Received: 1 June 2021 | Revised: 2 November 2021 | Accepted: 9 December 2021

DOI: 10.1002/sim.9301

RESEARCH ARTICLE

Statistics in Medicine WILEY

Power analysis for stepped wedge trials with multiple interventions

Phillip Sundin | **Catherine M. Crespi**

Department of Biostatistics, University of California Los Angeles (UCLA), Los Angeles, California

Correspondence
Phillip Sundin, Department of Biostatistics, University of California Los Angeles (UCLA), Los Angeles, CA, USA.
Email: phillip1492@ucla.edu

Funding information
PCORI PCS-201C1-6482, PI: Bastani

Stepped wedge design (SWD) trials are cluster randomized trials that feature staggered, unidirectional cross-over between treatment conditions. Existing literature on power for SWDs focuses primarily on designs with two conditions, typically a control and an intervention condition. However, SWDs with more than one treatment condition are being proposed and conducted. We present a linear mixed model for SWDs with two or more interventions, including both multiarm and factorial designs. We derive standard errors of the intervention effect coefficients, and present power calculation methods. We consider both repeated cross-sectional and cohort designs. Design features, with a focus on treatment allocations, are examined to determine their impact on power.

KEYWORDS
cluster randomized trial, factorial design, multiarm randomized trial, multiple comparison, power calculation, sample size calculation, stepped wedge design

1 | INTRODUCTION

Cluster randomized trials (CRTs) are trials in which groups of individuals, called clusters, are randomized to treatment conditions. CRTs can use parallel designs, in which clusters receive only one treatment condition, or crossover designs, in which clusters are assigned to a sequence of treatment conditions. A variation on the CRT crossover design is the stepped wedge design (SWD).[1] The most common SWDs involve unidirectional crossover in which clusters transition from a control condition to an intervention condition at different prespecified times. Clusters are randomized to the predetermined sequences.

SWDs have several advantages over parallel and crossover CRT designs. SWDs allow comparisons both within cluster and across cluster, which can yield efficiency gains.[2] It may be less costly and logistically easier to roll out the intervention over time instead of all at once, as would occur in many parallel CRTs.[3] Guaranteeing receipt of an intervention for all clusters may make clusters more willing to participate or alleviate ethical concerns.[2]

Most research on the design and analysis of stepped wedge trials has focused on SWDs with one intervention condition contrasted with a control condition. There is a small but growing body of literature on SWDs with more than one intervention condition. Grayling et al[4] focused on studies in which there is a nested natural order of D interventions such that intervention d consists of intervention $d-1$ plus some additional factor. The authors discuss the optimization of treatment sequence allocations and focus on optimal design for such trials. The variance of treatment effect estimates in SWDs with nested interventions has also been studied.[5] However, SWDs with multiple interventions that are not nested within one another have not been well studied, and interaction effects also have not received much attention.[6]

SWDs with multiple treatment arms are being conducted despite a scarcity of methodological literature. There are several examples of stepped wedge design trials that feature two interventions implemented alone and in combination, as

in a 2 × 2 factorial design. These examples include a trial of the comparative effectiveness of two interventions to promote human papillomavirus vaccination among adolescents,[7] a study examining two interventions for reducing hyperbilirubinaemia in infants,[8] and a study compared two interventions for addressing behavioral problems in children with cerebral palsy.[9] In these studies, clusters were assigned to sequences that could include periods spent in usual care, the two single intervention conditions, and/or a combined condition.

There are also examples in the literature of several related single-intervention stepped wedge trials conducted simultaneously. The FallDem study used two stepped wedge trials to examine two interventions for improving the lives of dementia patients,[6,10] and Durovni et al conducted two separate SWD trials for tuberculosis screening.[11,12] In some cases, it might be advantageous to combine two separate trials with stepped wedge designs into one trial with multiple treatment conditions, akin to a multiarm trial.

In this article, we consider stepped wedge design trials with more than one intervention, including both multiarm designs, which involve a control and two or more treatment conditions, and factorial designs, in which interventions are implemented alone and in combination. Multiarm trials have several advantages, such as allowing for direct comparison of alternative treatments (comparative effectiveness) and resource savings due to "reusing" the same control condition to compare to several interventions. Factorial designs also have potentially increased efficiency and can allow for the estimation of interaction effects.[13] Thus extending SWDs to incorporate multiarm and factorial design features could be quite beneficial. We develop power analysis methods for such trials and examine factors that influence power for stepped wedge designs with a normally distributed outcome variable.

The article is organized as follows. Section 2 introduces the models for the SWD with multiple treatment conditions and develops power analysis methods. Section 3 uses examples to examine the influence of different design features on power. Section 3 also presents results from a simulation study. Section 4 discusses the implications of our work, possible extensions, limitations and future work.

2 | METHODS

We first present a model for a stepped wedge design with a single intervention and then consider designs with any number of interventions. We focus on designs with only two interventions and an interaction effect as designs with more than two interventions have not yet been seen in practice. The section concludes with an overview of the derivation of the standard errors of the estimated treatment effect coefficients, with details in Appendix S1.

2.1 | Model specification

We begin with the classic stepped wedge design model with a single binary treatment factor.[1] For a design with I clusters observed at T times, and N different individuals per time per cluster, let Y_{ijk} be a continuous outcome for individual k in cluster i at time j. The model for Y_{ijk} is

$$Y_{ijk} = \mu + \alpha_i + \psi_{ik} + \nu_{ij} + \beta_j + X_{ij}\theta_1 + e_{ijk}, \tag{1}$$

where μ is an intercept, $\alpha_i \sim N(0, \sigma_\alpha^2)$ is a random intercept for cluster i, $\psi_{ik} \sim N(0, \sigma_\psi^2)$ is a random intercept for individual k in cluster i, $\nu_{ij} \sim N(0, \sigma_\nu^2)$ is a random intercept for cluster i in time j, β_j is a fixed effect for time j, X_{ij} is a {0,1} indicator for whether cluster i at time j receives treatment, θ_1 is the treatment effect, and $e_{ijk} \sim N(0, \sigma_e^2)$. The total variance of an individual level outcome is $\sigma_y^2 = \sigma_\alpha^2 + \sigma_\psi^2 + \sigma_\nu^2 + \sigma_e^2$.

It is straightforward to expand this model to include multiple binary treatment factors.[6] Assuming additive treatment effects, the model with R treatment factors is

$$Y_{ijk} = \mu + \alpha_i + \psi_{ik} + \nu_{ij} + \beta_j + \sum_{r=1}^{R} X_{ijr}\theta_r + e_{ijk}, \tag{2}$$

where X_{ijr} is a {0,1} indicator of whether cluster i at time j receives treatment r and θ_r is the treatment effect for treatment r. For the remainder of this section, we take $R = 2$ for simplicity, with results generalizable to $R > 2$. Adding an interaction

effect θ_3, the model becomes

$$Y_{ijk} = \mu + \alpha_i + \psi_{ik} + \nu_{ij} + \beta_j + X_{ij1}\theta_1 + X_{ij2}\theta_2 + X_{ij1}X_{ij2}\theta_3 + e_{ijk}. \quad (3)$$

Individual auto-correlation (IAC) is defined as the proportion of the individual-level variance (which in this model is $\sigma_\psi^2 + \sigma_e^2$) that is time-invariant. In model (3), the IAC is $\pi = \sigma_\psi^2/(\sigma_\psi^2 + \sigma_e^2)$. Setting $\pi = 0$ yields a repeated cross-sectional design. We can also define the cluster auto-correlation (CAC) as the proportion of cluster level variance that is time-invariant. In this model, the cluster-level variance is $\sigma_\nu^2 + \sigma_\alpha^2$ and CAC $= \sigma_\alpha^2/(\sigma_\nu^2 + \sigma_\alpha^2) = \rho_a/\rho_w$.[14,15] We also define two intraclass correlation (ICC) values. The within-period ICC, $\text{Corr}(y_{ijk}, y_{ijk'})$ is now $\rho_w = (\sigma_\nu^2 + \sigma_\alpha^2)/\sigma_y^2$ and the across-period ICC, $\text{Corr}(y_{ijk}, y_{ij'k'})$, is $\rho_a = \sigma_\alpha^2/\sigma_y^2$.

