

Power Calculation for Noninferiority Stepped-Wedge Cluster Randomized Trials

Justin Guo

Montgomery Bell Academy

Chih-Yuan Hsu

Vanderbilt University Medical Center

Yu Shyr

yu.shyr@vumc.org

Vanderbilt University Medical Center

Research Article

Keywords: intraclass correlation, mixed-effects model, noninferiority test, superiority test, stepped-wedge cluster randomized trial

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10831163/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Power Calculation for Noninferiority Stepped-Wedge Cluster Randomized Trials

Justin Guo^{1,#}, Chih-Yuan Hsu^{2,3,#}, Yu Shyr^{2,3,*}

¹ Montgomery Bell Academy, Nashville, TN 37205, USA

² Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37203, USA

³ Center for Quantitative Sciences, Vanderbilt University Medical Center, Nashville, TN 37203, USA

[#]Justin Guo and Chih-Yuan Hsu contributed equally to this work

^{*}Correspondence: Yu Shyr: yu.shyr@vumc.org

Abstract

Background: Stepped-wedge cluster randomized trials (SW-CRTs) are increasingly used when interventions need to be implemented sequentially across clusters. Despite substantial advances in the design and analysis of superiority SW-CRTs, methods for planning noninferiority SW-CRTs remain limited.

Methods: We propose swNIsim, a simulation-based framework for power calculation in noninferiority SW-CRTs using generalized linear mixed-effects models (GLMMs). The proposed approach accommodates both continuous and binary outcomes and supports practical design features, including unequal numbers of clusters across sequences, varying cluster-period sizes, delayed intervention effects, and analyses with or without adjustment for secular time effects. Noninferiority margins can be specified using several commonly applied strategies.

Results: Through simulation studies, we examined the impact of key design parameters on statistical power, including the number of periods, the number of clusters per sequence, the number of individuals per cluster-period, and the degree of between-cluster variability. For a fixed total sample size, designs with more time periods consistently achieved higher power. Increasing the number of clusters per sequence produced modest gains in power, whereas greater between-cluster variability reduced power. Type I error rates were well controlled across the scenarios examined. Balanced allocation of clusters across sequences was more efficient than unbalanced allocation schemes. Two real-world SW-CRT examples were used to illustrate application of the method.

Conclusions: The proposed swNIsim provides a flexible and practical approach for power calculation in noninferiority SW-CRTs. By accommodating a wide range of design configurations and outcome types, and through implementation as a freely available R package, swNIsim may facilitate the planning and design of noninferiority stepped-wedge trials.

Keywords: intraclass correlation; mixed-effects model; noninferiority test; superiority test; stepped-wedge cluster randomized trial

1 Introduction

A Stepped-Wedge Cluster Randomized Trial (SW-CRT) is an alternative to the conventional parallel cluster randomized design in which clusters (e.g., hospitals, clinics, or communities) initially receive the control condition and subsequently cross over to the intervention at predefined, staggered time points until all clusters have been exposed. This design is particularly advantageous when simultaneous implementation across all sites is impractical or when withholding the intervention from certain clusters for the entire study would be ethically or operationally challenging [1]. Examples include quality-improvement initiatives, clinic-based screening programs, and workflow modifications that require phased implementation because of limitations in training, staffing, equipment, or administrative capacity [1]. By accommodating gradual rollout, the SW-CRT enables rigorous evaluation of intervention effects under real-world implementation constraints.

A key statistical characteristic of the stepped-wedge design is that each cluster contributes data under both control and intervention conditions. This structure supports both between-cluster and within-cluster comparisons. Because each cluster serves, in part, as its own control, within-cluster comparisons can improve estimation precision, although the magnitude of this gain depends on factors such as the timing of intervention rollout, cluster size, and the underlying correlation structure [2].

Hussey and Hughes [3] established a linear mixed-effects model (LMM) framework for the design and analysis of SW-CRTs. Subsequent work by Hemming et al. [1,4] investigated design variations and their impact on statistical power, while later studies extended these methods to generalized linear mixed-effects models (GLMMs) for non-Gaussian outcomes [5]. Generalized

estimating equations have also been proposed as a population-averaged alternative that avoids specifying a distribution for random effects [6].

