

Information growth for sequential monitoring of clinical trials with a stepped wedge cluster randomized design and unknown intracluster correlation

Clinical Trials
2020, Vol. 17(2) 176–183
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1740774520901488
journals.sagepub.com/home/ctj



Siobhan P Brown¹  and Abigail B Shoben²

Abstract

Background/aims: In a stepped wedge study design, study clusters usually start with the baseline treatment and then cross over to the intervention at randomly determined times. Such designs are useful when the intervention must be delivered at the cluster level and are becoming increasingly common in practice. In these trials, if the outcome is death or serious morbidity, one may have an ethical imperative to monitor the trial and stop before maximum enrollment if the new therapy is proven to be beneficial. In addition, because formal monitoring allows for the stoppage of trials when a significant benefit for new therapy has been ruled out, their use can make a research program more efficient. However, use of the stepped wedge cluster randomized study design complicates the implementation of standard group sequential monitoring methods. Both the correlation of observations introduced by the clustered randomization and the timing of crossover from one treatment to the other impact the rate of information growth, an important component of an interim analysis.

Methods: We simulated cross-sectional stepped wedge study data in order to evaluate the impact of sequential monitoring on the Type I error and power when the true intracluster correlation is unknown. We studied the impact of varying intracluster correlations, treatment effects, methods of estimating the information growth, and boundary shapes.

Results: While misspecified information growth can impact both the Type I error and power of a study in some settings, we observed little inflation of the Type I error and only moderate reductions in power across a range of misspecified information growth patterns in our simulations.

Conclusion: Taking the study design into account and using either an estimate of the intracluster correlation from the ongoing study or other data in the same clusters should allow for easy implementation of group sequential methods in future stepped wedge designs.

Keywords

Group randomized clinical trial, group sequential design, information growth, sequential monitoring, stepped wedge

Background/aims

Stepped wedge design

In a cluster randomized trial, clusters, rather than individuals, are the unit of randomization. These designs are commonly used when the intervention of interest must be delivered at the group level, such as a school-based smoking prevention program or a study of infection control measures at a hospital. Stepped wedge cluster randomized trials are cluster randomized trials in which clusters cross from the control treatment to the investigational treatment at fixed times over the course

of the study. Most commonly, clusters start with the control treatment and later cross over to the investigational treatment, with all clusters ending on the

¹Department of Biostatistics, University of Washington, Seattle, WA, USA

²Division of Biostatistics, The Ohio State University, Columbus, OH, USA

Corresponding author:

Siobhan P Brown, Department of Biostatistics, University of Washington, 6200 NE 74th Street, Building 29, Suite 250, Seattle, WA 98115, USA.
Email: spes@uw.edu

investigational arm. The order in which clusters cross over is randomly determined at the start of the study. Because each cluster serves as its own control, such studies are often more efficient than parallel cluster designs.^{1–4} Stepped wedge trials may be cohort studies or cross-sectional; we deal with the latter case here. This trial design is especially useful when the intervention is expected to have a carryover effect, limiting the usefulness of a cluster-crossover study, or when logistical constraints limit the number of clusters in which the intervention can be implemented at one time.⁵ Often, the design is used to evaluate a promising intervention in a “real-world” setting, as all clusters eventually receive the intervention.⁶ Stepped wedge trials are increasingly popular, especially increasing in use since 2010 as illustrated dramatically by Grayling et al. in their 2017 *Trials* paper⁷; see also.^{5,6,8–11} Stepped wedge cluster randomized trials are an active area of research; much has been written on the advantages and disadvantages of their use,^{5,12,13} design and implementation,^{1,14–20} and analysis.^{21–26}

Because stepped wedge trials are cluster randomized trials, it is extremely important that the analysis of data from these studies accounts for the correlation of observations, usually with a mixed-effects model or generalized estimating equations. Cluster-level outcomes can also be used, though this approach is less common and results in a loss of efficiency unless the clusters are all of the same size.^{23,26–28} Because of the way the study treatment is rolled out, time is partially confounded with treatment effect and so must be included in the model, often by incorporating a fixed period effect.²² Period-by-cluster interactions can also be considered.²⁷

Sequential monitoring

Group sequential monitoring methods are used to conduct interim analyses while preserving study characteristics such as the Type I error and power.^{29,30} They allow for early stopping of a trial if the experimental therapy is demonstrably better or worse than the control, or if continuing the trial is unlikely to result in a statistically significant result (i.e. stopping for futility). These and related methods are especially important when the outcome of interest is mortality or severe morbidity, where one has an ethical imperative to stop the trial if either treatment is superior to the other.^{29–31} In addition, formal interim analyses can create significant efficiency gains for a research program; stopping trials once the conclusion is foregone reduces the average sample size and cost, while allowing other trials to begin earlier than would otherwise be possible.