Standard errors are needed to compute power. To derive standard errors, it is convenient to work with cluster-level outcomes. Let $\overline{Y}_{ij\cdot} = \frac{1}{N}\sum_{k=1}^{N} Y_{ijk}$ be the mean outcome of cluster i at time j across N individuals. The model for cluster-period means with two treatments and an interaction term is

$$\overline{Y}_{ij\cdot} = \mu + \alpha_i + \psi_i + \nu_{ij} + \beta_j + X_{ij1}\theta_1 + X_{ij2}\theta_2 + X_{ij1}X_{ij2}\theta_3 + e_{ij\cdot}, \quad (4)$$

where $e_{ij\cdot} = \frac{1}{N}\sum_{k=1}^{N} e_{ijk} \sim N(0, \sigma_c^2 = \frac{\sigma_e^2}{N})$ and $\psi_i = \frac{1}{N}\sum_{k=1}^{N}\psi_{ik} \sim N(0, \sigma_\zeta^2 = \frac{\sigma_\psi^2}{N})$. In this model, the variance of a cluster-period mean is $\text{Var}(\overline{Y}_{ij\cdot}) = \sigma_c^2 + \sigma_\alpha^2 + \sigma_\zeta^2 + \sigma_\nu^2$, and $\text{Cov}(\overline{Y}_{ij\cdot}, \overline{Y}_{ij'\cdot}) = \sigma_\alpha^2 + \sigma_\zeta^2$.

Define the outcome vector $\boldsymbol{Y} = (\overline{Y}_{11\cdot}, \ldots, \overline{Y}_{iT\cdot}, \ldots, \overline{Y}_{I1\cdot}, \ldots, \overline{Y}_{IT\cdot})'$. Assuming clusters are independent, the variance-covariance matrix of $\boldsymbol{Y}$ is a $IT \times IT$ matrix of the form

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V_1} & 0 & 0 & 0 \\ 0 & \boldsymbol{V_2} & \ldots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \ldots & \boldsymbol{V_I} \end{bmatrix},$$

with each $T \times T$ matrix $\boldsymbol{V_i}$ having structure

$$\boldsymbol{V_i} = \begin{bmatrix} \sigma_c^2 + \sigma_\alpha^2 + \sigma_\nu^2 + \sigma_\zeta^2 & \sigma_\alpha^2 + \sigma_\zeta^2 & \ldots & \sigma_\alpha^2 + \sigma_\zeta^2 \\ \sigma_\alpha^2 + \sigma_\zeta^2 & \sigma_c^2 + \sigma_\alpha^2 + \sigma_\nu^2 + \sigma_\zeta^2 & \ldots & \sigma_\alpha^2 + \sigma_\zeta^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 + \sigma_\zeta^2 & \sigma_\alpha^2 + \sigma_\zeta^2 & \ldots & \sigma_c^2 + \sigma_\alpha^2 + \sigma_\nu^2 + \sigma_\zeta^2 \end{bmatrix}.$$

Some practitioners find that standardization of the model can be convenient for power calculations. To standardize the model in (3), one divides through by σ_y. The cluster random intercept α_i now has standardized variance ρ_a, the cluster-by-time random intercept ν_{ij} has variance $\rho_w - \rho_a$ for $\rho_w > \rho_a$, the individual-level random intercept ψ_{ik} has variance $\pi(1 - \rho_w)$ and the error term e_{ijk} has variance $(1 - \pi)(1 - \rho_w)$. Thus the variances can be specified in terms of the parameters ρ_w, ρ_a and π. The matrix $\boldsymbol{V_i}$ will have diagonal elements $\rho_w + \frac{(1-\rho_w)}{N}$ and off-diagonal elements $\rho_a + \frac{\pi(1-\rho_w)}{N}$.

Now we turn to the design matrix of the fixed effects. While β_1 rather than β_T is often set equal to zero when the model is fit to data, we follow[1] and set $\beta_T = 0$ for identifiability. The choice is immaterial for power calculations. The $(T + 3) \times 1$ regression coefficient vector for the fixed effects is

$$\boldsymbol{\eta} = \begin{bmatrix} \mu & \beta_1 & \ldots & \beta_{T-1} & \theta_1 & \theta_2 & \theta_3 \end{bmatrix}'.$$

The full $IT \times (T + 3)$ design matrix $\boldsymbol{Z}$ becomes

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z_1} \\ \boldsymbol{Z_2} \\ \vdots \\ \boldsymbol{Z_I} \end{bmatrix},$$

where each matrix $\boldsymbol{Z_i}$ has dimension T × (T+3) and takes the form

$$\boldsymbol{Z_i} = \begin{bmatrix} \boldsymbol{1_T} & \boldsymbol{I_{T-1}} & \boldsymbol{X_{i1}} & \boldsymbol{X_{i2}} & \boldsymbol{(X_1X_2)_i} \\ & \boldsymbol{0'_{T-1}} & & & \end{bmatrix}.$$

The elements of the vector $\boldsymbol{X_{i1}} = (X_{i11}, X_{i21}, \dots, X_{iT1})'$ are indicators of whether cluster i at time j receives treatment 1, the elements of $\boldsymbol{X_{i2}} = (X_{i12}, X_{i22}, \dots, X_{iT2})'$ are indicators of receipt of treatment 2, and $\boldsymbol{(X_1X_2)_i}$ is the Hadamard product of $\boldsymbol{X_{i1}}$ and $\boldsymbol{X_{i2}}$, with a value of 1 if cluster i receives both treatments at time j and 0 otherwise. The matrix $\boldsymbol{I_{T-1}}$ contains indicators for each time j from $1, \dots, (T-1)$. The vector $\boldsymbol{0'_{T-1}}$ corresponds to time T. For designs with $R > 2$, $\boldsymbol{Z_i}$ can be expanded to include the additional indicators.

2.2 | Power analysis

Inference for fixed effects in linear mixed models can be conducted using Wald tests or likelihood ratio tests. We focus on Wald tests. For hypotheses of the form $H_0 : \eta = 0$, where η is a fixed effects coefficient, the Wald test statistic takes the form $\hat{\eta}/\sqrt{\text{Var}(\hat{\eta})}$, where $\hat{\eta}$ is the estimated coefficient, and has an approximate standard normal distribution when the null hypothesis is true.[16] The power to reject H_0 for a specific true value of η, denoted η_a, with Type I error rate α and a two-sided test, is approximately

$$P\left(\left|\frac{\eta_a}{\sqrt{\text{Var}(\hat{\eta})}}\right| \geq z_{1-\frac{\alpha}{2}} \,\middle|\, \eta = \eta_a\right),$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$th percentile of the standard normal distribution.

To calculate power, we need an expression for $\text{Var}(\hat{\eta})$. We derive expressions for $\text{Var}(\hat{\eta})$ using the cluster-period mean models in (4). We focus on models with $R = 2$ treatments and an interaction term assuming a factorial design; the results are generalizable to $R > 2$ and multiarm trials as discussed in Section 4. Given the linear mixed model formulation, the variance-covariance matrix of the estimated fixed effect coefficients has the form $\boldsymbol{C} = (\boldsymbol{Z'V^{-1}Z})^{-1}$, where $\boldsymbol{Z}$ is the fixed effects design matrix and $\boldsymbol{V}$ is the variance-covariance matrix of the outcome vector. Our approach is to find expressions for the variances and covariances of treatment effect coefficient estimates, $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$. We do so by calculating $\boldsymbol{Z'V^{-1}Z}$ then invert it to get the elements of $(\boldsymbol{Z'V^{-1}Z})^{-1}$, corresponding to the variances and covariances of the treatment effect coefficients.