Despite these methodological advances, existing approaches have been developed for superiority trials, whose objective is to determine whether an intervention performs better than the control. In contrast, methods for noninferiority SW-CRTs remain limited. Noninferiority designs are often appropriate when a new intervention may be slightly less effective than the standard treatment but offers other important advantages, such as reduced cost, easier implementation, improved accessibility, or fewer adverse effects.

To address this methodological gap, we propose swNIsim, a simulation-based framework for power calculation in noninferiority SW-CRTs using GLMMs. Unlike analytical approaches that often rely on restrictive assumptions, simulation-based power calculation provides flexibility for accommodating complex design features commonly encountered in practice. The proposed approach supports both continuous and binary outcomes, allows unequal numbers of clusters across sequences and varying cluster-period sizes, accommodates delayed intervention effects, and permits analyses with or without adjustment for time-period effects. In addition, swNIsim provides flexible options for specifying noninferiority margins. Using extensive simulations, we investigate the effects of major design characteristics on power and Type I error control. We further demonstrate the application of the proposed method through two real-world trial examples and provide an accompanying R package to facilitate its implementation.

2 Methods

Suppose the study comprises T time periods, corresponding to $T - 1$ intervention sequences. Each sequence s contains n_s clusters, $s = 1, \dots, T - 1$, yielding a total of $I = \sum_{s=1}^{T-1} n_s$ clusters. Every cluster is randomly assigned to one of the $T - 1$ sequences. For clusters assigned to sequence s , individuals remain in the control condition through period s and transition to the intervention beginning in period $s + 1$ (Figure 1). Let K_{ij} denote the number of individuals in cluster i during period j , $i = 1, \dots, I$ and $j = 1, \dots, T$. In designs with loss to follow-up or incomplete observations resulting from staggered participant enrollment, K_{ij} may decrease over time.

2.1 Statistical models for different outcome types

2.1.1 Continuous outcomes

Let $\bar{y}_{ij}$ denote the cluster–period mean, defined as the average outcome among individuals in cluster i at time period j , for $i = 1, \dots, I$ and $j = 1, \dots, T$. We assume

$$\bar{y}_{ij}|\mu_{ij} \sim N(\mu_{ij}, \sigma_w^2/K_{ij}) \quad (1)$$

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \theta\delta_{ij}$$

where the cluster-specific random effect α_i follows a normal distribution $N(0, \sigma_c^2)$. The parameters σ_w^2 and σ_c^2 denote the within-individual and between-cluster variances, respectively. The intraclass correlation (ICC), quantifying the correlation between individuals from the same cluster, is given by $\sigma_c^2/(\sigma_c^2 + \sigma_w^2)$. In this model, μ denotes the baseline mean under the standard treatment, and β_j is the fixed effect of time period j . The parameter θ denotes the treatment effect, defined as the difference between the intervention and the standard treatment. The term δ_{ij} reflects the level of intervention exposure for cluster i at time period j , allowing for the possibility of delayed intervention effects. In the absence of delayed intervention effects, $\delta_{ij} = 1$ for $j \geq t_i$ and $\delta_{ij} = 0$ otherwise, where t_i denotes the time period at which cluster i initiates the intervention. If delayed intervention effects are present, δ_{ij} may be defined, for example, as $\delta_{ij} = (j - t_i + 1)/(T - 1)$ for $j \geq t_i$ and $\delta_{ij} = 0$ for $j < t_i$.