Setting appropriate boundaries for an interim analysis relies on specification of the Type I error at each interim analysis (i.e. the error spending function) and the amount of statistical information available at that time relative to the expected final information (the

information fraction). Boundaries are typically extreme early in the study (with less information), appropriately requiring extreme estimates to stop the study for efficacy or futility, and are less extreme as information increases later in the study. At each analysis, the estimated information is crucial for determining the actual boundaries to be used. Overestimated information will lead to using boundaries less extreme than needed, potentially spuriously stopping too early for either superiority or futility/harm, thus increasing the Type I error and losing power. Conversely, with underestimated information, the boundaries used at the interim analysis will be more extreme than intended, leading to the trial continuing when it could be stopped, increasing the average sample size.

In a stepped wedge trial, the information fraction available at an interim analysis depends on the typically unknown variance components through the intracluster correlation coefficient (ICC; the portion of total variance attributable to between-cluster differences). Uniquely, no information is gained in the first period when all clusters are on the standard treatment nor in the last period when all clusters are on the experimental treatment while a period effect is used; optimal designs do not include any periods with all clusters on the same treatment arm.²⁰

Grayling et al.³² laid out a theoretical framework and justification for using group sequential methods in stepped wedge trials with the error spending approach, and found substantial gains in efficiency. However, their work used known variance components or fixed ICC, a setting where the information would be known at each analysis time, usually not the case in practice. More recently, Grayling et al.³³ published an examination of sample size reestimation in this setting when the variance components are unknown. This approach can preserve power in the case either variance component was greatly underestimated at the design phase, though it requires that the number of subjects enrolled in each cluster during each time period be increased and the final sample size may need to be greatly increased. In this article, we set out to evaluate how the study design impacts the rate of information growth and how misspecification of that information growth impacts the operating characteristics of group sequential monitoring methods.

Methods

We simulated cross-sectional stepped wedge study data in order to evaluate the impact of sequential monitoring on the Type I error and power when the ICC is unknown. We designed a study with 10 clusters, 6 time periods, and 10 subjects per cluster in each time period. All clusters begin the study on the control arm; two clusters switch to the interventional arm at the end of

each time period so that all clusters receive the interventional treatment in the final period. That results in a total of 600 subjects enrolled in the study, 300 on each of the treatment arms.

Let $i = 1, \dots, 10$ indicate the cluster, $j = 1, \dots, 6$ the period, and $k = 1, \dots, 10$ indicate the subject within each cluster period. With cluster effect $\alpha_i \sim N(0, \tau^2)$ and $\varepsilon_{ijk} \sim N(0, \sigma^2)$, the individual response was $y_{ijk} = \alpha_i + X_{ij}\theta + \varepsilon_{ijk}$, where X_{ij} is the indicator for treatment in the i th cluster during period j and θ is the treatment effect. The ICC is defined as $\rho = \tau^2/(\tau^2 + \sigma^2)$. Power calculations and the calculation of the information fraction after each period are based on the Wald test with known values of σ^2 and ρ using the weighted least squares as presented by Hussey and Hughes.²² In their notation

$$\text{Var}(\hat{\theta}) = \frac{I \frac{\sigma^2}{N} \left(\frac{\sigma^2}{N} + T\tau^2 \right)}{(IU - W) \frac{\sigma^2}{N} + (U^2 + ITU - TW - IV)\tau^2} \quad (1)$$

with I clusters, T time points, $U = \sum_{ij} X_{ij}$, $W = \sum_j (\sum_i X_{ij})^2$, and $V = \sum_i (\sum_j X_{ij})^2$,