Let $\boldsymbol{Z}$ be the $IT \times (T+3)$ design matrix and $\boldsymbol{V}$ be the $IT \times IT$ variance-covariance matrix of the cluster-level outcomes. Let $\sigma_{diag} = \sigma_c^2 + \sigma_v^2$ and $\sigma_{off} = \sigma_\alpha^2 + \sigma_\zeta^2$, and for standardized models, $\sigma_{diag} = \frac{(1-\pi)(1-\rho_w)}{N} + \rho_w - \rho_a$ and $\sigma_{off} = \rho_a + \frac{\pi(1-\rho_w)}{N}$. Assuming clusters are independent, $\boldsymbol{V}$ has block diagonal structure with elements $\boldsymbol{V_i} = \sigma^2_{diag}\boldsymbol{I_T} + \sigma^2_{off}\boldsymbol{1_T1'_T}$, where $\boldsymbol{I_T}$ is a $T \times T$ identity matrix and $\boldsymbol{1_T}$ is a $T \times 1$ vector of 1's. Using the Sherman-Morrison formula,[17,18] we can obtain its inverse as

$$\boldsymbol{V_i^{-1}} = \frac{1}{\sigma^2_{diag}\left(\sigma^2_{diag} + T\sigma^2_{off}\right)}\left[\left(\sigma^2_{diag} + T\sigma^2_{off}\right)\boldsymbol{I_T} - \sigma^2_{off}\boldsymbol{1_T1'_T}\right].$$

This matrix has off-diagonal elements $\frac{-\sigma^2_{off}}{\sigma^2_{diag}(T\sigma^2_{off}+\sigma^2_{diag})}$ and diagonal elements $\frac{(T-1)\sigma^2_{off}+\sigma^2_{diag}}{\sigma^2_{diag}(T\sigma^2_{off}+\sigma^2_{diag})}$. Due to the block diagonal structure of $\boldsymbol{V}$, we have

$$\boldsymbol{Z'V^{-1}Z} = \sum_{i=1}^{I} \boldsymbol{Z'_iV_i^{-1}Z_i},$$

where $\boldsymbol{Z_i}$ is the $T \times (T+3)$ part of the design matrix corresponding to cluster i. We can then rewrite

$$\boldsymbol{Z'_iV_i^{-1}Z_i} = \frac{1}{\sigma^2_{diag}\left(\sigma^2_{diag} + T\sigma^2_{off}\right)}\left[\left(\sigma^2_{diag} + T\sigma^2_{off}\right)\boldsymbol{Z'_iZ_i} - \sigma^2_{off}\boldsymbol{Z'_i1_T1'_TZ_i}\right]. \quad (5)$$

We then use block matrix inversion techniques to solve for the submatrix corresponding to the coefficients of interest. A full derivation is provided in Appendix S1.

In the case of a design with $R = 2$ interventions and no interaction term, a closed form solution for the variance of the estimated intervention effects for Intervention 1 and 2 can be calculated using the inverse of a 2×2 matrix, with variances written as

$$Var(\hat{\theta}_1) = \frac{l_2 - z_2 - \frac{y_2^2}{fT} - \frac{1}{f+gT}\left(w_2 - \frac{l_2^2}{T}\right)}{\left(l_2 - z_2 - \frac{y_2^2}{fT} - \frac{1}{f+gT}\left(w_2 - \frac{l_2^2}{T}\right)\right)\left(l_1 - z_1 - \frac{y_1^2}{fT} - \frac{1}{f+gT}\left(w_1 - \frac{l_1^2}{T}\right)\right) - \left(q_1 - \frac{y_1 y_2}{fT} - \frac{1}{f+gT}\left(w_{XW} - \frac{l_1 l_2}{T}\right)\right)^2},$$

$$Var(\hat{\theta}_2) = \frac{l_1 - z_1 - \frac{y_1^2}{fT} - \frac{1}{f+gT}\left(w_1 - \frac{l_1^2}{T}\right)}{\left(l_2 - z_2 - \frac{y_2^2}{fT} - \frac{1}{f+gT}\left(w_2 - \frac{l_2^2}{T}\right)\right)\left(l_1 - z_1 - \frac{y_1^2}{fT} - \frac{1}{f+gT}\left(w_1 - \frac{l_1^2}{T}\right)\right) - \left(q_1 - \frac{y_1 y_2}{fT} - \frac{1}{f+gT}\left(w_{XW} - \frac{l_1 l_2}{T}\right)\right)^2},$$

with all terms defined in Appendix S1. Standard errors are calculated by taking the square root of these variances. Closed form solutions for the model with the interaction effect are found in Appendix S1.

The standard errors thus derived enable power calculations for hypothesis testing. In factorial and multiarm design trials, there will typically be multiple hypotheses of interest. When multiple hypotheses are tested simultaneously, multiplicity adjustments should be taken into account in power analysis to control experimentwise Type I error. If a single-step method such as Bonferroni is used, the power calculations can be adjusted by adjusting the significance level for each test. Accounting for the use of other multiplicity adjustment procedures, such as the Hochberg or fixed sequence procedures, can be more complex.[19]

The calculations also enable the testing of linear contrasts. For example, a comparative effectiveness hypothesis comparing two active treatments may involve the hypothesis $H_0 : \theta_1 - \theta_2 = 0$, which can be tested using the Wald statistic $(\hat{\theta}_1 - \hat{\theta}_2)/\sqrt{\text{Var}(\hat{\theta}_1 - \hat{\theta}_2)}$, where $\text{Var}(\hat{\theta}_1 - \hat{\theta}_2) = \text{Var}(\hat{\theta}_1) + \text{Var}(\hat{\theta}_2) - 2\text{Cov}(\hat{\theta}_1, \hat{\theta}_2)$ and the variance and covariances can be obtained as described.

The power method described makes use of a normality-based z-test, which may not hold up well for a small number of clusters. However, for the examples we present in the next section, there was no evidence of small-sample bias, suggesting that this is not always an issue. The topic of small-sample bias corrections for stepped wedge designs has been explored elsewhere.[20,21]

3 | EXAMPLES

Since the formulas are complex, we present examples to illustrate how power is affected by design features of SWDs, focusing on the impact of sequencing of treatment conditions within clusters. The examples in Sections 3.1 and 3.2 use standardized effect sizes and realistic but arbitrary values of standardized variance parameters. The examples in Sections 3.3 and 3.4 use simple effect sizes (in original units) and variance parameter values derived from a real study. For all examples, we set the experimentwise Type I error rate to 0.05 and use a Bonferroni correction when conducting multiple simultaneous tests within the same design. Calculations were performed in R version 3.6.1[22] with code available at https://github.com/phillipsundin/SWFD.

3.1 | Two separate single-intervention SWDs vs a concurrent SWD

Several studies have conducted two related but separate single-intervention SWD trials.[10,11] We explore potential advantages of combining two single-intervention trials into one trial with two interventions, including efficiency gains and comparative effectiveness.

Consider the two single-intervention SWD trials, each with six clusters and four time periods, in Figure 1A. Figure 1B stacks the two designs into a single 12-cluster trial; such a design has been called a concurrent design.[6] Figure 1C shows a concurrent design with only 10 clusters. Let δ_1 and δ_2 denote the standardized effect sizes for interventions 1 and 2 compared to the control condition. We set $\delta_1 = \delta_2 = 0.4$, representing medium effect sizes.[23] Within each design, power for the two intervention effects is the same due to symmetry. We specify $N = 15$ individuals per cluster-period in a repeated