The noninferiority hypotheses are formulated as

$$H_0: \theta \leq -\Delta \text{ vs. } H_1: \theta > -\Delta,$$

where $\Delta > 0$ denotes the noninferiority margin. A Wald-type test statistic $z = (\hat{\theta} + \Delta) / SE(\hat{\theta})$ is used to test the one-sided hypotheses, where $\hat{\theta}$ is the estimated treatment effect of θ , obtained using a linear mixed-effect model that includes a cluster-level random intercept and, by default, adjusts for fixed effects of time periods, with $\bar{y}_{ij}$ weighted by K_{ij} . The corresponding p-value is $1 - \Phi(z)$, where Φ denotes the cumulative distribution function of the standard normal distribution. Noninferiority is concluded when the p-value is less than the prespecified significance level.

Statistical power is estimated using Monte Carlo simulation with a large number of replicates. For each replicate, a dataset is generated under the prespecified design, the planned mixed-effects model is fitted with the R package *lme4* [7], and the noninferiority test is performed. The estimated power is the proportion of replicates in which the null hypothesis is rejected. The swNIsim also reports the model-fitting failure rate. By default, replicates with failed model fits are treated as non-rejections and are therefore included in the denominator when calculating power.

2.1.2 Binary outcomes

Let s_{ij} denote the cluster–period total of binary outcomes, defined as the number of individuals with a response in cluster i at time period j , for $i = 1, \dots, I$ and $j = 1, \dots, T$. We assume

$$s_{ij}|p_{ij} \sim \text{Binomial}(K_{ij}, p_{ij}) \quad (2)$$

$$\text{logit}(p_{ij}) = b_1 + \alpha_i + \beta_j + \theta\delta_{ij}$$

where the cluster-specific random effect α_i follows a normal distribution $N(0, \sigma_c^2)$. The intercept b_1 is the baseline log odds under the standard treatment, and $(1 + \exp(-b_1))^{-1}$ is the corresponding baseline response probability. Time period effects, treatment exposure, and the remaining parameters are defined as in the continuous-outcome model. For this logistic random-intercept model, the latent-scale ICC is given by $\sigma_c^2/(\sigma_c^2 + \pi^2/3)$ [8]. For example, $\sigma_c = 1$ corresponds to a latent-scale ICC of approximately 0.233.

The same one-sided Wald framework is used for the binary outcome. Specifically, $\hat{\theta}$ is estimated using a logistic mixed-effects model with a cluster-specific random intercept and, by default, fixed effects for time periods. The noninferiority margin Δ is defined on the log-odds scale. Noninferiority is concluded when $1 - \Phi[(\hat{\theta} + \Delta) / SE(\hat{\theta})]$ is less than the prespecified significance level. As in the continuous-outcome setting, statistical power is estimated using Monte Carlo simulation with the same rejection and failure-handling rules.

2.2 Noninferiority margin determination

A noninferiority margin, Δ on the fitted-model scale is specified through the “margin” argument. Following [9], we provide two options for determining the nonnegative margin Δ .

2.2.1 Continuous outcomes

Preserved fraction of the mean under the standard treatment

The aim is for the intervention group to preserve at least a fraction f_1 of the mean response under the standard treatment, where $0 < f_1 < 1$. The preserved fraction f_1 represents the proportion of the standard treatment mean that is considered acceptable for the intervention group. This can be expressed as $(\mu + \theta) > f_1\mu$, which implies $\theta > -(1 - f_1)\mu$. Thus, the margin is $\Delta = (1 - f_1)\mu$.

Preserved fraction of the mean difference between the standard treatment and placebo

The aim is for the intervention to be superior to the placebo and to preserve at least a fraction f_2 of the mean difference between the standard treatment and placebo, where $0 < f_2 < 1$. The preserved fraction f_2 represents the proportion of this difference that is considered acceptable for the mean difference between the intervention and placebo. This can be expressed as $(\mu + \theta) - \mu_0 > 0$ and $(\mu + \theta) - \mu_0 > f_2(\mu - \mu_0)$, where μ_0 is the placebo mean. Typically, $\mu - \mu_0 > 0$ holds, so the margin is $\Delta = (1 - f_2)(\mu - \mu_0)$.

2.2.2 Binary outcomes

For binary outcomes, the treatment effect is quantified by the log odds ratio. Therefore, the noninferiority margin is specified on the log-odds scale. If a target margin is initially defined on the proportion scale, it will be converted to the corresponding log-odds scale. Throughout this article, a larger proportion indicates greater clinical benefit.