Note that, without loss of generality, we assumed that the data generation model did not have a period effect. We used four different values for ρ : 0, 0.01, 0.05, and 0.20. Those represent trials with no, small, intermediate, and large ICC, respectively.³⁴ We set $\sigma^2 = 1 - \rho$ to keep the variability of responses constant across simulation settings (implying that $\tau^2 = \rho$). Three different treatment effects were evaluated: $\theta = 0, 0.25$, and 0.43 . Thus, we were able to evaluate performance with a null effect; with an intermediate level of power, where departures from the model assumptions were likely to have the largest impact (estimated power ranged from 47% to 61% depending on ρ); with an estimated 90% power for $\rho = 0.05$, which is a more realistic trial design.

The outcome was modeled with a linear mixed-effects model (LMM) with the restricted log likelihood criterion (REML). The model included treatment and period effects (as factors); no treatment-by-period interaction was used. With no interim monitoring, the Kenward–Roger approximation to the degrees of freedom was necessary to preserve the Type I error, consistent with results in other cluster randomized studies.³⁵ For that reason, the Kenward–Roger approximation was used for all interim analyses; at each interim analysis, the Kenward–Roger adjusted p -value was compared to the monitoring boundaries to decide whether the study should stop or continue. In a recent dissertation, Tanner explored the performance of various degree-of-freedom adjustments and bias corrections in stepped wedge trials with a small number of clusters and unbalanced design; some of these methods may be more appropriate than Kenward–Roger for a particular planned study.³⁶

We calculated the information growth with three methods. The first, the naïve approach, used the period number relative to the total number of periods (i.e. j/J); this approach is equivalent to using the current sample size relative to the maximum planned sample size. In the second, we used a prespecified value of the ICC (ρ) and examined the performance of the analysis with both correctly and incorrectly specified values of the ICC. For the third approach, we estimated the ICC from the estimated variance components from the LMM model fit at the interim analysis. Each simulation was repeated 10,000 times; thus, with a true Type I error rate of 0.025, we expect the observed value to fall between 0.022 and 0.028 95% of the time (i.e. a Monte-Carlo predictive interval of $0.025 \pm \sqrt{(0.025 \times 0.975/10,000)}$).³⁷

We specified a monitoring plan that used a single interim analysis, after the fourth period. At that time, 6 of the 10 clusters will have subjects who have received the experimental therapies; all clusters will have at least 10 subjects who received the control therapy. We looked at both the Pocock and O’Brien–Fleming boundary shapes, as most boundaries fall somewhere between the two.^{31,38,39} The p -value scale was used for constraints.³⁹ Sequential boundaries were generated using SeqTrial for R (<http://www.rctdesign.org/Software.html>) (Figure 1). We investigated early stopping for efficacy only as well as early stopping for both efficacy and futility (i.e. stopping once a meaningful benefit has been ruled out).

Results

Table 1 gives the information fraction at the end of each period for this study design across a range of true ICC values. The information fraction was calculated using the approximate variance of the estimated treatment effect given in equation (1) above into $I(j) = \text{Var}(\hat{\theta}_j)/\text{Var}(\hat{\theta}_j)$ for each time period $j = 1, \dots, 6$. Note that the information depends on ρ but not the error variance in this context. We found that the study design has a large impact on the information, even with an ICC of 0. Generally, higher ICC leads to slower accumulation of information. At the end of the fourth period, when we performed the interim analysis, 67% of study subjects have been enrolled, but the true information fraction varies from 0.65 to 0.80 depending on the ICC. At the halfway point in the trial, the information fractions are a bit closer together, ranging from 0.41 to 0.52 for the values of ICC that were considered. Table 2 presents the Type I error and power observed in the simulations for both boundary shapes.