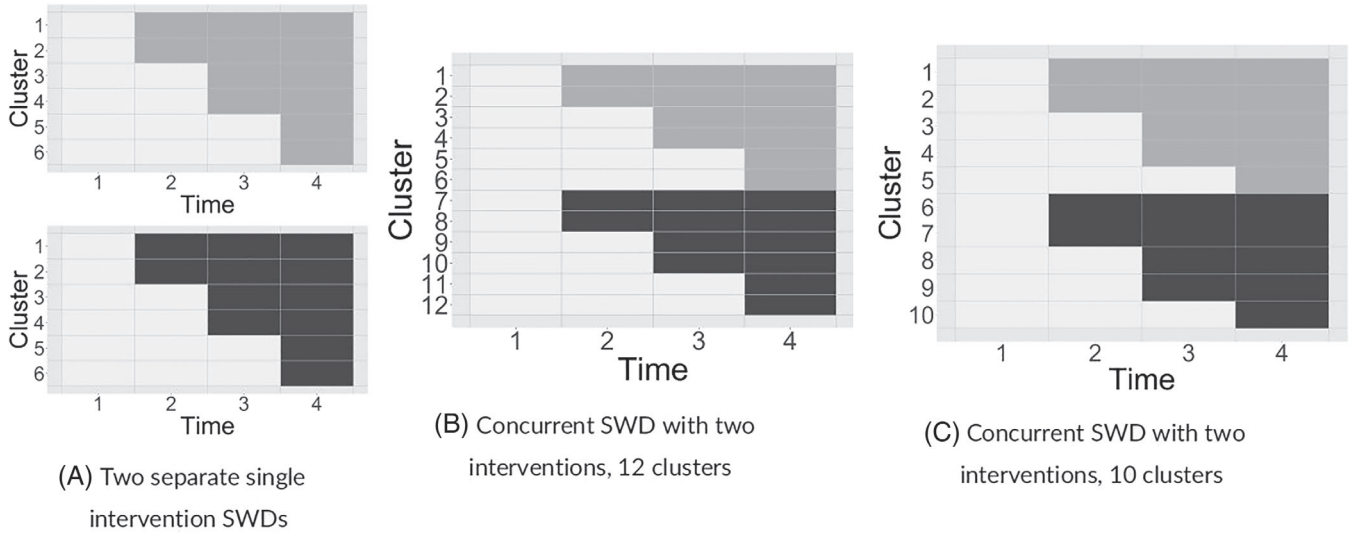

FIGURE 1 Examples of two single-intervention SWDs versus concurrent SWDs with two interventions. White cells indicate cluster-periods in the control condition. Light and dark gray cells indicate treatment conditions for Interventions 1 and 2, respectively

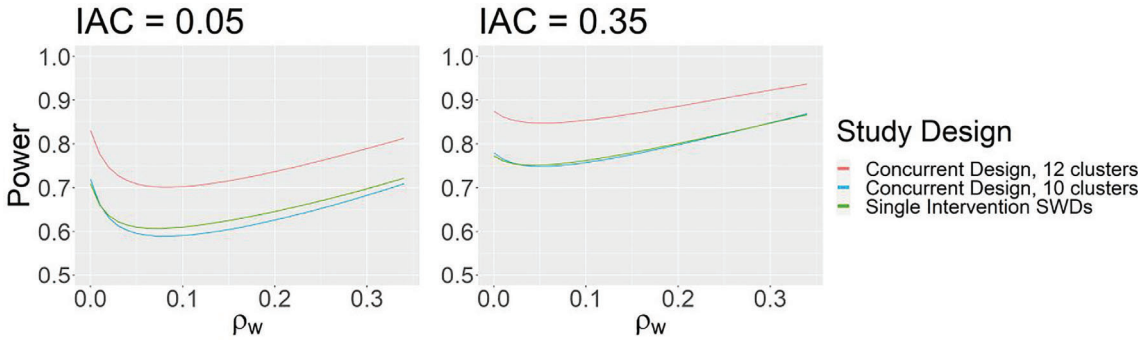

FIGURE 2 Comparison of power for either intervention effect for a single-intervention SWD, 12-cluster concurrent SWD and 10-cluster concurrent SWD for two values of IAC

cross-sectional design. Power for detecting an intervention effect in one of the single-intervention SWDs was calculated using model (1); for the concurrent SWDs, power was calculated using model (2). Type I error was set to 0.05 for hypothesis tests in the single-intervention SWDs and 0.025 for the concurrent designs. In these examples, we fix $\rho_w = \rho_a$, equivalent to setting $\sigma_v^2 = 0$, and examine power under two different values of π.

Figure 2 displays power for either intervention effect for the three designs as a function of ρ_w. The convex shapes of the power curves are similar to those observed for SWDs with only one treatment.[2,24,25] For both values of π, the 12-cluster concurrent design, which maintains the same total number of clusters as the two separate single intervention SWDs, has power gains ranging from 0.10 to 0.13 compared to the other designs for the values of ρ_w considered even with a reduced Type I error rate. The 10-cluster concurrent design, which reduces the total sample size by about 17% compared to the designs with 12 clusters, has power comparable to that of a single-intervention SWD when $\pi = 0.35$; for $\pi = 0.05$, power for the 10-cluster concurrent design is at most 0.02 lower.

Another advantage of including two interventions in one study is the ability to directly compare them. This can be accomplished using tests of linear contrast, which can be powered using our methods. Suppose we assume standardized effect sizes of 0.30 and 0.70 for the two interventions compared to control, entailing a difference of 0.4 between them (a difference this large may be unrealistic for some studies, but helps to illustrate the principle). We set $\rho_w = \rho_a$ and $\pi = 0.05$. Type I error was set to $0.05/3 = 0.0167$ for each of three tests: the two intervention-to-control comparisons and comparison between the two interventions. Figure 3 displays power for the linear contrast as a function ρ_w. The relationship between power and ρ_w for the comparative effectiveness contrast is similar to that for the intervention-to-control hypothesis tests.

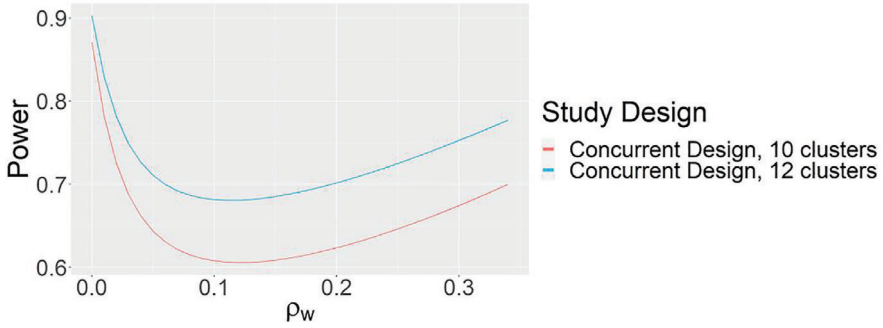

FIGURE 3 Power for comparison of two interventions in the concurrent SWDs

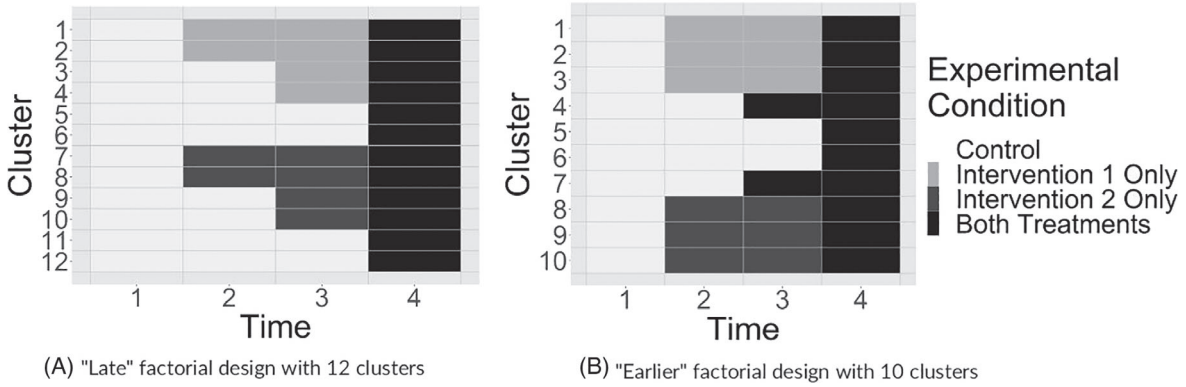

FIGURE 4 Stepped wedge factorial design examples

We note that in a concurrent design, the interventions are conducted in parallel and thus the intervention-to-intervention contrast is less susceptible to confounding by time than the intervention-to-control comparisons.

3.2 | Factorial designs with additive treatment effects

Our methods enable power analysis for factorial designs, which can be highly efficient when effects are additive, that is, when there is no interaction effect. To investigate stepped wedge factorial designs, we begin by comparing the 12-cluster, 4-period concurrent design in Figure 1B with designs that assign some cluster-periods to a combined condition. Figure 4A shows a 12-cluster "late" factorial design in which all clusters transition to the combined condition in the last period. Figure 4B shows an "earlier" factorial design with only ten clusters that introduces the combined condition earlier. Additive intervention effects are assumed for these examples.

Both designs feature six cluster-periods in each single intervention condition and twelve cluster-periods in the combined condition. We assume repeated cross-sectional observations, moderate effect sizes ($\delta_1 = \delta_2 = 0.4$) and $N = 15$ individuals per cluster-period.