Preserved fraction of the success rate under the standard treatment

The aim is for the intervention group to preserve at least a fraction f_1 of the success rate, p_1 , of the standard treatment, where $0 < f_1 < 1$. The preserved fraction f_1 represents the proportion of the standard treatment success rate that is considered acceptable for the intervention group. The lowest

acceptable success probability for the intervention is $p_2^L = f_1 p_1$. Therefore, the resulting margin on the log-odds scale is $\Delta = \text{logit}(p_1) - \text{logit}(p_2^L)$.

Preserved fraction of the success rate difference between the standard treatment and placebo

The aim is for the intervention to be superior to the placebo and to preserve at least a fraction f_2 of the success rate difference between the standard treatment and placebo, where $0 < f_2 < 1$. The preserved fraction f_2 represents the proportion of $p_1 - p_0$ (standard vs. placebo) that should be retained in $p_2 - p_0$ (intervention vs. placebo). The lowest acceptable success probability for the intervention is $p_2^L = p_0 + f_2(p_1 - p_0)$. Therefore, the resulting margin on the log-odds scale is $\Delta = \text{logit}(p_1) - \text{logit}(p_2^L)$.

3 Results

3.1 Simulation studies

We conducted a series of simulation studies to evaluate the impact of key design characteristics on statistical power, including 1) the number of time periods, 2) the number of clusters per sequence, 3) the number of individuals per cluster-period, and 4) the magnitude of between-cluster variability, while keeping the total number of individuals fixed. Specifically, we considered a range of stepped-wedge design configurations defined by different combinations of the number of time periods (T), the number of clusters per sequence ($n_s = n$, for all s), and the number of individuals per cluster-period ($K_{ij} = K$, for all i, j): $(T, n, K) = (7, 10, 20), (7, 4, 50), (6, 12, 20), (6, 8, 30), (5, 15, 20), (5, 10, 30), (4, 20, 20), (4, 10, 40)$. Each configuration included 1,200 individuals. The significance level was set at 0.025.

The true treatment effect and time period effects were set to zero. i.e., $\theta = 0$ and $\beta_j = 0$ for all j . For Model (1) with continuous outcomes, we set $\mu = 5$, $\sigma_w = 2$, and $\sigma_c = 1.0, 1.3$, or 1.6 , corresponding to ICC values of approximately 0.20, 0.30, and 0.39, respectively. The noninferiority margin was set to 0.15. For Model (2) with binary outcomes, we set $b_1 = 0.5$, corresponding to a conditional response probability of 0.622 under the reference condition, and $\sigma_c = 1.0, 1.5$, and 2.0 , corresponding to latent-scale ICC values of approximately 0.233, 0.406, and

0.549, respectively. The noninferiority margin was set to 0.18. Empirical power in each scenario was calculated using 100,000 replicates.

For both continuous and binary outcomes, simulations showed that, given the fixed total sample size, designs with a larger number of time periods generally produced higher power (Tables 1 and 2). Increasing the number of clusters per sequence while reducing the number of individuals per cluster-period slightly improved power, whereas greater between-cluster variability reduced power. The number of time periods appeared to be the most influential design factor.

We examined Type I error rates under the boundary null hypothesis ($\theta = -\Delta$). At the nominal one-sided level of 0.025, empirical Type I error rates ranged from 0.0254 to 0.0269 for continuous outcomes and from 0.0250 to 0.0256 for binary outcomes, indicating that the Type I error rate was well controlled (Supplemental Tables 1 and 2). The greatest inflation was observed in the smallest four-period, 30-cluster configuration for continuous outcomes.