The main concern with overestimated information fraction is inflation of the Type I error, which we saw only to a limited degree. The highest observed Type I error rate was 0.029, just above what would be expected with a true Type I error rate of 0.025, while the lowest

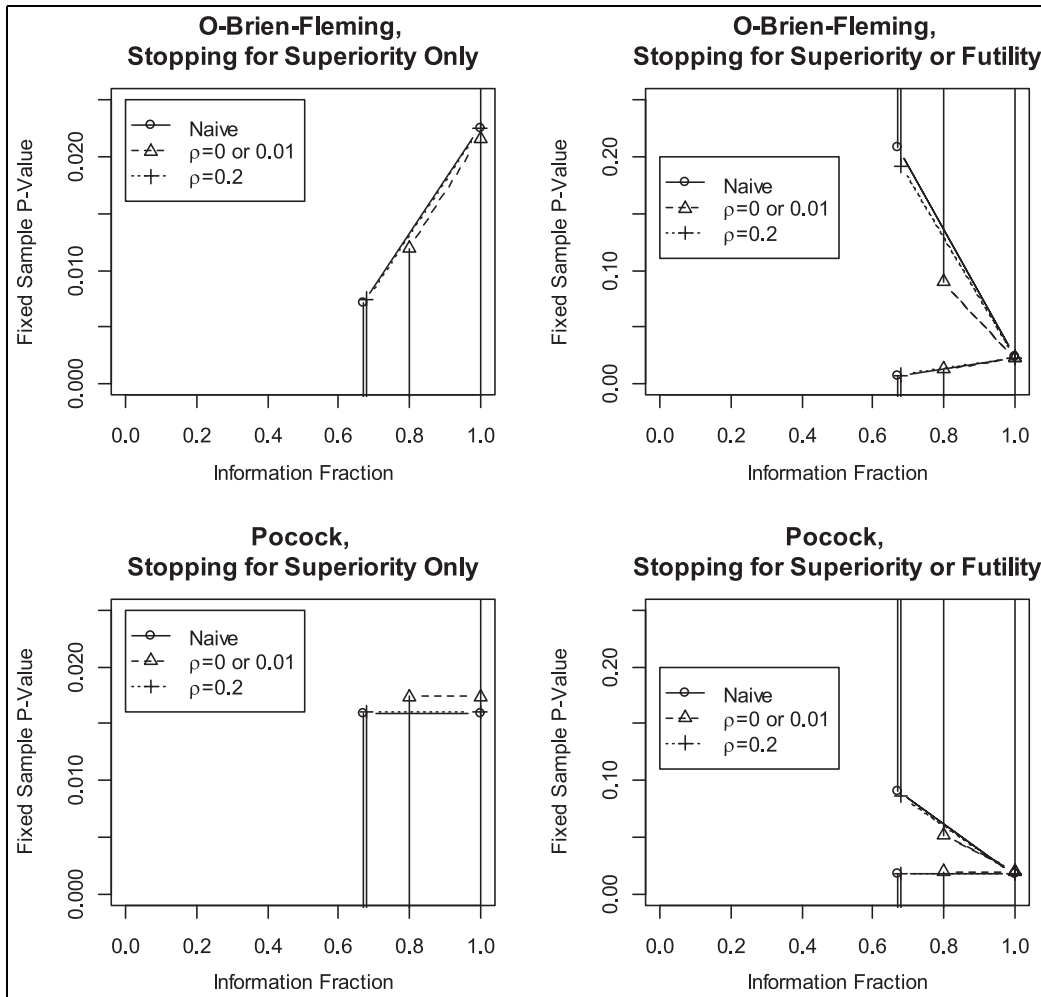


Figure 1. Sequential monitoring boundaries for a stepped wedge study.

Table 1. Information fraction after each period for different true ICCs.

Period	Proportion of final enrollment	ICC				
		$\rho = 0$	$\rho = 0.01$	$\rho = 0.05$	$\rho = 0.20$	$\rho = 0.50$
1	0.17	0.00	0.00	0.00	0.00	0.00
2	0.33	0.20	0.22	0.22	0.19	0.18
3	0.50	0.50	0.52	0.49	0.43	0.41
4	0.67	0.80	0.80	0.75	0.68	0.65
5	0.83	1.00	0.99	0.93	0.88	0.86
6	1.00	1.00	1.00	1.00	1.00	1.00

ICC: intraclass correlation coefficient.

was 0.013. This slight conservatism appears to result from the use of the Kenward–Roger adjustment, rather than the sequential monitoring or information misspecification, as the error rates are similar when monitoring is not used. The simulation setting with the highest potential to see an inflated Type 1 error is when the true ρ value is 0.20, but the information growth under $\rho = 0$ or 0.01 is used (i.e. the setting where information is most highly overestimated). In those settings, the

highest Type I error rates are observed: 0.029 with stopping for efficacy only and either shape; and 0.026 and 0.027 with stopping for both efficacy and futility. Those error rates are not consistent with a true Type I error rate of 0.025; they are also significantly higher than the levels found with no testing, though the absolute increase is relatively small.

