^{1/2} R_b A_b^{1/2}$. A_b *is a diagonal matrix with elements given by* $var(Y_{bkti})$ *and* R_b *is the assumed correlation matrix of the observations from batch* b.

If a binomial distribution for outcomes is assumed $\mu_{bkti} = P(Y_{bkti} = 1)$ *and diagonal elements of* A_b *will be given by* $var(Y_{bkti}) = \mu_{bkti}(1 - \mu_{bkti})$.

Proof. If Y_{bkti} is the outcome for participant i in period t in cluster k in batch b, then write $\mu_{bkti} = E[Y_{bkti}]$ and consider some link function g so that

$$g(\mu_{bkti}) = \beta_{bt} + \theta X_{bkt}.$$

Consider the vectors of all parameters (including the treatment effect θ) $\beta = (\beta_{11}, \beta_{12}, \ldots, \beta_{1T}, \ldots, \beta_{B1}, \beta_{B2}, \ldots, \beta_{BT}, \theta)^T$, all observations Y and all means μ. Then, by [2] the GEE estimator for β is given by the solution to

$$D^T V^{-1} (Y - \mu) = 0 \tag{13}$$

where $D = \frac{\partial \mu}{\partial \beta^T}$, $V = A^{1/2} R A^{1/2}$ where R is the working correlation matrix and A has diagonal elements given by $\phi var(Y_{bkti})$. ϕ is a dispersion parameter; for our derivations we will assume that this is equal to 1. R is supposed to have a block-diagonal structure, with the blocks R_b corresponding to batches. That is, we only assume that observations in distinct batches are independent but make no assumptions about the supposed correlation within batches. The estimator $\hat{\beta}$ will be approximately normally distributed with mean

β and covariance matrix given by $(D^T V^{-1} D)^{-1}$. Of interest is the variance of $\hat{\theta}$, which corresponds to the element in the lower right hand corner of this matrix.

Since separate period effects are assumed for each batch in the model in Equation 11, we can write

$$g(\mu) = \begin{pmatrix} W_1 & & & X_1 \\ & W_2 & & X_2 \\ & & \ddots & \vdots \\ & & W_B & X_B \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_B \\ \theta \end{pmatrix} \tag{14}$$

where $\beta_b = (\beta_{b1}, \ldots, \beta_{bT})^T$ is the vector of time effects for batch b, and W_b is the corresponding design matrix for these time effects. X_b is the vector of treatment effect indicators for batch b. With Y_b, μ_b and V_b defined similarly, Equation 13 can be written as

$$\sum_{b=1}^{B} D_b^T V_b^{-1} (Y_b - \mu_b) = 0$$

where

$$D_b = \begin{pmatrix} 0 & \cdots & 0 & \frac{\partial \mu_b}{\partial \beta_b} & 0 & \cdots & 0 & \frac{\partial \mu_b}{\partial \theta} \end{pmatrix}.$$

Hence we can write the variance of the estimator β as

$$\left(\sum_{b=1}^{B} D_b^T V_b^{-1} D_b \right)^{-1} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1}^T V_1^{-1} \frac{\partial \mu_1}{\partial \beta_1} & \cdots & & \frac{\partial \mu_1}{\partial \beta_1}^T V_1^{-1} \frac{\partial \mu_1}{\partial \theta} \\ \vdots & \ddots & \vdots & \\ & \cdots & \frac{\partial \mu_B}{\partial \beta_B}^T V_B^{-1} \frac{\partial \mu_B}{\partial \beta_B} & \frac{\partial \mu_B}{\partial \beta_B}^T V_B^{-1} \frac{\partial \mu_B}{\partial \theta} \\ \frac{\partial \mu_1}{\partial \beta_1}^T V_1^{-1} \frac{\partial \mu_1}{\partial \theta} & \cdots & \frac{\partial \mu_B}{\partial \beta_B}^T V_B^{-1} \frac{\partial \mu_B}{\partial \theta} & \sum_{b=1}^{B} \frac{\partial \mu_b}{\partial \theta}^T V_b^{-1} \frac{\partial \mu_b}{\partial \theta} \end{pmatrix}^{-1}. \tag{15}$$

Interest is in the variance of $\hat{\theta}$, which is given by the bottom right entry of this matrix:

$$\begin{aligned} var(\hat{\theta}) &= \left[\sum_{b=1}^{B} \left(\frac{\partial \mu_b}{\partial \theta}^T V_b^{-1} \frac{\partial \mu_b}{\partial \theta} - \frac{\partial \mu_b}{\partial \theta}^T V_b^{-1} \frac{\partial \mu_b}{\partial \beta_b} \left\{ \frac{\partial \mu_b}{\partial \beta_b}^T V_b^{-1} \frac{\partial \mu_b}{\partial \beta_b} \right\}^{-1} \frac{\partial \mu_b}{\partial \beta_b}^T V_b^{-1} \frac{\partial \mu_b}{\partial \theta} \right) \right]^{-1} \\ &= \left(\sum_{b=1}^{B} \frac{1}{var_b(\hat{\theta})} \right)^{-1} \end{aligned} \tag{16}$$

where $var_b(\hat{\theta})$ is the variance of the treatment effect estimator obtained when batch b is considered independently.

□

2 Code to replicate power calculations

This code uses version 1.6 of the `swdpwr` R package [1].

```
###################
# Demonstration for PACT-HF study #
###################
#2 batches of a 5-sequence, 6-period stepped wedge design
#54 patients in each cluster in each period
#Baseline prevalence of 28%
#ICC 0.01
#Reduction to 21%

library(swdpwr)

onebatch <- matrix(c(c(0,1,1,1,1,1),c(0,0,1,1,1,1), c(0,0,0,1,1,1),
                     c(0,0,0,0,1,1), c(0,0,0,0,0,1)),5,6,byrow=TRUE)

completedesign <- rbind(onebatch, onebatch)

PACTHF_K <- 54

# ICC=0.01, Hussey and Hughes within-cluster correlation structure
PACTHF_alpha0 <- 0.01
PACTHF_alpha1 <- 0.01

# P(outcome|control) = 0.28
# P(outcome|treatment) = 0.21

# First: assume that there is no underlying trend in probability of outcome
PACTHF_meanresponse_start = 0.28
PACTHF_meanresponse_end0 = 0.28
PACTHF_meanresponse_end1 = 0.21
PACTHFpower_nochange <- swdpower(K = PACTHF_K, design = completedesign,
                     family = "binomial", model = "marginal",
```

```
                        link = "logit", type = "cross-sectional",
                        meanresponse_start =  PACTHF_meanresponse_start,
                        meanresponse_end0 = PACTHF_meanresponse_end0,
                        meanresponse_end1 = PACTHF_meanresponse_end1,
                        typeIerror = 0.05, alpha0 = PACTHF_alpha0, alpha1 = PACTHF_alpha1)
#Power is 98.8%


#Allowing for baseline prevalence to change over time
PACTHF_batch1 <- swdpower(K = PACTHF_K, design = onebatch,
                                family = "binomial", model = "marginal",
                                link = "logit", type = "cross-sectional",
                                meanresponse_start =  0.30,
                                meanresponse_end0 = 0.29,
                                meanresponse_end1 = 0.2175,
                                typeIerror = 0.05, alpha0 = PACTHF_alpha0, alpha1 = PACTHF_alpha1)



PACTHF_batch2 <- swdpower(K = PACTHF_K, design = onebatch,
                                family = "binomial", model = "marginal",
                                link = "logit", type = "cross-sectional",
                                meanresponse_start =  0.29,
                                meanresponse_end0 = 0.28,
                                meanresponse_end1 = 0.21,
                                typeIerror = 0.05, alpha0 = PACTHF_alpha0, alpha1 = PACTHF_alpha1)


#Variance of treatment effect estimator for each batch:
PACTHFtreateff_batch1 <- abs(as.numeric(PACTHF_batch1$treatment.effect.beta))
PACTHFpower_batch1 <- as.numeric(PACTHF_batch1$Power)
PACTHFvar_batch1 <-  1/(( qnorm(1-PACTHFpower_batch1) + qnorm(0.025))/PACTHFtreateff_batch1)^2


PACTHFtreateff_batch2 <- abs(as.numeric(PACTHF_batch2$treatment.effect.beta]))
PACTHFpower_batch2 <- as.numeric(PACTHF_batch2$Power)
PACTHFvar_batch2 <-  1/(( qnorm(1-PACTHFpower_batch2) + qnorm(0.025))/PACTHFtreateff_batch2)^2
```

```
#The effect size we wish to detect is -0.38
PACTHF_var_batches <- 1/(1/PACTHFvar_batch1 + 1/PACTHFvar_batch2)
PACTHFpower_batches <- pnorm(qnorm(0.025)+0.38/sqrt(PACTHF_var_batches))
#Power is 80.7698%
```

References

[1] J. Chen, X. Zhou, F. Li, and D. Spiegelman. swdpwr: A SAS macro and an R package for power calculation in stepped wedge cluster randomized trials. *ArXiV*, page arxiv:2011.06031v1, 2020.

[2] S. L. Zeger and K.-Y. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130, 1986.

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

The batched stepped wedge design: a design robust to delays in cluster recruitment

Jessica Kasza[1] | Rhys Bowden[1] | Richard Hooper[2] | Andrew B. Forbes[1]

[1]School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

[2]Centre for Primary Care and Public Health, Queen Mary University of London, United Kingdom

Correspondence
*Jessica Kasza, Monash University, 553 St Kilda Road, Melbourne, Victoria 3004, Australia. Email: jessica.kasza@monash.edu

Summary

Stepped wedge designs are an increasingly popular variant of longitudinal cluster randomized trial designs, and roll out interventions across clusters in a randomized, but step-wise fashion. In the standard stepped wedge design, assumptions regarding the effect of time on outcomes may require that all clusters start and end trial participation at the same time. This would require ethics approvals and data collection procedures to be in place in all clusters before a stepped wedge trial can start in any cluster. Hence, although stepped wedge designs are useful for testing the impacts of many cluster-based interventions on outcomes, there can be lengthy delays before a trial can commence.