Figure 5 compares power for the main effect of each intervention for the three designs (Figures 1B and 4A,B) as a function of ρ_w for two values of the IAC. Power for the two main effects is identical due to symmetry. The 12-cluster "late" factorial design has the lowest power for all values of ρ_w and π. For $\rho_w < 0.02$, the 12-cluster concurrent and 10-cluster "earlier" factorial designs have similar power; for $\rho_w > 0.02$, the 10-cluster "earlier" design has highest power while still maintaining a 17% reduction in sample size compared to the concurrent design.

This example illustrates several points. First, as expected for a factorial design, when intervention effects are additive, including a combined condition can increase efficiency and reduce the sample size requirement. However, the timing of transitions to the combined condition matters. Simply assigning all clusters to the combined condition for the last period reduced power compared to a concurrent design. Rather, the combined condition needs to be introduced earlier to realize

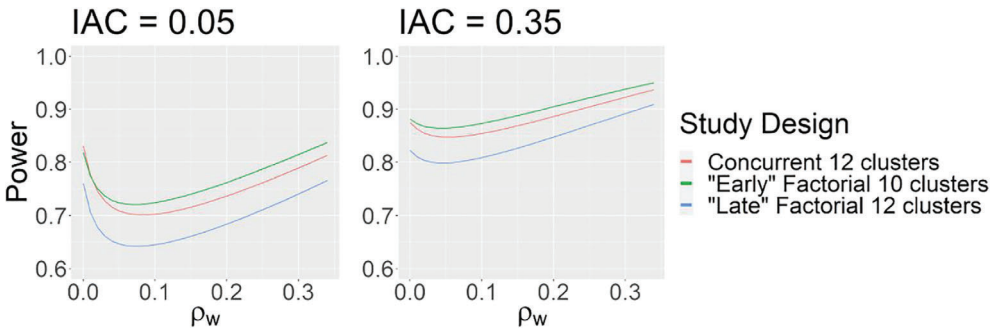

FIGURE 5 Comparison of power for main effects for three designs

efficiency gains. Further, if an interaction effect is present, the design in Figure 4A suffers from identifiability issues, as the interaction effect would be perfectly collinear with the last time period.

3.3 | Factorial designs with interaction effect

To study power for detecting an interaction, we consider hypothetical SWDs for evaluating two school-based interventions to reduce obesity among children. The primary outcome will be age- and sex-standardized BMI z-score. Variance parameter values were estimated using data from a previous study that measured BMI z-scores at three time points over 13 month among 286 children at nine schools (unpublished data). A linear mixed model based on Equation (1) was fit to the these data to obtain the estimates $\sigma_e^2 = 1.11$, $\sigma_v^2 = 0.14$, $\sigma_\psi^2 = 3.54$, and $\sigma_\alpha^2 = 0.24$ with total variance $\sigma_y^2 = 5.29$. On a standardized scale, these values correspond to $\rho_w = (0.24 + 0.14)/5.29 = 0.07$, $\rho_a = 0.24/5.29 = 0.05$, $\pi = 3.54/(3.54 + 1.11) = 0.76$ and CAC $= 0.24/(0.24 + 0.14) = 0.63$. The study is to be powered on detecting effect sizes of 1 on the z-score scale for each intervention and an interaction effect of 0.5 (also on the z-score scale), corresponding to standardized effect sizes of $1/\sqrt{5.29} = 0.44$ for main effects and 0.22 for the interaction effect. We note that the z-score outcomes are based on the 2000 Centers for Disease Control and Prevention (CDC) growth charts[26] and not our sample, which had substantially higher variance. The planned study will involve eight schools with 90 children at each school, and will have five 6-month periods. For simplicity, we assume no dropout.

We consider the designs displayed in Figure 6. Each of these designs has seven cluster-periods in Intervention 1 only, seven in Intervention 2 only, and ten in the combined condition. Design #1 is a concurrent design with the combined condition as another "stack". Design #2 is similar to a two-intervention concurrent design but has most clusters further transition to the combined condition. Designs #3 and #4 combine elements of Designs #1 and #2; they are distinguished by Design #4 having earlier introduction of the combined condition and featuring some clusters that never transition to the combined condition. Designs #1, #3, and #4 are symmetric in Interventions 1 and 2 and thus have equal power for these two effects. Design #2 is close to symmetric, but symmetry can be impossible to achieve with a small number of clusters. Type I error was set to $0.05/3 = 0.0167$ for three hypothesis tests.

The investigators considered the IAC of $\pi = 0.76$ in the prior study to be relatively high and thought that it might be lower in the planned study. To explore the impact of IAC on power, Figure 7A displays power for main effects for a range of plausible values of π. As shown by others,[27] power is an increasing function of π. Design #2 has the highest power for main effects for all values of π. In this design, power for Intervention 1 is slightly higher than that for Intervention 2 due to its more balanced sequencing over time (2 clusters receiving intervention in periods 2, 3, and 4 rather than 1, 2, then 3 clusters). For all values of π, Design #1 has lowest power. Focusing on power for the interaction, displayed in Figure 7B, Design #2 has by far the highest power; power for the three other designs is similar. Overall, power to detect the interaction is low.

Design #2 is clearly superior for detecting both main and interaction effects. In Design #2, clusters transition between conditions more than any other design. When there are more transitions, within-cluster comparisons are increased, and thus power to detect effects is increased. In Design #1, cluster transition only once, and thus this design has the lowest power for main effects. Beyond power, other drawbacks of the designs should be considered. For example, in Design #3, time in the combined condition occurs almost entirely during the last period, risking confounding with time. This example