We saw a reduction in power when a futility boundary was used and the information is overestimated

Table 3. Average sample size.

Stopping for alternative only														
			$\theta = 0.25$				$\theta = 0.43$							
$\theta = 0$			$\rho = 0$	$\rho = 0.01$	$\rho = 0.05$	$\rho = 0.20$	$\rho = 0$	$\rho = 0.01$	$\rho = 0.05$	$\rho = 0.20$	$\rho = 0$	$\rho = 0.01$	$\rho = 0.05$	$\rho = 0.20$
No monitoring		600	600	600	600	600	600	600	600	600	600	600	600	600
Naïve information fraction														
O'Brien-Fleming		599.6	599.4	598.8	598.4	557.7	559.2	561.6	560.8	455.5	462.8	480.5	479.7	
Pocock		598.6	598.1	597.1	596.6	534.0	536.5	542.4	542.0	433.6	440.8	457.0	457.2	
Information growth with $\rho = 0$ or 0.01 ^a														
O'Brien-Fleming		599.2	598.7	597.9	597.3	543.8	545.6	549.4	549.5	440.6	448.7	465.3	465.3	
Pocock		598.4	597.9	596.8	596.3	531.2	533.3	540.3	539.3	431.5	438.3	454.6	454.8	
Information growth with $\rho = 0.20$														
O'Brien-Fleming		599.6	599.3	598.7	598.3	556.1	557.7	560.4	559.8	453.7	461.3	478.7	477.8	
Pocock		598.5	598.1	597.0	596.6	533.6	536.2	542.2	541.7	433.4	440.6	456.7	456.9	
Estimate ρ														
O'Brien-Fleming		599.6	599.4	598.8	598.4	557.7	559.2	561.6	560.8	455.5	462.8	480.5	479.7	
Pocock		598.6	598.1	597.1	596.6	534.0	536.5	542.4	542.0	433.6	440.8	457.0	457.2	
Stopping for alternative or fertility														
			$\theta = 0.25$				$\theta = 0.25$							
$\theta = 0$			$\rho = 0$	$\rho = 0.01$	$\rho = 0.05$	$\rho = 0.20$	$\rho = 0$	$\rho = 0.01$	$\rho = 0.05$	$\rho = 0.20$	$\rho = 0$	$\rho = 0.01$	$\rho = 0.05$	$\rho = 0.20$
No monitoring		600	600	600	600	600	600	600	600	600	600	600	600	600
Naïve information fraction														
O'Brien-Fleming		437.4	438.2	441.7	440.2	528.4	525.5	516.7	517.1	453.0	459.5	473.6	472.6	
Pocock		412.6	413.0	415.8	414.8	470.0	466.2	459.7	458.5	425.8	429.8	438.4	437.9	
Information growth with $\rho = 0$ or 0.01 ^a														
O'Brien-Fleming		413.3	413.4	416.4	415.6	480.3	476.2	467.5	466.4	433.3	438.5	447.0	446.3	
Pocock		405.4	405.5	407.2	406.0	441.6	437.7	433.7	434.1	417.3	419.5	424.3	424.0	
Information growth with $\rho = 0.20$														
O'Brien-Fleming		434.4	435.1	438.8	437.2	524.1	521.0	512.3	512.1	450.6	457.3	471.3	470.0	
Pocock		411.7	411.9	414.5	413.5	467.3	463.1	457.0	455.8	425.0	428.7	437.6	436.4	
Estimate ρ														
O'Brien-Fleming		437.4	438.2	441.7	440.2	528.4	525.5	516.7	517.1	453.0	459.5	473.6	472.6	
Pocock		412.6	413.0	415.8	414.8	470.0	466.2	459.7	458.5	425.8	429.8	438.4	437.9	

^aInformation after the fourth period is the same for either value of ρ .