We also conducted simulations allowing the number of clusters to vary across sequences. Specifically, for a design with five sequences (i.e., $T = 6$), we considered three cluster allocations: $(n_1, \dots, n_5) = (8, 8, 8, 8, 8)$, $(4, 6, 8, 10, 12)$, and $(12, 10, 8, 6, 4)$. Other parameter settings were identical to those described above, with $K = 20$, and $\sigma_c = 1.3$ for continuous outcomes and 1.5 for binary outcomes (Table 3). The highest power was observed under the balanced design, $(n_1, \dots, n_5) = (8, 8, 8, 8, 8)$, indicating that equal allocation of clusters across sequences is the most efficient configuration.

Because no secular time effects were imposed ($\beta_j = 0$ for all j), we also evaluated power when period effects were not included in Models (1) and (2). As shown in Tables 4 and 5, the overall power patterns were similar to those obtained from models that adjusted for period effects. However, power was consistently higher when period effects were excluded, reflecting the efficiency loss incurred by estimating time-effect parameters that were truly zero.

3.2 Real Applications

Two real-world stepped-wedge cluster-randomized trials were used to illustrate the proposed swNIsim. These examples serve as parameterized demonstrations of the methodology rather than

exact reproductions of the original studies or re-evaluations of their superiority-based sample size calculations. The first example is based on the TheraPPP trial for continuous-outcome setting [10]; the second example is based on the PRIM-ER trial for binary-outcome setting [11,12].

3.2.1 Continuous outcome example: TheraPPP trial

The TheraPPP trial evaluated the effectiveness of a standardized care pathway for mechanically ventilated (MV) patients, including subgroups with hypoxemic respiratory failure (HRF) and acute respiratory distress syndrome (ARDS). The primary continuous outcome was 28-day ventilator-free days (VFDs). This trial was conducted across 17 adult intensive care units (ICU) in Alberta, Canada. In the full stepped-wedge design, 16 ICUs were organized into eight groups, each containing two ICUs, with one cluster initiating the intervention every two months.

For the overall MV patients, the study protocol reported a baseline mean of 21 VFDs, a standard deviation of 10 days, and an ICC of 0.15. The original trial was designed for 90% power at a two-sided significance level of 0.05 to detect a 0.9-day superiority difference [10]. For illustration of the proposed noninferiority framework, we adopted these design parameters and specified $T - 1 = 8$ (sequences), $n_s = 2$ (number of clusters per sequence), $\mu = 21$, $\sigma_w = 10$, $\sigma_c = 4.2$, $\beta_j = 0$, and $\theta = 0$. The noninferiority margin was defined as $\Delta = (1 - f_1)\mu = 4.2$, where $f_1 = 0.8$. Under these assumptions, the required cluster-period size $K_{ij} = K = 5$ was determined to achieve 90% power at a one-sided significance level of 0.025. Similarly, for the subgroup of patients with sustained ARDS, the protocol reported a baseline mean of 11 VFDs, a standard deviation of 10 days, and an ICC of 0.01. Using these parameters for methodological illustration, we specified $T - 1 = 8$, $n_s = 2$, $\mu = 11$, $\sigma_w = 10$, $\sigma_c = 1$, $\beta_j = 0$, and $\theta = 0$. The noninferiority margin was defined as $\Delta = (1 - f_1)\mu = 2.2$, where $f_1 = 0.8$. The required cluster-period size $K_{ij} = K = 14$ was determined to achieve 90% power at a one-sided significance level of 0.025.

3.2.2 Binary outcome example: PRIM-ER trial

The PRIM-ER trial evaluated the effect of a multicomponent intervention to initiate palliative care in the emergency department (ED) on hospital admission. The primary outcome was hospital admission following an ED visit. Based on the assumptions used in the original sample size

calculation (a baseline hospital admission rate of 70%, an ICC of 0.10, and 30 participating ED sites), the trial was estimated to have approximately 85% power to detect an absolute difference of 3 percentage points in hospital admission rates.