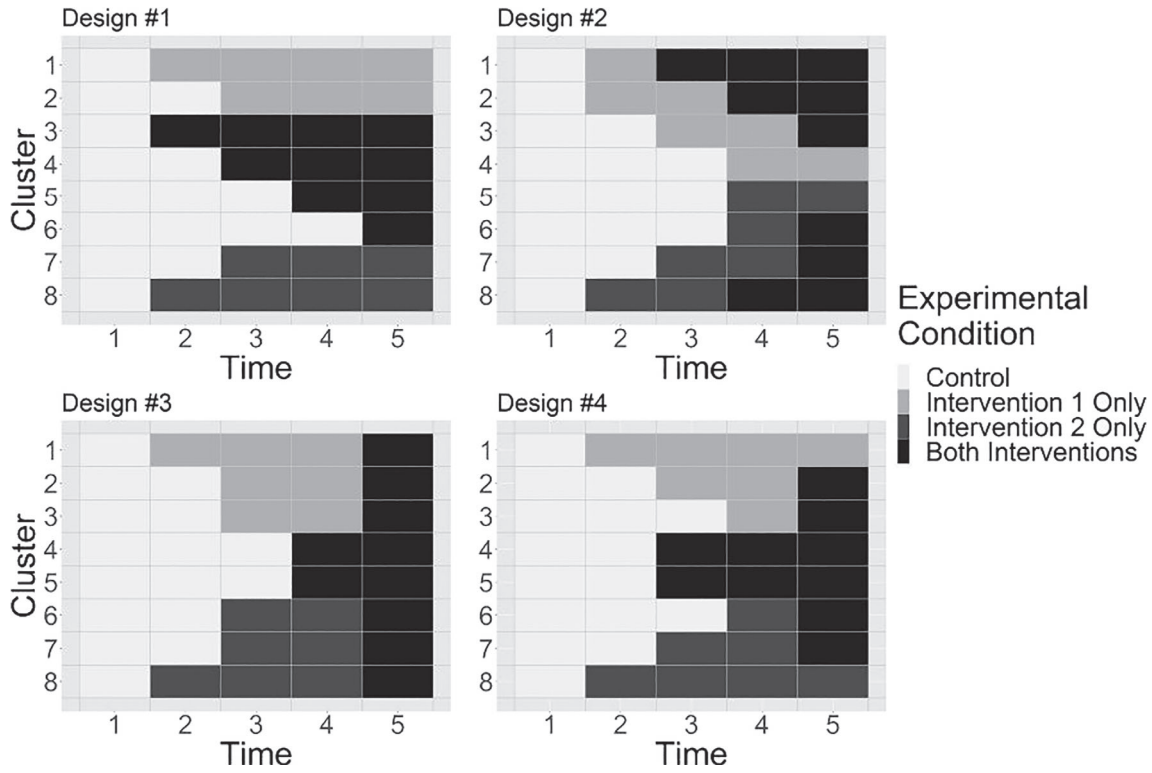

FIGURE 6 Variations of stepped wedge factorial design

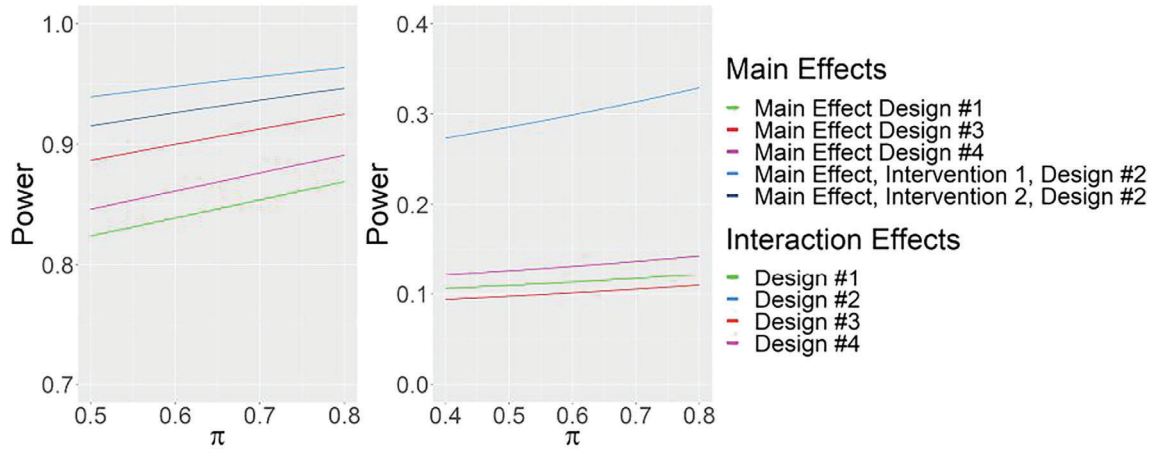

FIGURE 7 Comparison of power for detecting main and interaction effects

also illustrates that to power on the interaction term, designs should include two features: clusters that experience the control, single intervention and combined conditions, and relatively early introduction of the combined condition.

3.4 | Four-arm design

To study multiarm trials, we continue with SWDs for child obesity interventions using BMI z-score as the outcome variable and the variance parameter estimates from the previous similar study described in Section 3.3. We study the designs in Figure 6 but regard the combined condition as a third intervention (Intervention 3), and assume the goal is to compare each of the three interventions to the control condition. Other hypotheses could include direct comparisons of

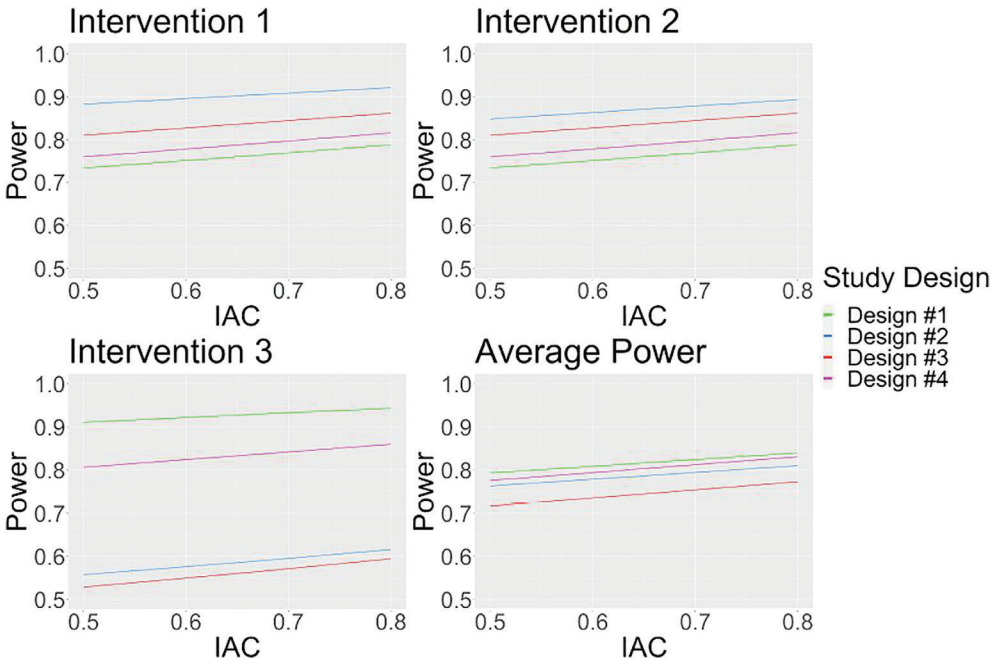

FIGURE 8 Comparison of power for each of three interventions for multiarm trials. Average power is calculated as the mean over all interventions

interventions. We assume a simple effect size of 0.92 for each of the intervention arm, corresponding to a standardized effect size of 0.4. Each cluster-period has $N = 90$ individuals. We use the same variance parameters as above, and allow the individual auto-correlation π to vary. Type I error is set to $0.05/3 = 0.0167$ for each of 3 tests. We compute power for each intervention as well as average power.

Power to detect all three interventions individually and averaged is shown in Figure 8 as a function of π. This example shows that unlike factorial designs, power in multiarm trials is less dependent on clusters transitioning to multiple intervention conditions and more dependent on when interventions are first introduced in the study. For Interventions 1 and 2, Design #2 yields the highest power across all values of π, as it introduces the intervention early in the study across multiple clusters. However, for Intervention 3, we see that Design #1 yields the highest power, as this design introduces Intervention 3 earlier in the study than any other design. We also note that for Intervention 3, Design #3 yields the lowest power, as only two cluster-periods are in this condition prior to the final time period, resulting in a significant amount of confounding between time and an intervention effect.

Design #2 has higher power than Design #3 for all interventions. However, Design #2 only outperforms Designs #1 and #4 for Interventions 1 and 2 and has lower power for Intervention 3. This can be attributed to the fact that Design #2 primarily features Intervention 3 in time periods 4 and 5. Looking at average power for all three interventions, Design #1 has the highest average power, but for higher π, has about average equal power with Design #4. Design #2 has about 0.03 lower average power compared with Design #1 for all π values shown, and Design #3 has about 0.07 lower average power than Design #1.

3.5 | Simulation study

We used simulation to verify the power calculations and Type I error rates for all examples in Section 3. We simulated 1000 data sets under the alternate hypothesis using representative values of the variance parameters that were allowed to vary to verify power for each example. Linear mixed models were fit to each simulated data set using restricted maximum likelihood as used in other stepped wedge simulation studies,[27] using the lme4 package in R.[28] No small sample size corrections were made. Power was calculated as the percentage of simulations in which the null hypothesis was rejected using a Wald test at the Bonferroni-corrected Type I error level. Type I error rates were estimated by simulating data under the null hypothesis, that is, setting all treatment effects to 0, and calculating the percentage of simulations in which

TABLE 1 Comparison of power based on simulation and proposed method (in parentheses)

	$\pi = 0.05$		$\pi = 0.35$	
	$\rho_w = 0.05$	$\rho_w = 0.30$	$\rho_w = 0.05$	$\rho_w = 0.30$
Design	$\delta_1 = 0.4$	$\delta_1 = 0.4$	$\delta_1 = 0.4$	$\delta_1 = 0.4$
Single intervention	0.61 (0.61)	0.71 (0.70)	0.78 (0.75)	0.86 (0.85)
12-cluster concurrent	0.73 (0.71)	0.80 (0.79)	0.83 (0.85)	0.93 (0.92)
10-cluster concurrent	0.60 (0.60)	0.68 (0.68)	0.77 (0.75)	0.88 (0.85)
Type I error (nominal error = 0.025)				
Single intervention	0.026	0.027	0.020	0.023
12-cluster concurrent	0.022	0.019	0.039	0.026
10-cluster concurrent	0.027	0.024	0.024	0.032

Note: Single-intervention and concurrent designs in Section 3.1.