(i.e. the true $\rho = 0.20$ but the information growth for $\rho = 0$ or 0.01 is used; this occurred to a lesser degree with $\rho = 0.05$ and the information growth under $\rho = 0$ or 0.01 is used). In those cases, we observed an absolute reduction in power up to 8% from the correctly specified information fraction. The biggest difference was seen with intermediate treatment effect and the Pocock boundary shape; using the overestimated information (information growth based on $\rho = 0$ or 0.01 when the true ρ value is 0.20) reduced the observed study power to 0.416 from 0.496 . Those reductions in power are substantial from a statistical perspective, though their practical impact can be debated for most settings.

We saw a modest increase in the average sample size with underestimated information, which would be the case with no or low ICC but boundaries that use the naïve approach or information growth assuming $\rho = 0.20$ (Table 3). With the O'Brien–Fleming boundaries, the average sample size increased 5%–10% in that scenario. The impact is smaller using the Pocock boundaries, where the observed increase was 0%–6%. Using the estimated ICC resulted in a similar loss of efficiency. As expected, using any monitoring boundary greatly reduced the average sample size, sometimes dramatically. The use of a futility boundary resulted in large efficiency gains with the intermediate treatment effect. In the case of the larger treatment effect, stopping for the alternative resulted in similar efficiencies. Reductions of the average sample size by 30% or more were common in both those settings.

Conclusion

The study design significantly impacts the information growth in stepped wedge trials, particularly early and late in the study. However, using a misspecified information fraction at the interim analysis did not have a large adverse impact on either the Type I error or the power of our simulated trials, though it did impact the average sample size. These results suggest that planning a single interim analysis approximately midway through a stepped wedge design can be safely implemented if researchers avoid a naïve estimate of the information fraction and use the estimated ICC at the time of the interim analysis unless a reasonable prespecified value is available. Another approach is to note that, for the sample design studied in the simulation and the more realistic values of the ICC used, the true information fractions at the end of the third period are relatively close; if the same holds for a time point and range of ICC values of a proposed trial, the timing of the interim analysis could be selected to minimize the impact of the ICC on the analysis. In the simulations, the information fraction depends only on the ICC and so the Type I error rate can be maintained so long as

the ICC is not dramatically overestimated at the interim analysis. However, if there is substantial uncertainty in general regarding the overall variance, researchers would be wise to consider sample size reestimation as part of a planned interim analysis to avoid potential loss of power.³⁵

In this brief report, we have only explicitly tested a limited set of design parameters, so these results may not be generalizable to all stepped wedge designs. In particular, unbalanced stepped wedge designs may show more variability in information growth as a function of the ICC. With an unbalanced design, particularly one with a small number of clusters, alternative degree-of-freedom adjustments or bias correction methods may be needed to maintain model performance in place of the Kenward–Roger method used here.³⁶ In addition, we considered only a single interim analysis, as that approach would be easier to do in practice than multiple interim analyses. Future work is needed to evaluate the cumulative impact of multiple interim analyses with a poorly estimated ICC in the stepped wedge design setting.


Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Siobhan P Brown  <https://orcid.org/0000-0002-4774-3122>

References

1. Girling AJ and Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Stat Med* 2016; 35(13): 2149–2166.
2. Hemming K and Girling A. The efficiency of stepped wedge vs. cluster randomized trials: stepped wedge studies do not always require a smaller sample size. *J Clin Epidemiol* 2013; 66(12): 1427–1428.
3. Hemming K and Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *J Clin Epidemiol* 2016; 69: 137–146.
4. Woertman W, de Hoop E, Moerbeek M, et al. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol* 2013; 66(7): 752–758.
5. Brown CA and Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol* 2006; 6: 54.
6. Mdege ND, Man MS, Taylor CA, et al. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions

- during routine implementation. *J Clin Epidemiol* 2011; 64(9): 936–948.
7. Grayling MJ, Wason JMS and Mander AP. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. *Trials* 2017; 18(1): 33.
 8. Martin J, Taljaard M, Girling A, et al. Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. *BMJ Open* 2016; 6: e010166.
 9. Taljaard M, Hemming K, Shah L, et al. Inadequacy of ethical conduct and reporting of stepped wedge cluster randomized trials: results from a systematic review. *Clin Trials* 2017; 14(4): 333–341.
 10. Barker D, McElduff P, D'Este C, et al. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. *BMC Med Res Methodol* 2016; 16: 69.
 11. Beard E, Lewis JJ, Copas A, et al. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials* 2015; 16: 353.
 12. de Hoop E, van der Tweel I, van der Graaf R, et al. The need to balance merits and limitations from different disciplines when considering the stepped wedge cluster randomized trial design. *BMC Med Res Methodol* 2015; 15: 93.
 13. Zhan Z, van den Heuvel ER, Doornbos PM, et al. Strengths and weaknesses of a stepped wedge cluster randomized design: its application in a colorectal cancer follow-up study. *J Clin Epidemiol* 2014; 67(4): 454–461.
 14. Baio G, Copas A, Ambler G, et al. Sample size calculation for a stepped wedge trial. *Trials* 2015; 16: 354.
 15. Copas AJ, Lewis JJ, Thompson JA, et al. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials* 2015; 16: 352.
 16. Hemming K, Haines TP, Chilton PJ, et al. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* 2015; 350: h391.
 17. Hemming K, Lilford R and Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Stat Med* 2015; 34(2): 181–196.
 18. Hooper R, Teerenstra S, de Hoop E, et al. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Stat Med* 2016; 35(26): 4718–4728.
 19. Lawrie J, Carlin JB and Forbes AB. Optimal stepped wedge designs. *Stat Probabil Lett* 2015; 99: 210–214.
 20. Thompson JA, Fielding K, Hargreaves J, et al. The optimal design of stepped wedge trials with equal allocation to sequences and a comparison to other trial designs. *Clin Trials* 2017; 14(6): 639–647.
 21. Cook AJ, DeLong E, Murray DM, et al. Statistical lessons learned for designing cluster randomized pragmatic clinical trials from the NIH Health Care Systems Collaboratory Biostatistics and Design Core. *Clin Trials* 2016; 13(5): 504–512.
 22. Hussey MA and Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007; 28(2): 182–191.
 23. Scott JM, deCamp A, Juraska M, et al. 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.
 24. Wang R and De Gruttola V. The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials. *Stat Med* 2017; 36(18): 2831–2843.
 25. Davey C, Hargreaves J, Thompson JA, et al. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials* 2015; 16: 358.
 26. Barker D, D'Este C, Campbell MJ, et al. 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(1): 119.
 27. Thompson JA, Fielding KL, Davey C, et al. Bias and inference from misspecified mixed-effect models in stepped wedge trial analysis. *Stat Med* 2017; 36(23): 3670–3682.
 28. Hughes JP, Granston TS and Heagerty PJ. Current issues in the design and analysis of stepped wedge trials. *Contemp Clin Trials* 2015; 45(Pt A): 55–60.
 29. Ellenberg SS, Fleming TR and DeMets DL. *Data monitoring committees in clinical trials: a practical perspective*. Chichester; Hoboken, NJ: John Wiley & Sons, 2002.
 30. Friedman LM, Furberg C and DeMets DL. *Fundamentals of clinical trials*. New York: Springer, 2010.
 31. Pocock SJ. Group sequential methods in design and analysis of clinical trials. *Biometrika* 1977; 64(2): 191–199.
 32. Grayling MJ, Wason JM and Mander AP. Group sequential designs for stepped-wedge cluster randomised trials. *Clin Trials* 2017; 14(5): 507–517.
 33. Grayling MJ, Mander AP and Wason JMS. Blinded and unblinded sample size reestimation procedures for stepped-wedge cluster randomized trials. *Biom J* 2018; 60(5): 903–916.
 34. Murray DM and Blistein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. *Eval Rev* 2003; 27(1): 79–103.
 35. Kenward MG and Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; 53(3): 983–997.
 36. Tanner WF. *Improved standard error estimation for maintaining the validities of inference in small-sample cluster randomized trials and longitudinal studies*. PhD Thesis, University of Kentucky, Lexington, KY, 2018.
 37. Koehler E, Brown E and Haneuse SJ. On the assessment of Monte Carlo error in simulation-based statistical analyses. *Am Stat* 2009; 63(2): 155–162.
 38. O'Brien PC and Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; 35(3): 549–556.
 39. Burington BE and Emerson SS. Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics* 2003; 59(4): 770–777.