To illustrate the proposed noninferiority framework, we specified $T - 1 = 30$, $n_s = 1$, $b_1 = -0.847$ (corresponding to a baseline non-admission probability of $p_1 = 0.3$), $\sigma_c = 0.605$ (determined to yield a latent-scale ICC of 0.1), $\beta_j = 0$, and $\theta = 0$. The noninferiority margin was defined as $\Delta = 0.305$, derived using $f_1 = 0.8$ and $p_1 = 0.3$. Under these assumptions, the required cluster-period size $K_{ij} = K = 6$ was determined to achieve 85% power at a one-sided significance level of 0.025.

4 Discussion

In this study, we proposed swNIsim, a simulation-based framework for power calculation in noninferiority stepped-wedge cluster randomized trials. While substantial methodological research has focused on the design of superiority SW-CRTs, methods for planning noninferiority SW-CRTs remain limited. The proposed framework addresses this gap by providing a flexible approach for estimating power under generalized linear mixed-effects models for both continuous and binary outcomes. The method accommodates several practical design features frequently encountered in applied studies, including unequal numbers of clusters across sequences, varying cluster-period sizes, delayed intervention effects, and analytic models with or without adjustment for period effects.

The simulation studies provide several insights into the design of noninferiority SW-CRTs. For a fixed total sample size, increasing the number of time periods consistently improved power for both continuous and binary outcomes. This finding suggests that the number of periods is a particularly influential design parameter in stepped-wedge trials. Increasing the number of clusters per sequence while reducing the number of individuals per cluster-period resulted in modest improvements in power, whereas increasing between-cluster variability reduced power. We also found that balanced allocation of clusters across sequences provided higher power than unequal allocation schemes, supporting the use of balanced designs whenever feasible.

The simulations further demonstrated adequate control of the Type I error rate across the scenarios examined. Although only slight inflation was observed, the largest deviations occurred in designs

with relatively few clusters. This observation is consistent with previous studies reporting that asymptotic inference in mixed-effects models may exhibit inflated Type I error rates when the number of clusters is small [13,14]. Consequently, investigators should exercise caution when designing SW-CRTs with a limited number of clusters and may wish to evaluate operating characteristics through simulation before finalizing the design.

Another notable finding concerns adjustment for secular time effects. When no true period effects were present in the data-generating mechanism, analyses that omitted period effects achieved higher power than models that included them. This result reflects the loss of efficiency associated with estimating unnecessary parameters. In practice, however, secular trends often exist and should generally be incorporated into the primary analysis. Therefore, the observed gains in power should not be interpreted as a recommendation to routinely omit period effects, but rather as an illustration of the trade-off between model complexity and efficiency.

Several limitations should be noted. The proposed approach models correlations among observations from the same cluster across periods through a shared cluster-level random intercept. Although this assumption is widely used in the stepped-wedge literature, it may not adequately capture more complex correlation structures, such as decay or autoregressive patterns over time [15,16], that arise in some applications. Future extensions could incorporate alternative correlation structures and additional random effects [17,18]. In addition, simulation-based methods are computationally more intensive than analytical approaches, particularly when large numbers of replicates are required to obtain precise power estimates. Nevertheless, simulation-based approaches offer greater flexibility and can accommodate complex design features and modeling assumptions that are difficult to address analytically.

5 Conclusions

We propose swNIsim, a simulation-based framework for power calculation in noninferiority stepped-wedge cluster randomized trials. The proposed method enables investigators to evaluate realistic design scenarios and estimate power under a wide range of assumptions that may be difficult to address analytically. Implemented as a freely available R package, swNIsim provides a practical and flexible tool for the planning and design of noninferiority SW-CRTs.

Abbreviations

ARDS: acute respiratory distress syndrome

ED: emergency department

GLMMs: generalized linear mixed-effects models

HRF: hypoxemic respiratory failure

ICC: intraclass correlation

ICU: intensive care units

LMM: linear mixed-effects model

MV: mechanically ventilated

SW-CRT: stepped-wedge cluster randomized trial

VFDs: ventilator-free days

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Clinical trial number

Not applicable.

Availability of data and materials

Additional file 1: Online Supplementary Material. A freely available R package *swNIsim* is at <https://github.com/justing-uo/swNIsim>.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Institutes of Health [P50CA236733, U54CA163072], the VA Tennessee Valley Healthcare System [VUMC120842], and Lifespan Rhode Island Hospital [U01CA287008].