In this paper we introduce "batched" stepped wedge designs. Batched stepped wedge designs allow clusters to commence the study in batches, instead of all at once, allowing for staggered cluster recruitment. Like the stepped wedge, the batched stepped wedge rolls out the intervention to all clusters in a randomized and step-wise fashion: a series of self-contained stepped wedge designs. Provided that separate period effects are included for each batch, software for standard stepped wedge sample size calculations can be used. With this time parameterization, in many situations including when linear models are assumed, sample size calculations reduce to the setting of a single stepped wedge design with multiple clusters per sequence. In these situations sample size calculations will not depend on the delays between the commencement of batches. Hence, the power of batched stepped wedge designs is robust to unexpected delays between batches.

KEYWORDS:
cluster randomised trial; intracluster correlation; sample size calculation; within-cluster correlation structure

1 | INTRODUCTION

The stepped wedge cluster randomised trial design, where clusters are randomised to switch from a control to an intervention condition at different pre-specified time points, has found application in a wide variety of research areas (examples in Mdege et al.[1]). Figure 1 displays an example of a conventional stepped wedge design with four periods and three treatment sequences. The period lengths are typically of equal duration and define the times at which different clusters cross from the control to the intervention condition. Stepped wedge designs are useful when intervention conditions applied at the level of the cluster

cannot be removed once implemented, e.g. educational interventions, or when assessing changes in policy that will be rolled out across systems. A crucial advantage of stepped wedge trials is that they may require fewer clusters and smaller total sample sizes than standard cluster randomised trials, due to the within-cluster comparisons enabled by the stepped wedge design[2]. It is well-recognised that the grouping of participants in clusters must be accounted for in sample size calculations and analysis of data from stepped wedge trials. In addition, due to the dependence between time and treatment in the stepped wedge, it is also essential to account for time in these calculations and analyses[3].

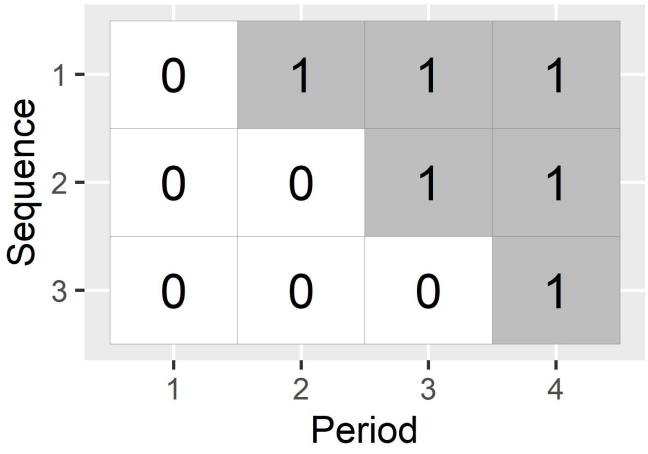

FIGURE 1 An example of a standard stepped wedge design, with 4 periods and 3 sequences (0 indicates periods in which the control condition is implemented; 1 indicates periods in which the intervention is implemented).

In the most commonly-used sample size formulas and statistical models for the design and analysis of stepped wedge trials (e.g[4,5,6]) it is assumed that the effect of "time" on outcomes is identical across clusters, and that time is divided up into distinct trial periods. If clusters commence study participation at different times (i.e. not all on the same date), then a distinction must be drawn between "calendar time" and "time-on-trial" (the amount of time since a cluster commenced trial participation). This distinction is particularly important for stepped wedge designs, where time and treatment are confounded. When clusters commence participation in a trial at the same calendar time, then calendar time and time-on-trial will be aligned: this is the case in Figure 1, where all clusters commence the study at the same time. When clusters are not aligned in calendar time, researchers must consider the distinction between calendar time and time-on trial, and parameterize time to align with their assumptions about the effect of time on outcomes in their statistical models. Three different time parameterizations that could be chosen when clusters are not aligned in calendar time are displayed in Figure 2: calendar time effects could be shared across clusters; time-on-trial effects could be shared across clusters; or separate period effects could be assumed in each batch. If clusters are not aligned in calendar time, but a standard stepped wedge sample size formula is applied, the assumption is that time-on-trial has an identical impact across all clusters, and there is no impact of calendar time (corresponding to time-on-trial effects that are shared across batches as in the middle panel of Figure 2). Further, when the standard sample size formulas are applied, it is required that there are no systematic differences between the batches of clusters that commence trial participation at different time points.

In practice, the great majority of cluster randomised stepped wedge trials have been designed so that all the clusters commence their participation at the same calendar time. This is likely to be for two reasons: firstly because clusters may all have expressed an interest in collaborating from an early stage in the development of the trial, and are all ready to go when the trial begins, but secondly, perhaps, because of concerns about the most appropriate way to model calendar time versus time-on-trial, and the lack of methodological guidance. There may be situations where it is to a triallist's advantage to stagger the commencement of different clusters.

In this paper we formalize the situation where different groups of clusters commence trial participation at different calendar times, defining the "batched stepped wedge design". In the batched stepped wedge design, different groups of clusters commence participation in a stepped wedge trial at different times, in a "batched" structure. The models we consider allow for systematic

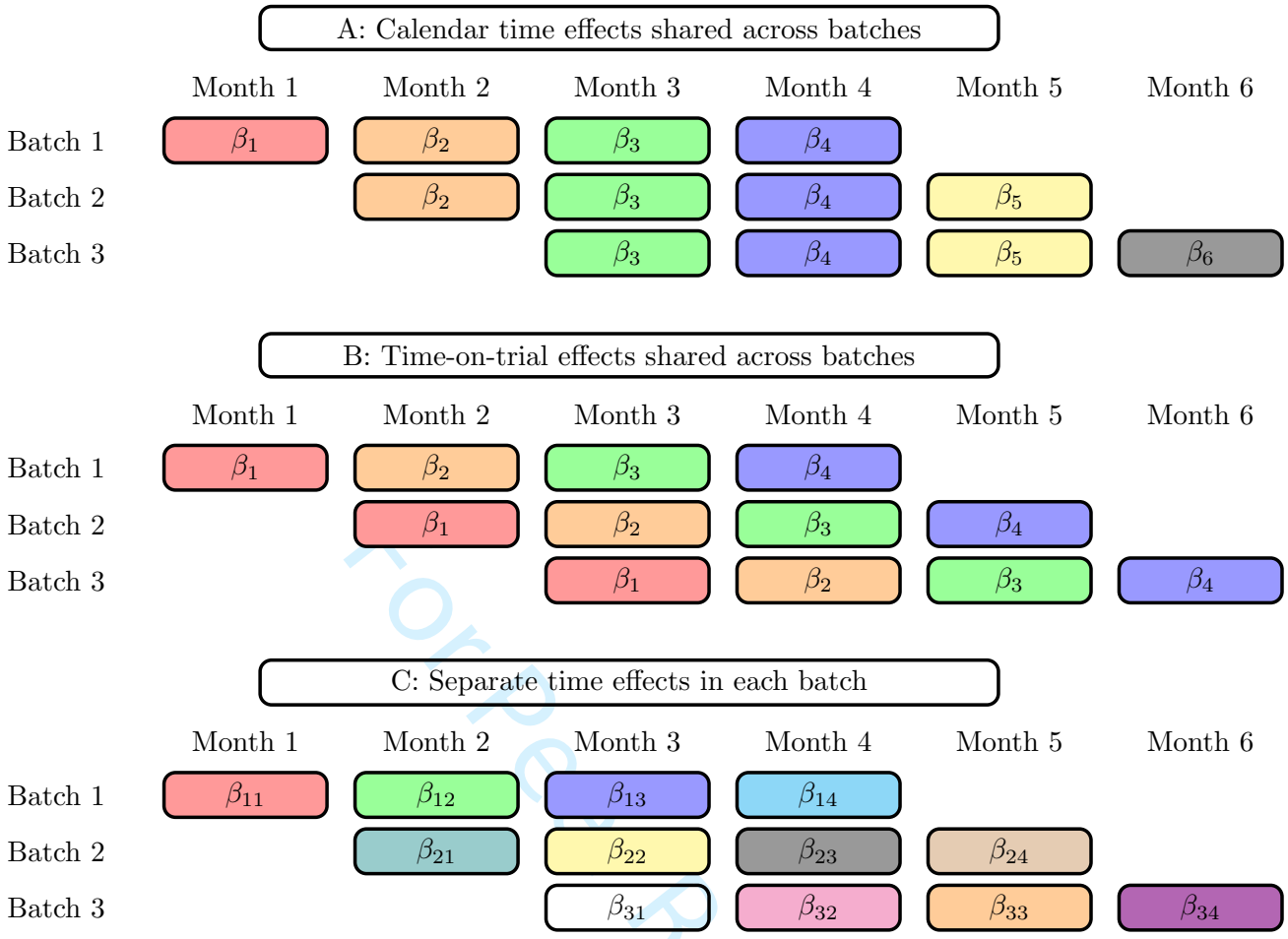

FIGURE 2 Three different ways in which the effect of time can be parameterised in a design where clusters commence study participation in three batches*: the βs parameterise the time effects. For example, in the top panel, β_1 parametrises the effect of month 1 on outcomes*. The top panel indicates how effects of time are shared across batches when the effects of calendar time are assumed to be constant across batches; the middle panel indicates how the effects of time are shared across batches when time-on-trial effects are assumed to be constant across batches; the bottom panel indicates that no time effects are shared across batches when separate time effects are estimated in each batch.

differences between clusters that commence study participation at different time points, and for differences in the effects of calendar and time-on-trial across these batches. This batched stepped wedge design is an alternative to the standard stepped wedge design: the batched design shares some of the benefits of the stepped wedge but allows for randomization of clusters to stepped wedge trial sequences in batches or blocks. Like the standard stepped wedge, the batched stepped wedge design ensures that all clusters eventually receive the intervention; treatment switches are unidirectional (i.e. the intervention is never removed once implemented); and the intervention is rolled out to each cluster in a randomized order. Examples when all batches are identical are shown in Figure 3; examples when batches differ are shown in Figure 4. This batched stepped wedge design can thus be conducted similarly to standard cluster randomized trials, where clusters may be randomised to the control or the intervention condition in groups as clusters are recruited to the trial, rather than all at once.