TABLE 2 Comparison of power based on simulation and proposed method (in parentheses)

	$\pi = 0.05$		$\pi = 0.35$	
	$\rho_w = 0.05$	$\rho_w = 0.30$	$\rho_w = 0.05$	$\rho_w = 0.30$
Design	$\delta_1 = 0.4$	$\delta_1 = 0.4$	$\delta_1 = 0.4$	$\delta_1 = 0.4$
12-cluster concurrent	0.73 (0.71)	0.80 (0.79)	0.83 (0.85)	0.93 (0.92)
12-cluster late design	0.66 (0.65)	0.76 (0.75)	0.79 (0.80)	0.89 (0.89)
10-cluster early design	0.71 (0.72)	0.83 (0.79)	0.88 (0.86)	0.94 (0.94)
Type I error (nominal error = 0.025)				
12-cluster concurrent	0.019	0.018	0.036	0.022
12-cluster late design	0.018	0.024	0.028	0.016
10-cluster early design	0.026	0.026	0.023	0.020

Note: Concurrent designs and factorial design in Section 3.2.

the null hypothesis was falsely rejected using an experimentwise Type I error rate of 0.05 and Bonferroni corrections as described in the examples.

Tables 1,2,3, and 4 compare power calculated using our method to power estimated by simulation for each set of examples. For all examples, power calculated numerically using our method and simulated power were similar, with no indication of systematic under- or over-estimation of power. Similarly, Type I error rates from the simulations were reasonably close to the nominal level, and did not appear to be systematically over- or under-estimated.

4 | DISCUSSION

Stepped wedge designs with more than one intervention are being used in practice despite a paucity of literature on their statistical design and analysis. We have presented power calculation methods for stepped wedge design trials that have multiple interventions, both as multiarm and factorial designs. We focus on studies that include a relatively small number of clusters, which is common for stepped wedge trials.[29] In our examples, it was not feasible to explore all possible design options. However, the examples demonstrate several principles. We found that a concurrent design, in which two one-treatment stepped wedge trials are conducted as a single study, is more efficient than two separate one-treatment studies, which is supported by Lyons et al.[6] Our methods enable power calculations for such studies. In concurrent designs, cluster-periods in the control condition perform "double duty" by serving as controls for both treatment conditions. Such trials are essentially three-arm trials in which two interventions are each compared to a control condition.

TABLE 3 Comparison of power based on simulation and proposed method (in parentheses)

	$\rho_w = 0.07, \rho_a = 0.05, \pi = 0.5$			$\rho_w = 0.07, \rho_a = 0.05, \pi = 0.7$		
	Intvn 1	Intvn 2	Interaction	Intvn 1	Intvn 2	Interaction
Design	$\delta_1 = 0.44$	$\delta_1 = 0.44$	$\delta_3 = 0.22$	$\delta_1 = 0.44$	$\delta_2 = 0.44$	$\delta_3 = 0.22$
1	0.83 (0.82)	0.85 (0.82)	0.15 (0.11)	0.85 (0.85)	0.89 (0.85)	0.15 (0.11)
2	0.94 (0.94)	0.91 (0.92)	0.28 (0.29)	0.95 (0.96)	0.94 (0.94)	0.31 (0.31)
3	0.88 (0.89)	0.89 (0.89)	0.13 (0.10)	0.91 (0.91)	0.92 (0.91)	0.14 (0.11)
4	0.87 (0.85)	0.86 (0.85)	0.15 (0.13)	0.90 (0.88)	0.89 (0.88)	0.16 (0.14)
Type I error (nominal = 0.0167)						
1	0.016	0.019	0.014	0.014	0.018	0.013
2	0.011	0.013	0.015	0.009	0.014	0.014
3	0.016	0.021	0.017	0.015	0.021	0.011
4	0.011	0.017	0.011	0.012	0.016	0.008

Note: Factorial designs in Section 3.3.

TABLE 4 Comparison of power based on simulation and proposed method (in parentheses)

	$\rho_w = 0.07, \rho_a = 0.05, \pi = 0.5$			$\rho_w = 0.07, \rho_a = 0.05, \pi = 0.7$		
	Intvn 1	Intvn 2	Intvn 3	Intvn 1	Intvn 2	Intvn 3
Design	$\delta_1 = 0.5$	$\delta_1 = 0.4$	$\delta_3 = 0.4$	$\delta_1 = 0.4$	$\delta_2 = 0.4$	$\delta_3 = 0.4$
1	0.75 (0.74)	0.79 (0.74)	0.91 (0.91)	0.79 (0.77)	0.81 (0.77)	0.93 (0.93)
2	0.89 (0.88)	0.85 (0.85)	0.58 (0.56)	0.91 (0.91)	0.88 (0.88)	0.62 (0.60)
3	0.81 (0.81)	0.81 (0.81)	0.52 (0.53)	0.84 (0.85)	0.85 (0.85)	0.57 (0.57)
4	0.80 (0.76)	0.79 (0.76)	0.80 (0.81)	0.83 (0.80)	0.82 (0.80)	0.84 (0.84)
Type I error (nominal = 0.0167)						
1	0.016	0.019	0.016	0.014	0.018	0.016
2	0.011	0.013	0.010	0.009	0.014	0.009
3	0.016	0.021	0.024	0.015	0.021	0.025
4	0.011	0.017	0.010	0.012	0.016	0.011

Note: Multiarm designs in Section 3.4.

Our results also illustrate that stepped wedge factorial designs that include cluster-periods in a combined condition can increase power substantially compared to concurrent designs when treatment effects are additive. However, since the presence of an interaction generally decreases power for detecting main effects in factorial designs,[30] power may end up being inadequate if a potential interaction was not taken into account in power calculations. One approach for guarding against this eventuality is to conduct sensitivity analyses that assume some interaction between interventions when designing the study. Our power calculation methods can be used for this purpose.

In some studies, detecting an interaction effect may be of interest. Our work shows that in a stepped wedge factorial design where the aims include detecting an interaction effect, treatment sequencing is critical. We found that in general, designs in which clusters transition from control to single treatment to a combined treatment will be more powerful than designs in which clusters make only one transition, from control to a single treatment or control to combined condition. Such multiple-transition designs allow for more within-cluster comparisons, which are a driving factor in power for stepped wedge trials in general.[31]

A common method of handling interactions in factorial designs is to test for an interaction and drop it if it is not significant. This approach has been shown to lead to biased results.[32] We follow Kahan in recommending reporting results

both as a factorial design and as a multiarm analysis, where a condition with multiple treatments is considered as a separate treatment condition altogether.

Our examples included both repeated cross-sectional and cohort designs. Power for repeated cross-sectional versus cohort designs has been addressed by others;[15] in general, cohort designs have higher power than cross-sectional designs.[27,33] However, there is often a lack of information about parameter values to support power analysis for cohort designs. ICCs are typically reported as the within-time, within-cluster correlation, ρ_w; the across-time, within-cluster correlation ρ_a, and individual auto-correlation π are often not reported. Given this lack of information, it may be sensible to make the simplifying assumption that $\rho_w = \rho_a$, which corresponds to the repeated cross-sectional design.

When conducting multiple hypothesis tests in stepped wedge trials with multiple interventions, investigators should consider the need to control the experimentwise Type I error rate. We note that when multiple treatment groups are each compared to a common control group, Dunnett's method may be used for experimentwise Type I error rate control.[34] For other multiarm or factorial designs, there are several possible methods to control for familywise error rate.[35] In our examples, we used a Bonferroni correction. As the number of hypotheses increases, the familywise error rate may be better addressed using other techniques.

In this article, we focus on SWDs with 2 or 3 treatment conditions. However, our results are generalizable to designs with more interventions. Consider a model with M main effects and B two-way interaction terms, where $M \geq 2$ and $0 \leq B \leq \frac{M(M-1)}{2}$. The variance-covariance matrix of the regression coefficients would be a $(M+B) \times (M+B)$ matrix. The elements of this matrix would have the same form as the elements of the 3×3 matrix in Appendix S1 for diagonal and off-diagonal elements for both main and interaction effects. Solving for $[(M+B) \times (M+B)]^{-1}$ would yield the variance-covariance matrix for the estimated coefficients. Note that this approach holds for two-way interactions only; higher-order interactions are not considered.