Authors' contributions

JG and CYH were major contributors in writing the manuscript and developing the R package. YS was the major contributor in the conception and design of the work.

References

1. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*. 2015;350:h391.
2. Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol*. 2013;66(7):752-758.
3. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials*. 2007;28(2):182-191.
4. Hemming K, Lilford R, Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Stat Med*. 2015;34(2):181-196.
5. Li F, Hughes JP, Hemming K, Taljaard M, Melnick ER, Heagerty PJ. Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: an overview. *Stat Methods Med Res*. 2021;30(2):612-639.
6. Scott JM, deCamp A, Juraska M, Fay MP, Gilbert PB. Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. *Stat Methods Med Res*. 2017;26(2):583-597.
7. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1):1-48.
8. Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. *International Statistical Review*, 2009;77: 378-394.

9. Shih A, Hsu CY, Shyr Y. Power and sample size calculation for non-inferiority trials with treatment switching in intention-to-treat analysis comparing RMSTs. *BMC Med Res Methodol.* 2025;25(1):157.
10. Parhar KKS, Soo A, Knight G, et al. Protocol and statistical analysis plan for the identification and treatment of hypoxemic respiratory failure and acute respiratory distress syndrome with protection, paralysis, and proning: a type-1 hybrid stepped-wedge cluster randomised effectiveness-implementation study. *Crit Care Resusc.* 2023;25(4):207-215.
11. Grudzen CR, Siman N, Cuthel AM, et al. Palliative care initiated in the emergency department: a cluster randomized clinical trial. *JAMA.* 2025;333(7):599-608.
12. Grudzen CR, Brody AA, Chung FR, et al. Primary palliative care for emergency medicine (PRIM-ER): protocol for a pragmatic, cluster-randomised, stepped wedge design to test the effectiveness of primary palliative care education, training and technical support for emergency medicine. *BMJ Open.* 2019;9(7):e030099.
13. Nugent JR, Kleinman KP. Type I error control for cluster randomized trials under varying small sample structures. *BMC Medical Research Methodology.* 2021;21:65.
14. 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;18:119.
15. Kenny A, Voldal EC, Xia F, Heagerty PJ, Hughes JP. Analysis of stepped wedge cluster randomized trials in the presence of a time-varying treatment effect. *Stat Med.* 2022;41(22):4311-4339.
16. Li F, Chen X, Tian Z, Wang R, Heagerty PJ. Planning stepped wedge cluster randomized trials to detect treatment effect heterogeneity. *Stat Med.* 2024;43(5):890-911.
17. 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(26):4718-4728.
18. 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. *Stat Methods Med Res.* 2019;28(3):703-716.

	Time period
--	-------------

	1	2	3	4	...	$T - 1$	T
Sequence 1 n_1 clusters							
Sequence 2 n_2 clusters							
Sequence 3 n_3 clusters							
$\vdots$					$\ddots$		
$\vdots$							
Sequence $T - 1$ n_{T-1} clusters							

Figure 1. A stepped-wedge design comprising T time periods and $T - 1$ sequences, with n_s clusters for each sequence. Clusters in sequence s initiate the intervention at time period $s + 1$.

Table 1. Power for continuous outcomes under models with adjusting for period effects. $\mu = 5$, $\sigma_c = (1.0, 1.3, 1.6)$, $\sigma_w = 2$, $\theta = 0$, and the noninferiority margin = 0.15. A one-sided $\alpha = 0.025$ was used.