Although guidance for researchers and statisticians in sample size and power calculations for batched stepped wedge and related designs is lacking, researchers have already sought to implement designs similar to batched stepped wedge designs. For example, in a study assessing the impact of a mobility program for patients aged 60+ years across 8 veterans affairs hospitals in the USA on discharge destination of patients, Hastings et al.[7] sought to implement a batched stepped wedge design, with two batches of clusters. In reference to this batched structure, Hastings et al.[7] stated, "The full implications of a blocked randomization from a statistical perspective require further study". Similarly, the EAGLE study[8], investigating the impact of a quality

improvement intervention on the reduction of anastomotic leak following right colectomy, is randomizing hospitals in batches to a series of dog-leg designs (the dog-leg can be considered as an incomplete three-sequence stepped wedge design)[9].

The batched stepped wedge design may be appealing to researchers for three reasons: (1) depending on assumptions made about the outcome regression model, the power of a batched stepped wedge design will be unaffected by delays to the commencement of subsequent batches; (2) it can allow trials to get started sooner, by allowing clusters to come on-line in batches (i.e. ethics approvals and data collection procedures can be rolled out across clusters after the study has commenced); and (3) it does not require data to be collected by all clusters in all periods (i.e. this design can be thought of as an "incomplete" stepped wedge design. Expanding on the first reason: in this paper we show that for linear mixed models, if separate period effects are included for each batch of a batched stepped wedge (as in the bottom panel of Figure 2), then the power of the batched stepped wedge design is equivalent to the power of a standard stepped wedge design with multiple clusters assigned to each sequence. We also show that this will also hold when further assumptions about the effect of time are made when binary outcomes are analysed using non-linear link functions and generalised estimating equations. That is, under these assumptions, the designs in Figure 3 would have equivalent power to detect a difference as the design in Figure 1 with 3 clusters per sequence - although we would encourage trialists to include more than 9 clusters in any stepped wedge trial[10].

In this paper we provide researchers with guidance regarding the statistical aspects of batched stepped wedge designs, making recommendations regarding the inclusion of batch and period effects in the outcome regression model. In Section 2 we describe the batched stepped wedge design; in Section 3 we consider sample size calculations for the batched stepped wedge assuming linear mixed models for outcomes; in Section 4 we consider binary outcomes modeled with generalized linear models fit via generalized estimating equations. In Sections 3 and 4 we discuss what assumptions must be made in the specification of the outcome regression model to ensure the robustness property of the batched stepped wedge (where study power is robust to delays in the recruitment and/or commencement of the next batch) will hold. In Section 5 we discuss under what conditions standard stepped wedge sample size software can be applied to batched stepped wedge designs and demonstrate this calculation via an example. *In Section 6 we present the results of a simulation study, and conclude* with a discussion of our results in Section 7.

2 | WHAT IS A BATCHED STEPPED WEDGE DESIGN?

Simply put, a batched stepped wedge design is a series of stepped wedge cluster randomised trials. There may be some overlap in time between the successive stepped wedge components of the batched stepped wedge design, i.e. some trial periods during which data is being collected from more than one batch of the design. The component stepped wedge trials may be identical (as in the batched stepped wedge designs in Figure 3), or they may differ (as in Figure 4). Different sets of clusters contribute data in different batches of the study, and within each batch, clusters are randomised to the different sequences of the component stepped wedge design. A batched stepped wedge design allows for the recruitment of clusters throughout the duration of a study: once enough clusters for one of the component stepped wedges have been recruited, these clusters can be randomised to the sequences of the next batch, and the next stepped wedge component can commence. The models that we propose account for systematic differences between the clusters in different batches.

In a standard stepped wedge design, the implicit assumption is that all clusters commence participation in the trial at the same time[11] (or that there are assumptions made regarding the effect of time-on-trial as discussed in the Introduction). This is in contrast to the way in which parallel, or standard, cluster randomised trials are conducted. When parallel cluster randomised trials are conducted, clusters are often recruited throughout the duration of the trial. As is well known, in the (unstratified) parallel cluster randomised trial, so long as equal numbers of clusters (with equal numbers of participants) are assigned to the control and intervention arms at each randomisation point, this successive recruitment has no impact on the power of the study. This also holds for cluster randomised crossover designs provided equal numbers of clusters with equal numbers of participants implement the control and the intervention arm at each time point (again, this observation is limited to unstratified designs). The reason for this is that for these parallel and cluster randomised trial designs, treatment condition and time are independent: at each time point of the study, half of the clusters and participants will be in the control condition, and the other half will be in the intervention. In the stepped wedge design, treatment and time are not independent: the proportion of clusters in the intervention condition increases as time passes[2]. Depending on how time and randomisation batch are accounted for in the outcome regression model used to inform sample size calculations, the batched randomisation could have an impact on study power for batched stepped wedge designs, due to the confounding of time and treatment.

It is now well-recognised that time/period effects need to be accounted for in sample size and power calculations for stepped wedge designs[3]. Time/period effects must similarly be accounted for in sample size calculations for batched stepped wedge designs; researchers must provide adequate justification if they do not account for period effects in this calculation. Further, clusters that are included in different batches of the design may differ from each other, and thus it is recommended that batch effects be included in the outcome regression model. When there is an overlap between batches (e.g. the middle and bottom panels of Figure 3; the top panel of Figure 4), we also recommend that separate fixed period effects be included for each batch (equivalent to fixed batch-by-period interaction terms being included in the model). There are three key reasons for this recommendation: the first is that it requires making the fewest assumptions about the effects of time on outcomes and whether these effects are shared across batches; the second is that under this assumption, the variance of the treatment effect estimator for the batched stepped wedge is a combination of the variance for each component stepped wedge; the third (and most important) is that under this assumption, in many situations, study power will be robust to delays in the commencement of batches

We now consider the variance of the treatment effect estimator for batched stepped wedge designs. We first consider linear mixed models in Section 3, discussing batched stepped wedge designs with identical and non-identical components separately, and then discuss generalized linear models fit via GEE in Section 4.

3 | BATCHED STEPPED WEDGE DESIGNS AND LINEAR MIXED MODELS

3.1 | Batched stepped wedges with identical components

We first suppose that the batched stepped wedge design being considered is composed of B batches of identical T-period and K-sequence designs (for the standard stepped wedge $K = T - 1$), and initially suppose that one cluster is assigned to each sequence of each batch. We consider the following linear mixed model for the outcome Y_{bkti} from participant $i = 1, \dots, m$ in period $t = 1, \dots, T$ from cluster $k = 1, \dots, K$ in batch $b = 1, \dots, B$:

$$Y_{bkti} = \beta_{bt} + \theta X_{bkt} + \alpha_{bkt} + \epsilon_{bkti}, \quad \epsilon_{bkti} \sim N(0, \sigma^2_\epsilon). \tag{1}$$

In this model we have numbered the periods within each batch separately and thus period is identical to time-on-trial: Y_{11Ti} represents the outcome for the ith participant in cluster 1 in the final period (period T) of the first batch. If there is one period of overlap between successive batches, period T of batch 1 would correspond to period 1 of batch 2. Our model set-up automatically allows for separate period effects in each batch through the inclusion of the β_{bt} period terms (i.e. the scenario in the bottom panel of Figure 2): there are $B \times T$ period terms in total. These fixed period effects could alternatively be parameterised as period effects (where the effect for each period is shared by all clusters contributing data in that period, no matter their batch), batch effects, and terms for the period-by-batch interaction. This would require either constraining some of the β_{bt} to be identical, or re-numbering period from 1 to the total number of periods in the entire study (e.g. in Design 1 of Figure 3, the period subscript would range from 1 to 12; in Design 2 of Figure 3, the period subscript would range from 1 to 10). Given that later in this paper we recommend that separate period effects be included for each batch, throughout this paper we will number period within each batch (conceiving of time as time-on-trial; although this distinction from calendar time is immaterial when including an interaction with the batch term).

The treatment effect of interest is θ, assumed to be constant across batches, and the treatment group of cluster k in batch b at time period t is indicated by the binary variable X_{bkt}. The T-length vector of random effects $\alpha_{bk} = \left(\alpha_{bk1}, \dots, \alpha_{bkT}\right)^T$ for cluster k in batch b is assumed to have a multivariate normal distribution, centered around zero, with a variance matrix such that $var(\alpha_{bkt}) = \sigma^2_\alpha$ and $cov(\alpha_{bkt}, \alpha_{bks}) = r_{ts}\sigma^2_\alpha$, with $0 \leq r_{ts} \leq 1$. If $r_{ts} = r^{|t-s|}$ for some $0 < r < 1$, the discrete-time decay model of[12] is returned; if $r_{ts} = r$ for some $0 < r \leq 1$, the nested exchangeable model is returned, with $r = 1$ corresponding to the Hussey and Hughes model[4].

It is mathematically convenient to collapse Model 1 to cluster-period means when investigating the statistical power of designs[13]:

$$Y_{bkt} = \frac{1}{m}\sum_{i=1}^{m} Y_{bkti} = \beta_{bt} + \theta X_{bkt} + \alpha_{bkt} + \epsilon_{bkt}, \quad \epsilon_{bkt} \sim N\left(0, \frac{\sigma^2_\epsilon}{m}\right). \tag{2}$$

In the following result we consider the variance of the treatment effect estimator for models of the form given in Equations 1 and 2.

Result 1. *Suppose that each batch of the batched stepped wedge design is identical, and models of the form in Equations 1 and 2 are considered, so that the cluster-period means from each cluster share a common variance matrix, denoted by V. V is a $T \times T$ matrix, with the (t, s) element given by $cov(Y_{bkt}, Y_{bks})$. If X_{bk} is the $T \times 1$ vector containing the treatment indicators of cluster k in batch b for all T periods, then $X_{bk} = X_{b'k} = X_k$ for all pairs of batches b and b', and the variance of the treatment effect estimator $\hat{\theta}$ is given by:*

$$var(\hat{\theta}) = \frac{1}{B}\left[\sum_{k=1}^{K} X_k^T V^{-1} X_k - \frac{1}{K}\left(\sum_{k=1}^{K} X_k^T V^{-1} \sum_{k=1}^{K} X_k\right)\right]^{-1} = \frac{1}{B} var_0(\hat{\theta}), \quad (3)$$

where $var_0(\hat{\theta})$ is the variance of the treatment effect estimator for one of the components of the batched design with one cluster per sequence. This result can be generalised to the situation where C_b clusters are assigned to each sequence of batch b. When this is the case,

$$var(\hat{\theta}) = \frac{1}{\sum_{b=1}^{B} C_b} var_0(\hat{\theta}). \quad (4)$$

Result 1 indicates that when batch-by-time interaction terms are included in the model for the outcome, the treatment effect estimator is simply a weighted combination of treatment effect estimators obtained from each batch separately. Specifically, the estimator from each batch is weighted by its variance.