There are several limitations to our work. We use standardized effect sizes. Standardized effect sizes may be misleading if underlying distributions are skewed.[36] A more extensive discussion of advantages and disadvantages of simple versus standardized effect sizes is found elsewhere.[37] We consider continuous outcomes only; further development is needed for noncontinuous outcomes, including binary, survival, categorical and count outcomes. In the model we present, the cluster autocorrelation is constrained to be the same for cluster means across time periods, regardless of the length of time between observing cluster level outcomes. This may not be an accurate assumption, as cluster means observed closer in time may be more correlated than those that are farther apart.[38] There are models for one treatment SWDs that allow the correlation between cluster means to decay over time.[21,39,40] For linear mixed models with a decaying correlation structure, the covariance matrix is a Toeplitz matrix and requires the use of the Trench algorithm to numerically invert.[39] We did not include this feature in our work here as we focused on the derivation of closed form variances and covariances of treatment and interaction effects. We only consider complete SWDs. Incomplete designs, in which data are not collected from some clusters in some periods, have been addressed for stepped wedge trials with a single treatment.[41,42] Another topic that could be explored further would be determining the minimum number of clusters, individuals per clusters, or design sequences that yield a certain level of power, which has been explored for stepped wedge designs with a single intervention,[43] but is out of the scope of our current work here. Finally, we have assumed that treatment effects are instantaneous and do not consider delays in treatment effects, which have been considered for SWDs with a single treatment.[1,40,44] Future work could explore how delays in one or both treatment effects may impact power of main and interaction effects.

CONFLICT OF INTEREST

No potential conflict of interest was reported by the authors.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in at https://github.com/phillipsundin/SWFD.

ORCID

Phillip Sundin https://orcid.org/0000-0003-2610-4644
Catherine M. Crespi https://orcid.org/0000-0002-6150-2181

REFERENCES

1. Hussey M, Hughes J. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials*. 2007;28:182-191.

2. Woertman W, de Hoop E, Moerbeek M, Zuidemac S, Gerritsen D, Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol*. 2013;66:752-758.
3. Grayling M, Wason J, Mander A. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. *Trials*. 2017;18(33).
4. Grayling M, Mander A, Wason J. Admissible multiarm stepped-wedge cluster randomized trial designs. *Stat Med*. 2018;38:1103-1119.
5. Zhang P, Shoben A, Jackson R, Fernandez S. Variance formulae for multiphase stepped wedge cluster randomized trial. *Stat Med*. 2020;39:4147-4168.
6. Lyons V, Li L, Hughes J, Rowhani-Rahbar A. Proposed variations of the stepped-wedge design can be used to accommodate multiple interventions. *J Clin Epidemiol*. 2017;86:160-167.
7. Comparing strategies for health clinics to increase HPV vaccinations in youth; 2017; Patient-Centered Outcomes Research Institute. https://www.pcori.org/research-results/2017/comparing-strategies-health-clinics-increase-hpv-vaccinations-youth.
8. van der Geest B, de Graaf J, Bertens L, et al. Screening and treatment to reduce severe hyperbilirubinaemia in infants in primary care (STARSHIP): a factorial stepped-wedge cluster randomised controlled trial protocol. *BMJ Open*. 2019;9.
9. Whittingham K, Sanders M, McKinlay L, Boyd R. Interventions to reduce behavioral problems in children with cerebral palsy: an RCT. *Pediatrics*. 2014;133:1249-1257.
10. Reuther S, Holle D, Buscher I, et al. Effect evaluation of two types of dementia-specific case conferences in German nursing homes (FallDem) using a stepped-wedge design: study protocol for a randomized controlled trial. *Trials*. 2014;15(319).
11. Durovni B, Saraceni V, Moulton L, et al. Effect of improved tuberculosis screening and isoniazid preventive therapy on incidence of tuberculosis and death in patients with HIV in clinics in Rio de Janeiro, Brazil: a stepped wedge, cluster-randomised trial. *Lancet Infect Dis*. 2013;13:852-858.
12. Durovni B, Saraceni V, van den Hof S, et al. Impact of replacing smear microscopy with Xpert MTB/RIF for diagnosing tuberculosis in Brazil: a stepped-wedge cluster-randomized trial. *PLoS Med*. 2014;12.
13. Oelhert G. *A First Course in Design and Analysis of Experiments*. New York, NY: W. H. Freeman; 2010:170-171.
14. Teerenstra S, Eldridge S, Graff M, de Hoop E, Borma G. A simple sample size formula for analysis of covariance in cluster randomized trials. *Stat Med*. 2012;31:2169-2178.
15. Feldman H, McKinlay S. Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. *Stat Med*. 1994;13:61-78.
16. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. Vol 2. New York, NY: Springer-Verlag; 2009:56-57.
17. Sherman J, Morrison W. Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix. *Ann Math Stat*. 1949;20(4):620-624.
18. Bartlett MS. An inverse matrix adjustment arising in discriminant analysis. *Ann Math Stat*. 1951;22(1):107-111.
19. Grayling M, Wason J. A web application for the design of multi-arm clinical trials. *BMC Cancer*. 2020;20(80).
20. Ford W, Westgate P. Maintaining the validity of inference in small-sample stepped wedge cluster randomized trials with binary outcomes when using generalized estimating equations. *Stat Med*. 2020;39:2779-2792.
21. Li F. Design and analysis considerations for cohort stepped wedge cluster randomized trials with a decay correlation structure. *Stat Med*. 2019;39:438-455.
22. R Core Team. *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing, 2019, https://www.R-project.org/.
23. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: L. Erlbaum Associates; 1988:24-27.
24. Hemming K, Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *J Clin Epidemiol*. 2016;69:137-146.
25. Baio G, Copas A, Ambler G, Hargreaves J, Beard E, Omar R. Sample size calculation for a stepped wedge trial. *Trials*. 2015;16(354).
26. Kuczmarski R, Ogden C, Guo S, et al. 2000 CDC growth charts for the United States: methods and development. *Vital Health Stat*. 2002;11:1-190.
27. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Stat Med*. 2016;35:4718-4728.
28. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1):1-48.
29. Taljaard M, Teerenstra S, Ivers N, Fergusson D. Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. *Clin Trials*. 2016;13:459-463.
30. Green S, Liu P, O'Sullivan J. Factorial design considerations. *J Clin Oncol*. 2002;20:3424-3230.
31. 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. *Int J Epidemiol*. 2020;49:979-995.
32. Kahan B. Bias in randomised factorial trials. *Stat Med*. 2013;32:4540-4549.
33. Feldman HA, McKinlay SM. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for designs. *Stat Med*. 1992;11:1685-1704.
34. Dunnett C. A multiple comparison procedure for comparing several treatments with a control. *Biometrics*. 1964;20:482-491.
35. Soulakova J. Resampling-based and other multiple testing strategies with application to combination drug trials with factorial designs. *Stat Methods Med Res*. 2011;20:505-521.
36. Botta-Dukat Z. Cautionary note on calculating standardized effect size (SES) in randomization test. *Community Ecol*. 2016;19:77-83.
37. Botta-Dukat Z. Standardized or simple effect size: what should be reported. *Br J Psychol*. 2009;100:603-617.

38. Hemming K, Taljaard M, Forbes A. Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. *Trials*. 2017;18(101).
39. Grantham K, Kasza J, Heritier S, Hemming K, Forbes A. Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. *Stat Med*. 2018;38:1918-1934.
40. Li F, Hughes J, Hemming K, Taljaard M, Melnick E, Heagerty P. Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: an overview. *Stat Methods Med Res*. 2020;30:1-28.
41. Hemming K, Lilford R, Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple level designs. *Stat Med*. 2014;34:181-196.
42. Kasza J, Taljaard M, Forbes A. Information content of stepped-wedge designs when treatment effect heterogeneity and/or implementation periods are present. *Stat Med*. 2019;39:4686-4701.
43. Barker D, D'Este C, Campbell MJ, McElduff P. Minimum number of clusters and comparison of analysis methods for cross sectional stepped wedge cluster randomised trials with binary outcomes: a simulation study. *Trials*. 2017;18.
44. Hughes J, Granston T, Heagerty P. Current issues in the design and analysis of stepped wedge trials. *Contemp Clin Trials*. 2015;45(Part A):55-60.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Sundin P, Crespi CM. Power analysis for stepped wedge trials with multiple interventions. *Statistics in Medicine*. 2022;41(8):1498-1512. doi: 10.1002/sim.9301