T	n	K	Power		
			$\sigma_c = 1.0$	$\sigma_c = 1.3$	$\sigma_c = 1.6$
7	10	20	0.4999	0.4966	0.4943
	4	50	0.4958	0.4932	0.4923
6	12	20	0.4345	0.4339	0.4313
	8	30	0.4326	0.4320	0.4294
5	15	20	0.3661	0.3627	0.3613
	10	30	0.3640	0.3613	0.3601
4	20	20	0.2868	0.2834	0.2819
	10	40	0.2857	0.2833	0.2764

Table 2. Power for binary outcomes under models with adjusting for period effects. Baseline log odds $b_1 = 0.5$, $\sigma_c = (1.0, 1.5, 2.0)$, $\theta = 0$, and the noninferiority margin = 0.18. A one-sided $\alpha = 0.025$ was used.

T	n	K	Power		
			$\sigma_c = 1.0$	$\sigma_c = 1.5$	$\sigma_c = 2.0$
7	10	20	0.5570	0.4935	0.4357
	4	50	0.5493	0.4867	0.4308
6	12	20	0.4890	0.4295	0.3783
	8	30	0.4853	0.4259	0.3768
5	15	20	0.4118	0.3602	0.3161
	10	30	0.4079	0.3575	0.3137
4	20	20	0.3213	0.2811	0.2466
	10	40	0.3158	0.2775	0.2453

Table 3. Power under scenarios with unequal numbers of clusters across sequences. All models adjusted for period effects. $T = 6$, $K = 20$, $\theta = 0$, and $(n_1, \dots, n_4) = (8, 8, 8, 8, 8)$, $(4, 6, 8, 10, 12)$, or $(12, 10, 8, 6, 4)$, yielding 40 total clusters and 4,800 total observations. For continuous outcomes, $\mu = 5$, $\sigma_w = 2$, $\sigma_c = 1.3$, and the noninferiority margin was 0.15; while for binary outcomes, $b_1 = 0.5$, $\sigma_c = 1.5$, and the noninferiority margin was 0.18. A one-sided $\alpha = 0.025$ was used.

Outcome	Sequence allocation	Clusters by sequence	σ_c	Power
Continuous	Balanced	8, 8, 8, 8, 8	1.3	0.3164
Continuous	Increasing	4, 6, 8, 10, 12	1.3	0.2944
Continuous	Decreasing	12, 10, 8, 6, 4	1.3	0.2960
Binary	Balanced	8, 8, 8, 8, 8	1.5	0.3152
Binary	Increasing	4, 6, 8, 10, 12	1.5	0.2872
Binary	Decreasing	12, 10, 8, 6, 4	1.5	0.2854

Table 4. Power for continuous outcomes under models without adjusting for period effects. $\mu = 5$, $\sigma_c = (1.0, 1.3, 1.6)$, $\sigma_w = 2$, $\theta = 0$, and the noninferiority margin = 0.15. A one-sided $\alpha = 0.025$ was used.

T	n	K	Power		
			$\sigma_c = 1.0$	$\sigma_c = 1.3$	$\sigma_c = 1.6$
7	10	20	0.8529	0.8522	0.8519
	4	50	0.8508	0.8513	0.8504
6	12	20	0.8048	0.8032	0.8027
	8	30	0.8034	0.8020	0.8015
5	15	20	0.7417	0.7408	0.7391
	10	30	0.7402	0.7406	0.7387
4	20	20	0.6643	0.6631	0.6617
	10	40	0.6627	0.6604	0.6607

Table 5. Power for binary outcomes under models without adjusting for period effects. Baseline log odds $b_1 = 0.5$, $\sigma_c = (1.0, 1.5, 2.0)$, $\theta = 0$, and the noninferiority margin = 0.18. A one-sided $\alpha = 0.025$ was used.

T	n	K	Power		
			$\sigma_c = 1.0$	$\sigma_c = 1.5$	$\sigma_c = 2.0$
7	10	20	0.8996	0.8515	0.7964
	4	50	0.8955	0.8476	0.7910
6	12	20	0.8596	0.8034	0.7421
	8	30	0.8570	0.8007	0.7391
5	15	20	0.8021	0.7424	0.6760
	10	30	0.8011	0.7380	0.6743
4	20	20	0.7292	0.6636	0.5991
	10	40	0.7231	0.6594	0.5938

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [swNlsimSupplementary2026.08.26.docx](#)