Result 1 is a consequence of the more general result discussed in Section 3.2, with the proof provided in Section 1 of the Supplementary Material available online. Equation 3 indicates that the variance of the treatment effect estimator from the batched stepped wedge design with B batches of identical T-period stepped wedge designs on K clusters is equivalent to the variance of the treatment effect estimator for a single T-period stepped wedge design with $B \times K$ clusters. When C_b clusters are assigned to each sequence of batch b, the variance of the treatment effect estimator for the batched design is equivalent to that of the single component design with $\sum_{b=1}^{B} C_b$ clusters per sequence. When all batches are identical and a model such as that in Equation 1 is assumed, sample size calculations for batched stepped wedge designs are thus straightforward. We demonstrate such calculations in Section 5.

Model 1 can be extended to allow for closed or open cohort schemes (as described in Kasza et al.[14], for example), to incorporate treatment effect heterogeneity (as described in Kasza et al.[15], for example), and to allow for differing numbers of subjects in each cluster in each period (as described in Kasza at al.[16], for example). When treatment effect heterogeneity is included in the model, the variance matrix V will not be identical across the clusters within a batch. However, the variance of the treatment effect estimator for a batch (denoted by $var_0(\hat{\theta})$ in Equation 3) will be common across batches. Hence when treatment effect heterogeneity is included in the model, the variance of the treatment effect estimator from the batched stepped wedge design with B batches of identical component designs with C_b clusters per sequence in batch b is again equivalent to the variance of the treatment effect estimator for a single component with $\sum_{b=1}^{B} C_b$ clusters per sequence.

When different clusters have different numbers of participants in each cluster-period, there may not be a common joint variance matrix $var_0(\hat{\theta})$ across batches. When cluster sizes differ, but each cluster is expected to collect the same number of observations in each of their data collection periods (i.e. cluster k collects m_k observations in each period), researchers could calculate the mean and coefficient of variation of cluster sizes and use the approximation presented in[17] to obtain a common $var_0(\hat{\theta})$ for each batch of the design.

3.2 | Batched stepped wedges with non-identical components

We now consider batched stepped wedge designs with non-identical components (examples in Figure 4): we suppose that there are B batches of stepped wedge designs, where batch b is a T_b-period stepped wedge design, with K_b clusters. *Result 2 provides the variance of the treatment effect estimator when batches are no longer identical.*

Result 2. If Y_{bkti} is the outcome for participant $i = 1, \ldots, m_{bkt}$ in period $t = 1, \ldots, T_b$ in cluster $k = 1, \ldots, K_b$ in batch $b = 1, \ldots, B$, and Y_b is the $M_b = \sum_{k=1}^{K_b} \sum_{t=1}^{T_b} m_{bkt}$-length vector of outcomes from all clusters in batch b, we suppose that

$$Y_b \sim N(Z_b \gamma_b + \theta X_b, \Sigma_b)$$

where γ_b is the T_b-length vector of period effects for batch b, Z_b is the design matrix associated with these period effects for cluster b (of dimension $M_b \times T_b$), θ is the treatment effect of interest (assumed to be shared across all batches), X_b is the M_b-length vector indicating if a participant is in a cluster-period in the control condition ($X_{bkti} = 0$) or the intervention condition

($X_{bkti} = 1$), and Σ_b is the $M_b \times M_b$ covariance matrix of the outcomes from all clusters in batch b. Then if $\hat{\theta}$ is the generalised least squares estimator of θ,

$$var(\hat{\theta}) = \left(\sum_{b=1}^{B} \frac{1}{var_b(\hat{\theta})} \right)^{-1}, \tag{5}$$

where $var_b(\hat{\theta})$ is the variance of the generalised least squares estimator of θ obtained by considering batch b only. Further, if $var_b(\hat{\theta}) = var_0(\hat{\theta})$ then $var(\hat{\theta}) = \frac{1}{B} var_0(\hat{\theta})$.

The proof of Result 2 is provided in Section 1 of the Supplementary Material. *This result assumes normally distributed outcomes where only the treatment effect is shared across batches; no assumptions are made regarding the equality of within-cluster correlation structures within or between batches. However,* if $var_b(\hat{\theta}) = var_0(\hat{\theta})$ for some $var_0(\hat{\theta})$ for all $b = 1, \ldots, B$ then Equation 5 collapses to the *result given in Result 1, i.e. the situation where all batches are identical. Once again, the treatment effect estimator is the weighted sum of the estimators obtained for each batch, with each batch's estimator weighted by its variance.*

4 | BINARY *AND COUNT* OUTCOMES AND BATCHED STEPPED WEDGES

Several modelling options are available when researchers are interested in binary, rather than continuous outcomes. One option, discussed in Hussey and Hughes[4], and taken in Hemming et al.[5], is to apply Equation 1 to binary outcomes, setting σ^2_ϵ equal to $p(1-p)$, where $p = P(Y_{bkti} = 1)$. The generalized least squares estimator of θ is then considered, and results in Section 3 apply. However, it has been pointed out that the variance of the treatment effect estimator may not be adequately approximated when this approach is applied[18]. Zhou et al.[18] developed an alternative approach, assuming a truncated normal distribution for cluster random effects, to ensure that estimated probabilities lie between 0 and 1. When batch and period are parameterised as in Equation 1 (i.e. separate period effects are included for each batch in the outcome regression model), the variance of the treatment effect estimator can be written as the sum of variances for each stepped wedge component as for the continuous outcome.

When binary *or count* outcomes are of interest, researchers are frequently interested in estimating a marginal treatment effect instead of the conditional treatment effect. The use of generalized estimating equations (GEE) for the analysis of longitudinal cluster randomized trials allows for estimation of such marginal effects, and implications of this analysis approach for sample size calculations have previously been discussed[6,19]. *When the GEE approach is used, a working correlation matrix structure must be assumed. This working correlation structure describes the pattern of within-cluster correlations; an exchangeable correlation structure would imply equal correlations between all observations in a cluster, for example. As discussed in Li et al.[6], when GEE is the intended analysis approach, power calculations can proceed via generalized least squares. We now state the main result for this scenario.*

Result 3. *If Y_{bkti} is the outcome for participant $i = 1, \ldots, m_{bkt}$ in period $t = 1, \ldots, T_b$ in cluster $k = 1, \ldots, K_b$ in batch $b = 1, \ldots, B$, with $\mu_{bkti} = E[Y_{bkti}]$, we assume*

$$g(\mu_{bkti}) = \beta_{bt} + \theta X_{bkt}. \tag{6}$$

where g is the link function, β_{bt} is the fixed effect for period t in batch b, θ is the treatment effect of interest, and X_{bkt} is the indicator for whether cluster k in batch b and period t is in the intervention or control condition. Let μ_b be the vector of means for batch b. If $\beta_b = (\beta_{b1}, \ldots, \beta_{bT_b})^T$ is the set of time effects for batch b, $\hat{\beta}_b$ is the generalised least squares estimator of β_b, and $\hat{\theta}$ is the generalised least squares estimator of θ then

$$\begin{aligned} var(\hat{\theta}) &= \left(\sum_{b=1}^{B} \frac{\partial \mu_b}{\partial \hat{\theta}}^T W_b^{-1} \frac{\partial \mu_b}{\partial \hat{\theta}} - \frac{\partial \mu_b}{\partial \hat{\theta}}^T W_b^{-1} \frac{\partial \mu_b}{\partial \hat{\beta}_b} \left[\frac{\partial \mu_b}{\partial \hat{\beta}_b}^T W_b^{-1} \frac{\partial \mu_b}{\partial \hat{\beta}_b} \right]^{-1} \frac{\partial \mu_b}{\partial \hat{\beta}_b}^T W_b^{-1} \frac{\partial \mu_b}{\partial \hat{\theta}} \right)^{-1} \\ &= \left(\sum_{b=1}^{B} \frac{1}{var_b(\hat{\theta})} \right)^{-1} \end{aligned} \tag{7}$$

where $var_b(\hat{\theta})$ is the variance of the treatment effect estimator obtained via GEE when batch b is considered separately, and W_b is the covariance matrix of the observations from batch b. W_b has the form $A_b^{1/2} R_b A_b^{1/2}$. A_b is a diagonal matrix with elements

given by var(Y_{bkti}) and R_b is the assumed correlation matrix of the observations from batch b. If a binomial distribution for outcomes is assumed $\mu_{bkti} = P(Y_{bkti} = 1)$ and diagonal elements of A_b will be given by var(Y_{bkti}) = $\mu_{bkti}(1 - \mu_{bkti})$.

The proof of this result is shown in the Appendix. As was the case in the linear mixed model scenario, the treatment effect estimator is again a weighted sum of each batch's treatment effect estimator. However, in contrast to the linear model discussed in Section 3, the variance of the treatment effect estimator depends on the assumed period effects through μ_{bkti} in the model in Equation 6. *For example,* when calculating sample sizes for a batched stepped wedge trial where a marginal model will be used to analyse binary outcomes, researchers must include predicted prevalences of the outcome in each period of the trial in sample size calculations. For binary outcomes that will be analysed in this way, the variance of the batched stepped wedge design will collapse to the simplified form given in Equation 4 only if no period effects are included in the model. This is a strict assumption, requiring the prevalence of the outcome in the control arm to be the same for the entire trial duration and across batches (i.e. there are no secular time effects and no batch effects).

Result 3 applies not only to binary outcomes analysed with a logit link function; the proof given in the Appendix does not rely on the choice of link function or the outcome type. Thus, Result 3 holds for binary outcomes with a linear link function, or for count outcomes with a log link function, to name just two alternatives.

5 | DEMONSTRATION OF POWER CALCULATIONS FOR BATCHED STEPPED WEDGE DESIGNS

We demonstrate how to calculate power for a batched stepped wedge design with two identical batches. The example we consider is from Unni et al.[20]: in that paper, various different stepped wedge-like designs were considered for the Patient-Centered Care Transitions in Heart Failure Trial (PACT-HF), including a batched design with "early" and "late" blocks, shown in Figure 5. Each of these blocks was a 5-sequence, 6-period stepped wedge design, with one cluster assigned to each sequence, with 54 patients in each cluster in each of these periods. The primary outcome considered was a binary outcome, that was a composite of a number of clinical outcomes, with a prevalence of 28% under the control condition, and an intracluster correlation of 0.01. We assume that this intracluster correlation was conditional on the inclusion of a "batch" term in the model (a point we return to in the Discussion), but in practice we would recommend assessing the impact of varying this correlation on study power. The aim was to detect a 25% reduction in the prevalence of the outcome: that is, a reduction from 28% to 21%. The effect on the logit link scale is -0.38.

We consider two methods for calculating the power of this design: first, we assume a linear model for the binary outcome (applying the results of Section 3.1); second, we assume that a generalised estimating equations approach will be taken to fitting a model with a logit link (applying the results of Section 4).

To perform the power calculation for the first approach, one need only calculate the power of a standard 6-period stepped wedge design with 2 clusters assigned to each sequence. The Shiny CRT calculator[5] accommodates this by allowing users to set the number of clusters assigned to each sequence; the Stata `steppedwedge` program[21] accommodates this through the "*k*" option. When the linear model is assumed, this study has a power of 77% to detect the difference. There are two additional ways to use the Shiny CRT calculator to calculate the power of the batched design. The user could get the precision of each of the component designs separately using the the "Precision" tab on the Shiny CRT calculator, and then combine these according to Result 1. Alternatively, the user could upload the design matrix for the batched design (ensuring that there is no overlap between successive batches) and obtain the power of the design directly. Were a design matrix uploaded with an overlap between successive batches, the Shiny CRT calculator would assume that batches with overlapping periods share period effects (that is, calendar time effects would be shared across batches as in the top panel of Figure 2).

For the second method we use the R `swdpwr` package[22] to calculate the power of the batched stepped wedge design. Since the GEE approach depends on the baseline prevalence of the outcome, we consider two different scenarios:

1. The prevalence of the outcome under the control condition remains at 28% for the entire duration of the trial.
2. The prevalence of the outcome under the control condition is initially 30%, but decreases to 28% by the final period of the trial, in a linear fashion. That is, at the time that the second batch starts data collection, the prevalence of the outcome under the control condition is 29%.

Since there is no change in the underlying prevalence of the condition over time in the first scenario, the power of the batched stepped wedge using the GEE approach is equivalent to the power of the 6-period stepped wedge with 2 clusters assigned to

each sequence. The power of this design can be obtained directly by using the `swdpower` command in the `swdpwr` package, and is 98.8%. Power is high due to the omission of period effects in this calculation.

For the second scenario, the variance of the treatment effect estimator must be obtained for each of the two component designs separately. We assume that the treatment effect is -0.38 on the logit link scale for both batches, with the aim to detect a reduction from 29% to 21.75% for the first batch, and from 28% to 21% in the second batch. The `swdpower` command cannot provide the power of the batched design directly. However, the variance of the treatment effect estimator can be obtained for each batch separately from the power calculated by the command. These variances are then combined using Result 3. The power of the batched design to detect a change from 28% to 21% is 80.8%. Commands to replicate this calculation are provided in Section 2 of the Supplementary Material.

6 | *SIMULATION STUDY*

We conducted a simulation study to verify our theoretical results, inspired by the PACT-HF design discussed in Section 5. As in the design schematic in Figure 5, we consider a design consisting of two batches, each a 6-period stepped wedge design. However, we vary the number of overlapping periods between the two batches from 0 (indicating no overlap between the two batches, as in Figure 5) to 5 (batches that overlap completely); the key aim of this simulation study is to assess whether inclusion of separate period effects for each batch has an impact on empirical power. Does power decrease as the number of batch-by-time terms in the model increases? In the simulation we increase the total number of clusters to 40 (4 clusters assigned to each of the 10 sequences). We simulate both binary and continuous outcomes for a range of correlation parameter values. Code to replicate this simulation study and the nested loop plots is available at `https://github.com/jkasza/BatchSW`.

Table 1 lists the parameters considered for the simulation study for the continuous outcomes. Along with varying the number of periods of overlap between successive batches, the intracluster correlation and the cluster autocorrelation, datasets were simulated with an effect size of 0 (to allow an examination of significance level) and 0.15. For each combination of parameters in Table 1, 1000 datasets were simulated, with separate time effects in each batch (as in the bottom panel of Figure 2). The period effects for each batch in each period were simulated from a normal distribution with mean 0 and variance 1. With 1000 simulated datasets, the Monte Carlo standard error associated with a power of 80% is expected to be around ±1.3%[23]. Each simulated dataset was analysed using a linear mixed-effects model with random effects for cluster and cluster-period, and separate categorical fixed period effects for each batch (i.e. period effects, batch effects, and period by batch interaction terms). Our focus here is on the comparison of theoretical and simulated power, so for each set of parameters, we calculated the percentage of hypothesis tests $H_0 : \theta = 0$ *rejected at the two-sided 5% significance level.*

TABLE 1 The continuous outcome simulation settings. 1000 datasets were simulated for each combination of parameters (108 combinations).

Parameter	Meaning	Values
T	Number of periods in each stepped wedge design	6
B	Number of batches	2
K	Number of clusters assigned to each sequence	4
m	Number of observations in each cluster in each sequence	10
N_O	Number of periods of overlap between successive batches	5, 4, 3, 2, 1, 0
ρ	Intra-cluster correlation	0.01, 0.05, 0.1
r	Cluster autocorrelation	1, 0.95, 0.75
θ	Effect size	0, 0.15

Figure 6 displays the empirical type I error rates and power for each set of parameters using nested loop plots[24]. This figure indicates that the number of periods of overlap does not have an impact on empirical type I error rates and power: as the number of periods of overlap changes, there is no pattern to the variation in empirical type I error rates or power. This provides support for our theoretical result, which indicates that if period, batch, and batch by period interaction terms are included in the

model, the degree of overlap has no impact on study power. As expected, Figure 6 does indicate that the intracluster correlation and cluster autocorrelation do have an impact on empirical power levels. It is interesting to note that for some combinations of the cluster autocorrelation and intracluster correlation (e.g. when the cluster autocorrelation is 0.75, and the intracluster correlation is equal to 0.05 or 0.1), the empirical power is slightly inflated. However, this does not change as the overlap between batches decreases. That is, empirical power does not decrease as the number of time effects included in the model increases.

Table 2 lists the parameters considered for the simulation study for the binary outcomes. As was the case for the simulation study for continuous outcomes, 1000 datasets were simulated for each combination of parameters. Binary data was simulated using the method of Qaqish[25], as coded by Li et al[19]. The range of intracluster correlations permitted by the simulation method of Qaqish is limited, so we only consider intracluster correlations of 0.01 and 0.05 for the binary outcomes. Each simulated dataset was analysed via GEE with a logit link with an exchangeable working correlation structure, with separate coefficients for period (treated as a continuous covariate) in each batch (this choice is to match the sample size calculation in the R `swdpwr` *package[22]). Again, our focus is on the comparison of theoretical and simulated power, so for each set of parameters and each analysis choice, we calculated the percentage of hypothesis tests* $H_0 : \theta = 0$ *rejected at the two-sided 5% significance level. Theoretical power for each combination of parameters was also calculated, using the* `swdpwr` *package. Figure 7 displays the empirical type I error rates and power for each set of parameters for the binary outcomes analysed via GEE with an exchangeable working correlation. As for the simulation study for continuous outcomes, the simulated power and type I error rates do not depend on the degree of overlap between successive batches, aligning with our theoretical results.*

TABLE 2 The binary outcome simulation settings. 1000 datasets were simulated for each combination of parameters (24 combinations).

Parameter	Meaning	Values
T	Number of periods in each stepped wedge design	6
B	Number of batches	2
K	Number of clusters assigned to each sequence	4
m	Number of observations in each cluster in each sequence	10
$P(Y_{bkt} = 1 \mid X_{bkt} = 0)$	Probability of the outcome in a non-treatment period	$0.4 + t \times 0.01$
N_O	Number of periods of overlap between successive batches	5, 4, 3, 2, 1, 0
ρ	Intra-cluster correlation	0.01, 0.05
r	Cluster autocorrelation	1
$P(Y_{bkt} = 1 \mid X_{bkt} = 1)$ $-P(Y_{kt} = 1 \mid X_{kt} = 0)$	Change in probability of the outcome caused by the intervention	0, 0.025

7 | DISCUSSION

The batched stepped wedge design is a promising alternative to the standard stepped wedge design. By allowing clusters to come on-line to the study in batches, the batched stepped wedge design has the potential to get started sooner than a standard stepped wedge, which typically requires all clusters to commence at the same point in time. If separate period effects are included for each batch (as in the bottom panel of Figure 2), the power of the batched stepped wedge design will, depending on the assumed outcome regression model, be robust to delays in the commencement of batches. This holds when linear models for the outcome are assumed, or when the prevalence of the outcome in the control condition is not expected to change over time. Hence, in these settings, study power will be unaffected if there is an unanticipated delay before the next batch commences study involvement when separate period effects are assumed for each batch. Under this assumed model, standard stepped wedge software can be used to calculate the required sample size and study power for batched stepped wedge designs.

Our key result indicates that in certain situations a batched stepped wedge design consisting of B identical stepped wedge designs provides the same power to detect an effect as one of the stepped wedge components with B clusters assigned to each sequence. However, the choice of variance components will have an impact on sample size and power calculations for all batched

stepped wedge designs. Inclusion of the batch term implies that variance components must now be treated as "within-batch" variance components, and will likely be smaller than if a model without batch effects was considered. When batch effects are included in the outcome regression model, variance components will be conditional on the inclusion of these batch effects in the model. Hence, researchers must consider the impact of batches on variance components and intracluster correlations when considering sample size and power.

Our key results have broad applicability. They generalise to batches of any other type of longitudinal cluster randomised trial design. and do not rely on the design type. For example, our results apply to a "batched dog-leg" design. Provided that separate period effects are included for each batch, the variance of such a design would have the form given in Sections 3 and 4: summing over the variances of treatment effect obtained for each of the individual component designs. Further, our key results do not depend on the precise form of the variance of the treatment effect estimator for each batch. The models considered in the Results could be extended to allow for closed or open cohorts, treatment effect heterogeneity, etc. The key result only requires that separate period effects are included in the outcome model for each batch: if this is the case, then the variances of the treatment effect estimators for each batch can be combined according to Results 1, 2, or 3 as appropriate.

We recommend that separate period effects are included for each stepped wedge batch (i.e. the time parameterization as in the bottom panel of Figure 2) to provide robustness to the sample size calculation in case of unexpected recruitment and set-up delays. In addition to allowing for robustness to delays and making the fewest assumptions about the effect of time on outcomes, assuming separate period effects across batches would be appropriate for trials where clusters in different batches are from geographically distinct areas, or where clusters in different batches are otherwise distinct. Additionally, if batch effects are not included in the outcome regression model and batches are assumed to have shared period effects for overlapping periods, then study power will depend on the separation between successive batches. If unexpected delays between batches occur, the power of the study will not be robust to this change, in that it will differ from that calculated a priori. Future work will investigate the impact of increasing degrees of overlap between successive batches when period effects are shared across batches.

Adaptive variants of the batched stepped wedge design are a logical next step of this work. Such adaptations may include sample size re-estimation, or more formal stopping rules for efficacy or futility of the intervention based on assessments at suitable time points, for example after participants in each batch have completed their followup. While adaptive variants of the stepped wedge design have been discussed in the literature, these do require that all clusters commence data collection at the same time. These adaptive variants will be explored in future work.

ACKNOWLEDGMENTS

This work was supported by the Australian Research Council Discovery Project DP210101398 and National Health and Medical Research Council of Australia Project Grant ID 1108283.

Conflict of interest

The authors declare no potential conflict of interests.

Data sharing

Results of simulations and the code to replicate the simulation study in Section 6 is available at `https://github.com/jkasza/BatchSW`.

References

1. Mdege ND, Man MS, Taylor CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *Journal of Clinical Epidemiology* 2011; 64(9): 936-948.

2. Matthews JNS, Forbes AB. Stepped wedge designs: insights from a design of experiments perspective. *Statistics in Medicine* 2017; 36(24): 3772-3790.

3. Hemming K, Taljaard M, McKenzie JE, et al. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. *BMJ* 2020; 363: k1614.

4. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials* 2007; 28: 182-191.

5. Hemming K, Kasza J, Hooper R, Forbes A, Taljaard M. A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator. *International Journal of Epidemiology* 2020; 49: 979-995.

6. Li F, Turner EL, Preisser JS. Sample size determination for GEE analyses of stepped wedge cluster randomized trials. *Biometrics* 2018; 74: 1450-1458.

7. Hastings SN, Stechuchak KM, Choate A, et al. Implementation of a stepped wedge cluster randomized trial to evaluate a hospital mobility program. *Trials* 2020; 21: 863.

8. ESCP EAGLE Safe Anastomosis Collective . ESCP Safe Anastomosis ProGramme in CoLorectal SurgEry (EAGLE): Study protocol for an international cluster randomised trial of a quality improvement intervention to reduce anastomotic leak following right colectomy. *Colorectal Disease* 2021: doi:10.1111/codi.15806.

9. Hooper R, Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *BMJ* 2015: 350.

10. Taljaard M, Teerenstra S, Ivers NM, Fergusson DA. Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. *Clinical Trials* 2016; 13: 459-463.

11. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford R. The stepped-wedge cluster randomised trial: rationale, design, analysis and reporting. *BMJ* 2015; 350: h391.

12. Kasza J, Hemming K, Hooper R, Matthews JNS, Forbes AB. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Statistical Methods in Medical Research* 2019; 28: 703-716.

13. Grantham KL, Kasza J, Heritier S, Hemming K, Forbes AB. Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. *Statistics in Medicine* 2019; 38: 1918-1934.

14. Kasza J, Hooper R, Copas A, Forbes AB. Sample size and power calculations for open cohort longitudinal cluster randomized trials. *Statistics in Medicine* 2020; 39: 1871-1883.

15. Kasza J, Taljaard M, Forbes AB. Information content of stepped-wedge designs when treatment effect heterogeneity and/or implementation periods are present. *Statistics in Medicine* 2019; 38: 4686-4701.

16. Kasza J, Bowden R, Forbes AB. Information content of stepped wedge designs with unequal cluster-period sizes in linear mixed models: Informing incomplete designs. *Statistics in Medicine* 2021: DOI: 10.1002/sim.8867.

17. Harrison LJ, Chen T, Wang R. Power calculation for cross-sectional stepped wedge cluster randomized trials with variable cluster sizes. *Biometrics* 2020; 76: 951-962.

18. Zhou X, Liao X, Kunz LM, Normand SLT, Wang M, Spiegelman D. A maximum likelihood approach to power calculations for stepped wedge designs of binary outcomes. *Biostatistics* 2020; 21: 102-121.

19. Li F, Forbes A, Turner EL, Preisser JS. Power and sample size requirements for GEE analyses of cluster randomized crossover trials. *Statistics in Medicine* 2019; 38: 636-649.

20. Unni RR, Lee SF, Thabane L, Connolly S, Van Spall HG. Variations in stepped-wedge cluster randomized trial design: Insights from the Patient-Centered Care Transitions in Heart Failure trial. *American Heart Journal* 2020; 220: 116-126.

21. Hemming K, Girling A. A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. *Stata Journal* 2014; 14: 363-380.

22. Chen J, Zhou X, Li F, Spiegelman D. swdpwr: A SAS macro and an R package for power calculation in stepped wedge cluster randomized trials. *ArXiV* 2020: arxiv:2011.06031v1.

23. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019; 38: 2074-2102.

24. Kammer M. *looplot: Create nested loop plots*. 2022. R package version 0.5.0.9002.

25. Qaqish BF. A family of multivariate binary distributions for simulating correlated binary variables. *Biometrika* 2003; 90: 455-463.

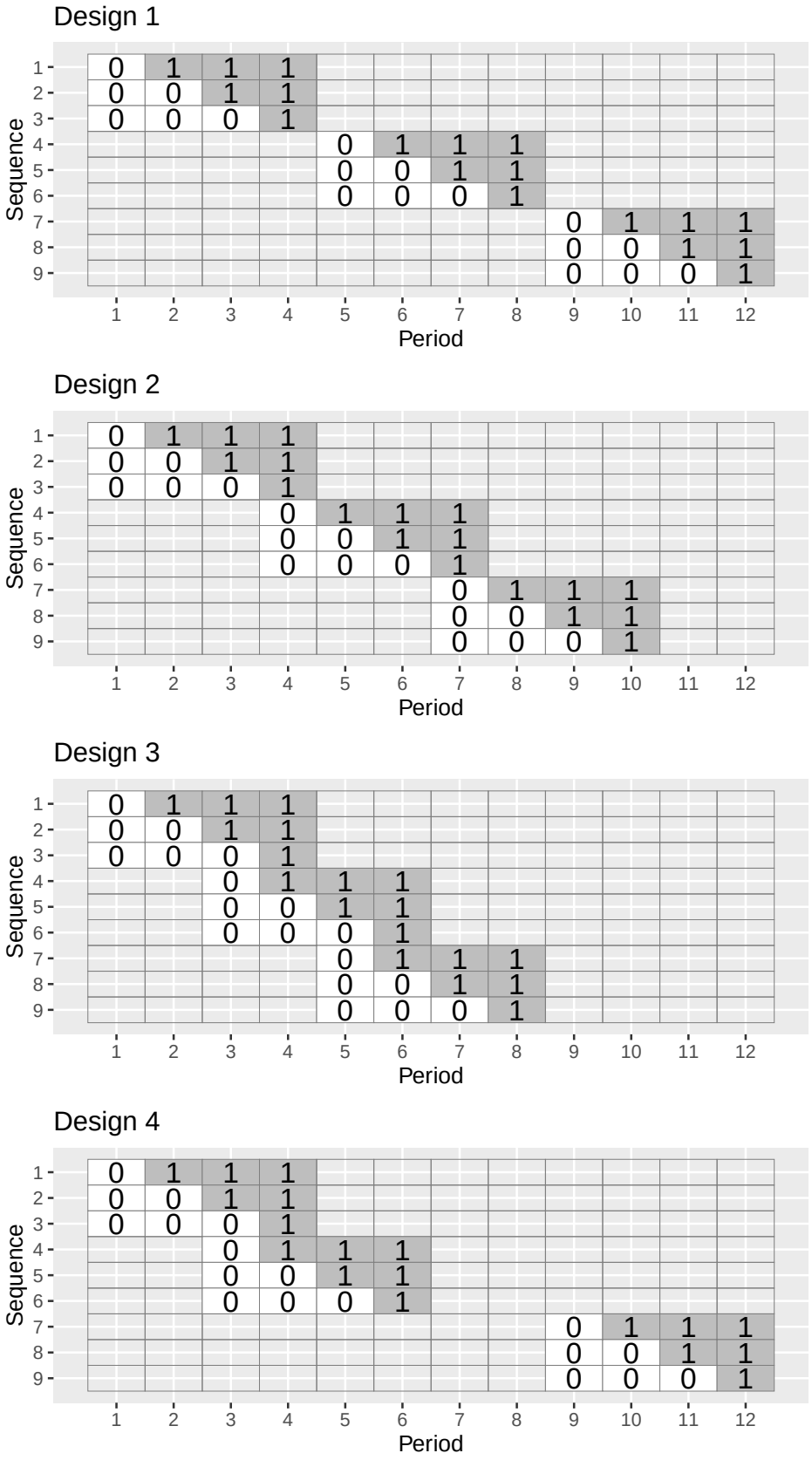

FIGURE 3 Four examples of batched stepped wedge designs with identical component designs (0 indicates control periods; 1 indicates intervention periods). Each of these designs has three batches of three-period stepped wedge designs, with differing degrees of overlap between successive batches. Design 1 (top row): no overlap between successive batches; Design 2 (second from top): overlap of one period between successive batches; Design 3 (second from bottom): overlap of two periods between successive batches; Design 4 (bottom row): variable overlap between successive batches.

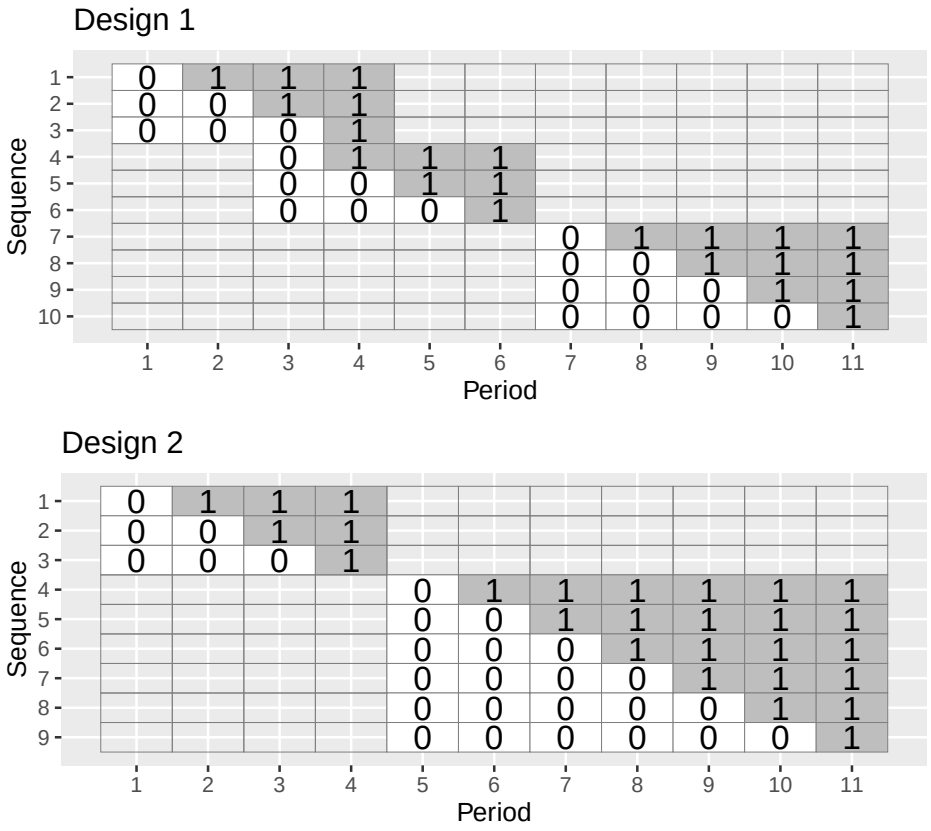

FIGURE 4 Two examples of batched stepped wedge designs without identical component designs (0 indicates control periods; 1 indicates intervention periods).

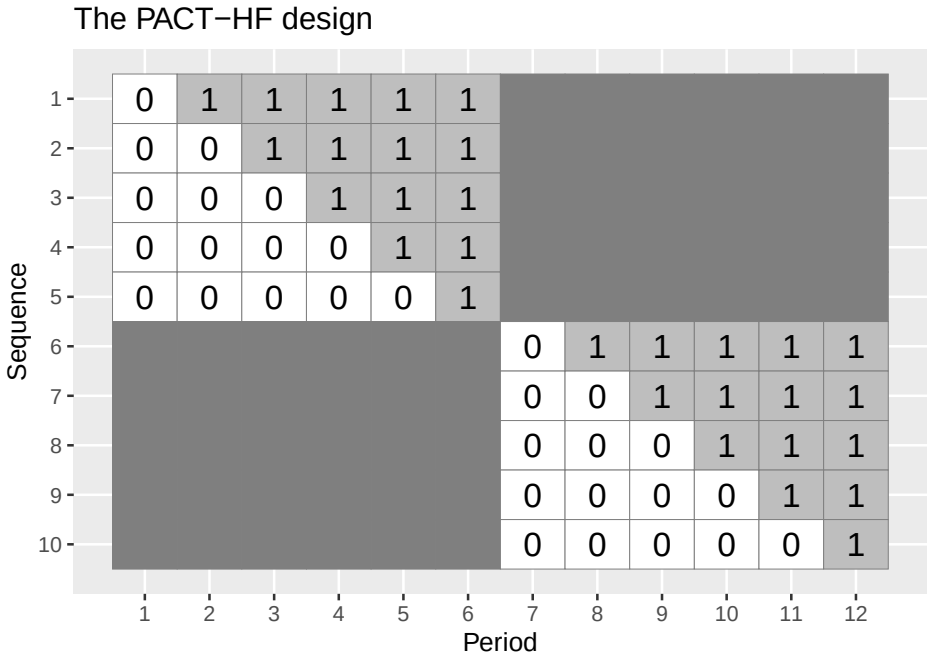

FIGURE 5 The design schematic for the PACT-HF trial: two batches of a 5-sequence stepped wedge design.

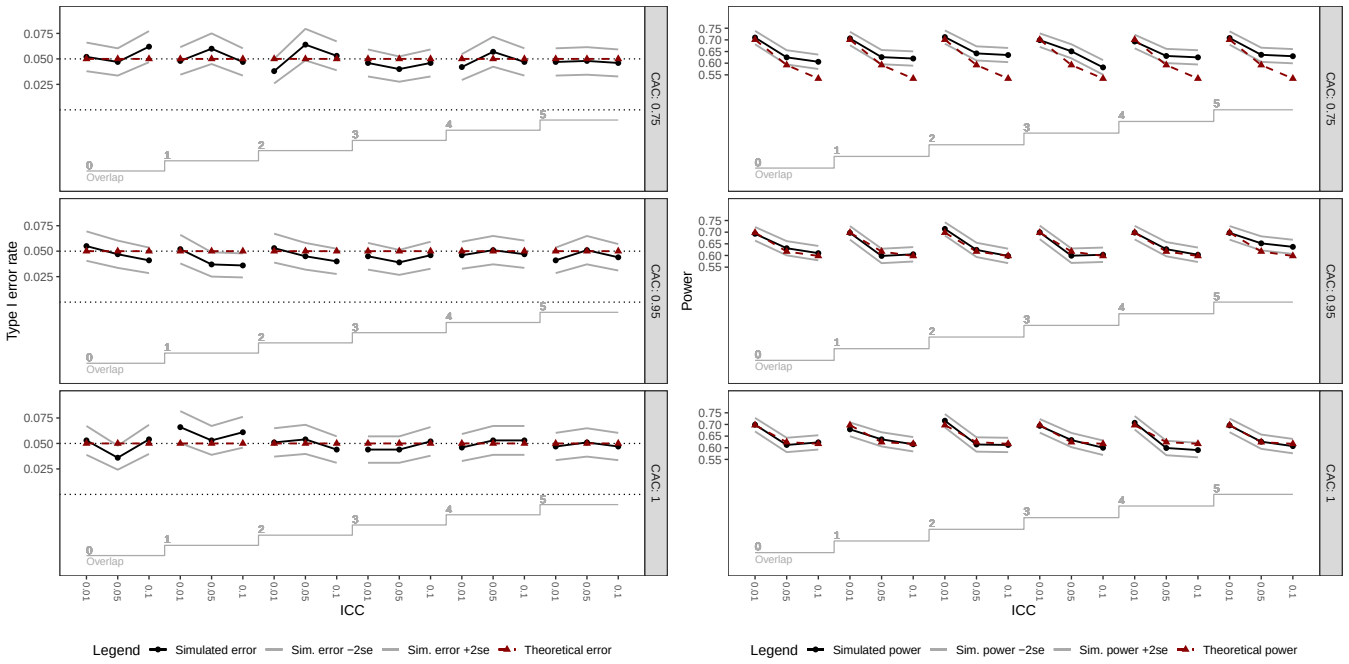

FIGURE 6 Empirical and theoretical type I error rates (left panel) and power (right panel) for the simulated continuous outcomes. ICC=intracluster correlation; CAC = cluster autocorrelation. Within each panel, sub-panels correspond to a different value of the CAC. The theoretical and empirical Type I error rate or power is displayed for each combination of number of periods of overlap, ICC, and CAC, with the empirical result plus and minus 2 standard errors also displayed.

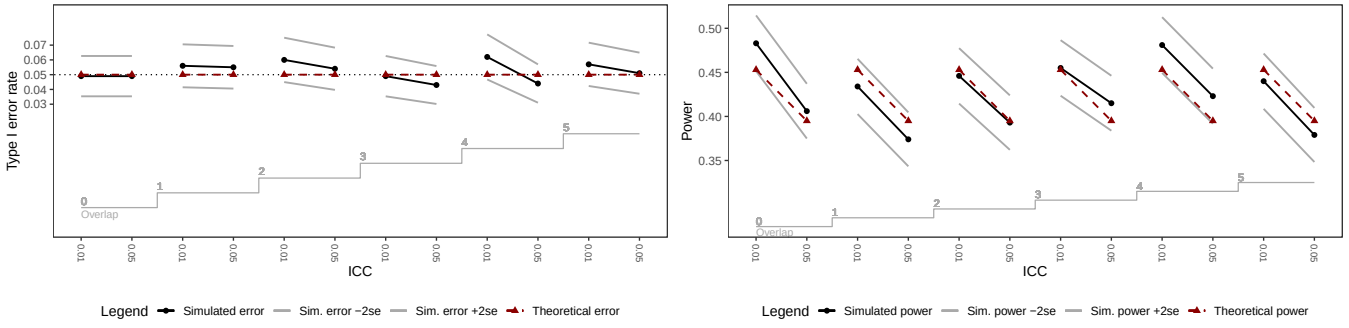

FIGURE 7 Empirical and theoretical type I error rates (left panel) and power (right panel) for the simulated binary outcomes analysed via GEE. ICC=intracluster correlation. The theoretical and empirical Type I error rate or power is displayed for each combination of number of periods of overlap and ICC, with the empirical result plus and minus 2 standard errors also displayed.

Jessica Kasza
Monash University
Melbourne, Australia

April 13, 2022

Professor N. Stallard
Editor, Statistics in Medicine

Dear Professor Stallard,

Thank you for the opportunity to revise our manuscript (SIM-21-0997, "The batched stepped wedge design: a design robust to delays in cluster recruitment") and resubmit to Statistics in Medicine. We also thank the Reviewers for their useful suggestions, and provide point-by-point responses to each comment below. Our main changes are as follows:

- Our Results (Result 1, 2 and 3), formerly only included in the Appendix, are now directly included in the main paper (with the proofs remaining in the Appendix).
- We now emphasise the fact that our key result is not dependent on a specific model for the outcomes. That is, provided that separate time effects are included for each batch, the variance of the treatment effect estimator for the batched stepped wedge is a weighted combination of the variances of the treatment effect estimators from the component stepped wedges: this holds for models within the generalised linear model family.
- The inclusion of a simulation study, for which we also provide R code, available at `https://github.com/jkasza/BatchSW`.

Please note that we have also updated the code in Section 2 of the Appendix to align with version 1.6 of the `swdpwr` R package. We have ensured that our manuscript satisfies all style guidelines. Each comment is reproduced below with our responses beneath; text added to our manuscript is written in italics. In addition to a clean revised version of the manuscript, we have also uploaded a tracked-changes version to the submission system as additional material for review. In that version of the manuscript, additions to the text are italicized and deletions are struck through.

Kind regards,
Jessica Kasza

Reviewer 1 Comments

This very well-written paper introduces a variant of the stepped wedge design, the batched stepped wedge design, that increases flexibility in study design by allowing an investigator to launch a stepped wedge study in stages (batches). Importantly, the authors provide relatively straightforward adaptations to existing methods to determine power in such trials. The paper is succinct and clear and hints at areas for future research, particularly adaptive designs and interim analyses that might reasonably occur at the end of one of the batches.

The authors mention that the batched SW can be thought of as a type of incomplete SW design. It might be useful, if possible, to talk in the discussion how the principles laid out in this paper for finding the treatment effect variance might generalize to other types of incomplete stepped wedge designs with repeating units. For instance, could we do a "batched dog-leg" design and add up the variances of the individual pieces?

- We thank the Reviewer for their comments and careful reading of our paper. In response to the question about the applicability of our results to other types of designs, we have added the following to the Discussion section:
"*Our key results have broad applicability. They generalise to batches of any other type of longitudinal cluster randomised trial design. and do not rely on the design type. For example, our results apply to a "batched dog-leg" design. Provided that separate period effects are included for each batch, the variance of such a design would have the form given in Sections 3 and 4: summing over the variances of treatment effect obtained for each of the individual component designs. Further, our key results do not depend on the precise form of the variance of the treatment effect estimator for each batch. The models considered in the Results could be extended to allow for closed or open cohorts, treatment effect heterogeneity, etc. The key result only requires that separate period effects are included in the outcome model for each batch: if this is the case, then the variances of the treatment effect estimators for each batch can be combined according to Results 1, 2, or 3 as appropriate.*"

Minor Comments

1. In the legend for figure 2, explicitly say that the betas parameterize the time effect.
    - The first two sentences in the legend for Figure 2 now read (additions in italics): "Three different ways in which the effect of time can be parameterised in a design where clusters commence study participation in three batches*: the βs parameterise the time effects. For example, in the top panel, β_1 parametrises the effect of month 1 on outcomes.*"
2. Page 7, line 21 (and elsewhere) talks about the marginal variance of μ_{bkti}. But I think you mean the marginal variance of Y_{bkti}, since parameters dont have a variance.
    - We have corrected this throughout Section 4 and the Appendix.
3. Page 7, line 39 uses the notation $P(Y_{bkti} = 1)$ but (in the context of binary outcomes) that is equal to μ_{bkti}, correct? I wasnt sure why you changed notation in that one

spot?

- This was intended to ensure clarity around the interpretation of μ_{bkti}. We have now generalised these results; they now apply to any outcome type. Please see the updated version of Section 4.

4. Page 8, line 11. I think this is the first time the Shiny CRT calculator is mentioned ... give a reference or link.
   - We now provide a reference the Shiny CRT tutorial paper, and the paper describing the Stata `steppedwedge` program at this location in the manuscript.
5. Page 8, line 21. Perhaps say "the **R** swdpwr package" to give context.
   - The suggested change has been made.

Reviewer 2 Comments

In this work, Kasza et al proposed the batched stepped wedge design and presented sample size methods for this new type of design with both a continuous and binary outcome. Specific considerations for time effects parameterization were developed, and an example sample size calculation has been provided. Overall, I think this is a very innovative idea and represents a useful addition to the current literature on stepped wedge designs. I have a few comments below.

- We thank the Reviewer for their comments and suggestions. We respond to each in turn below.

1. The innovation of this paper/idea comes from the theoretical results the author derived, but unfortunately they are hidden in the appendix. I feel it would be better if the authors can move Result 1 and Result 2 to the main paper (leaving the proof of Result 1 in the appendix), and provide discussions around these interesting technical results, and how they advance our current knowledge on sample size calculation for stepped wedge designs.
   - We have now moved our key results from the appendix to the main paper (leaving the proofs in the Appendix). To facilitate this, we have split Result 1 into two results (the first corresponding to the batched design with identical batches; the second to the setting of non-identical batches). This has led to several additions to Sections 3 and 4 of the paper, and given the extensive nature of these changes, we do not replicate them here. In the tracked changes version of the manuscript changes are indicated by italics.
2. The cluster-period mean parameterization (2) is useful when each batch of the batched stepped wedge design is identical, as the author has indicated. But when each batch of the batched stepped wedge design is not identical, why is this model not adequate? Also, it is currently a bit hard to tell, from the dense writing style, whether this model can be applied to all cases the authors have mentioned (e.g. open-cohort designs, different types of covariance structures, unequal cluster-size). Some clarification on when and when not to use this model for deriving sample size results is helpful.
   - We had not intended to suggest that the cluster-period mean parameterisation is

not adequate when the batches are not identical. Indeed, collapsing to cluster-period means is appropriate whenever the the vector of cluster-period means forms a sufficient statistic for θ. To clarify this we have changed the following sentence from the first paragraph of Section 3.2 from:
"The model for outcomes can then be written as in Equation 1 and that for cluster-period means can be written as in Equation 2, but different batches will no longer have the same variance matrix for outcomes (that is, there will be no shared $var_0(\hat{\theta})$)."
to
"The model for outcomes can then be written as in Equation 1 and that for cluster-period means can be written as in Equation 2. *However, when batches are non-identical,* different batches will no longer have the same variance matrix for outcomes (that is, there will be no shared $var_0(\hat{\theta})$)."

- In the last two paragraphs of Section 3, we mention that Model 1 (the model for the individual-level outcomes) can be extended to allow for closed or open cohorts, treatment effect heterogeneity, and unequal cluster sizes. The proof of our main results does not depend on the precise form of the variance matrix for each batch of the design, so applies to all of these potential variations to Model 1. We clarify the generalisability of our result through the addition of the following paragraph to the Discussion (please note that this is as in our response to Reviewer 1's first point):
"*Our key results have broad applicability. They generalise to batches of any other type of longitudinal cluster randomised trial design. and do not rely on the design type. For example, our results apply to a "batched dog-leg" design. Provided that separate period effects are included for each batch, the variance of such a design would have the form given in Sections 3 and 4: summing over the variances of treatment effect obtained for each of the individual component designs. Further, our key results do not depend on the precise form of the variance of the treatment effect estimator for each batch. The models considered in the Results could be extended to allow for closed or open cohorts, treatment effect heterogeneity, etc. The key result only requires that separate period effects are included in the outcome model for each batch: if this is the case, then the variances of the treatment effect estimators for each batch can be combined according to Results 1, 2, or 3 as appropriate.*"

3. For binary outcomes, authors have considered GEE as basis for sample size calculation (why not generalized linear mixed models?), but the working correlation structure does not seem to be sufficiently explained? This has been explained under the linear mixed model setup, but what are appropriate correlation models for stepped wedge trials with a binary outcome? Do the current result 2 also apply to count outcome, and other choices of the link function?

   - In response to the question "why not generalized linear mixed models?", we note that generalized linear mixed models for binary outcomes are discussed in the first paragraph of Section 4, where we note that the result from Section 3 will apply for GLMMs when separate period effects are included for each batch. To further explain the working correlation structure in the GEE framework, we

now include the following at the start of Section 4:
"*When the GEE approach is used, a working correlation matrix structure must be assumed. This working correlation structure describes the pattern of within-cluster correlations; an exchangeable correlation structure would imply equal correlations between all observations in a cluster, for example. As discussed in Li et al. (2018), when GEE is the intended analysis approach, power calculations can proceed via generalized least squares. We now state the main result for this scenario.*"

- A generalised version of what was previously referred to as Result 2 (now Result 3) does indeed apply for other outcome types and link functions: the Result that appears in Section 4 is now generalised to apply to all outcome types and link functions in the generalized linear model family. Essentially, when separate period terms are included in the model for each batch, the estimates of the treatment effects from each batch of the design are independent, and the estimator weights the estimate from each batch by its variance. We have clarified this in the paper through the inclusion of this more general result and the addition of the following to the end of Section 4:
"*Result 3 applies not only to binary outcomes analysed with a logit link function; the proof given in the Appendix does not rely on the choice of link function or the outcome type. Thus, Result 3 holds for binary outcomes with a linear link function, or for count outcomes with a log link function, to name just two alternatives.*"

4. For equation (8), is it derived under the identity link function, or the logistic link function? If it is the latter, then there seems to be an error in this equation. This is because the variance involves differentiating the mean function with respect to the parameters (e.g., $D = d\mu/d\theta = \mu(1-\mu) \times X$ for logistic link), and the form is slightly different from (8). Please double check to see if the expression coincides what is in Liang and Zeger (1986).

    - We have now generalised this result to apply to any choice of link function, so please see the updated version of the manuscript.

5. Does result 2 also hold for any other type of outcomes (count) and other choice of link function (e.g log link)? If so, it should be stated more generally. If not, then the authors should explain why.

    - This result does indeed hold more generally: the result now quoted in the manuscript is now a more general version.

6. The paper, as it currently writes, is a little thin without additional numerical/simulation results to support the main arguments. For example, the batched stepped wedge designs requires quite many time effects parameters (if either B or T is large), compared to standard stepped wedge designs that commence at the same time. So the authors should consider including a Section on Monte Carlo simulations to examine whether the proposed approach to calculate sample size is precise with so many time effects parameters in the model (of course when the analysis is also done this way). This may also help inform when the approach may break down and alternative time effects parameterization is needed. Such results will inevitably make the paper's results much

more convincing, and at the same time helps verify the derivations are all correct.

- We now include a simulation study in Section 6 of the updated version of the manuscript. In this simulation study we consider both binary and continuous outcomes and compare simulated power and type I error rates to the theoretical predictions. Simulated power and type I error rates align with the theoretical values. The key takeaway from this simulation study is that for both continuous and binary outcomes, the power and type I error rates in the simulations do not depend on the degree of overlap between successive batches, as suggested by our theoretical derivations. Code to replicate this simulation study is available at `https://github.com/jkasza/BatchSW`.