Article

CLINICAL TRIALS

Clinical Trials
2017, Vol. 14(5) 507–517

DOI: 10.1177/1740774517716937
journals.sagepub.com/home/ctj
SAGE

Group sequential designs for stepped-wedge cluster randomised trials

Michael J Grayling, James MS Wason and Adrian P Mander

Abstract

Background/Aims: The stepped-wedge cluster randomised trial design has received substantial attention in recent years. Although various extensions to the original design have been proposed, no guidance is available on the design of stepped-wedge cluster randomised trials with interim analyses. In an individually randomised trial setting, group sequential methods can provide notable efficiency gains and ethical benefits. We address this by discussing how established group sequential methodology can be adapted for stepped-wedge designs.

Methods: Utilising the error spending approach to group sequential trial design, we detail the assumptions required for the determination of stepped-wedge cluster randomised trials with interim analyses. We consider early stopping for efficacy, futility, or efficacy and futility. We describe first how this can be done for any specified linear mixed model for data analysis. We then focus on one particular commonly utilised model and, using a recently completed stepped-wedge cluster randomised trial, compare the performance of several designs with interim analyses to the classical stepped-wedge design. Finally, the performance of a quantile substitution procedure for dealing with the case of unknown variance is explored.

Results: We demonstrate that the incorporation of early stopping in stepped-wedge cluster randomised trial designs could reduce the expected sample size under the null and alternative hypotheses by up to 31% and 22%, respectively, with no cost to the trial's type-I and type-II error rates. The use of restricted error maximum likelihood estimation was found to be more important than quantile substitution for controlling the type-I error rate.

Conclusion: The addition of interim analyses into stepped-wedge cluster randomised trials could help guard against time-consuming trials conducted on poor performing treatments and also help expedite the implementation of efficacious treatments. In future, trialists should consider incorporating early stopping of some kind into stepped-wedge cluster randomised trials according to the needs of the particular trial.

Keywords

Stepped wedge, cluster randomised trial, group sequential, interim analyses, error spending

Introduction

In a stepped-wedge (SW) cluster randomised trial (CRT), an intervention is introduced across several time periods, with the time period in which a cluster begins receiving the experimental intervention assigned at random. Although the SW-CRT design was actually first proposed over 30 years ago,[1] it has only been in recent years that it has gained substantial attention in the trials community.

Numerous papers have now been published containing new research on the design. Methodology[2–5] and software[6] now exist to determine required sample sizes, and several results on optimal SW-CRT designs have been established,[7,8] while extensions to the standard design to allow for multiple levels of clustering have also been presented.[9]

However, as has been noted, little is known about the design of SW-CRTs with interim analyses.[10] In an individually randomised trial setting, it has been well established that group sequential methods can bring

MRC Biostatistics Unit Hub for Trials Methodology Research, Cambridge Institute of Public Health, Cambridge, UK

Corresponding author:
Michael J Grayling, MRC Biostatistics Unit Hub for Trials Methodology Research, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge CB2 0SR, UK.
Email: mjg211@cam.ac.uk

substantial ethical benefits and efficiency gains to a trial.[11] Explicitly, allowing the early stopping of a trial for either efficacy or futility can reduce the number of patients administered an inferior intervention and allow efficacious interventions to either move to later phase testing or to be rolled out across a population with greater speed. Given that SW-CRTs can be highly expensive because of the large number of time periods and measurements they can require, it would be advantageous to be able to incorporate interim analyses into the design.

In this article, we present methodology for establishing such designs. We then conclude with a discussion of the practical and methodological considerations associated with the use of interim analyses.

Methods

Notation, hypotheses, and analysis

We assume that a SW-CRT is to be carried out on C clusters over T time periods, with m measurements per cluster per time period. We do not make a distinction as to whether across the time periods these m measurements are on different patients (a cross-sectional design) or the same patients (a cohort design). Moreover, we note that our methodology could be easily extended to allow the number of measurements per cluster to vary across the time periods according to some pre-specified rule. For simplicity, we do restrict focus to the classical case of a 'balanced complete-block' SW-CRT however. In this case, a single experimental intervention is compared to a single control or placebo, each cluster is present in every time period, each cluster begins in the control condition and finishes in the experimental condition, and an equal (or as equal as possible) number of clusters switch to the intervention in each time period.

We next assume that the accrued data from our SW-CRT trial will be normally distributed, and a linear mixed model has been specified for analysis as

$$\boldsymbol{y} = D\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{\epsilon}$$

where

- $\boldsymbol{y}$ is the vector of responses, that is, the measurements taken as part of the trial;
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of p fixed effects. For example, this may commonly contain among other factors fixed effects for the time period;
- D is the design matrix which links $\boldsymbol{y}$ to $\boldsymbol{\beta}$. That is, D ensures the correct fixed effects are included in the formulae for each measurement;
- $\boldsymbol{u}$ is a vector of random effects which follows a specified multivariate normal distribution, $\boldsymbol{u} \sim N(\boldsymbol{0}, G)$. Commonly, one may expect $\boldsymbol{u}$ to contain random effects for cluster for example;
- Z is the design matrix which links $\boldsymbol{y}$ to $\boldsymbol{u}$, that is, it performs the same job for $\boldsymbol{u}$ as D does for $\boldsymbol{y}$;
- $\boldsymbol{\epsilon}$ is a vector of residuals which follows a specified multivariate normal distribution, $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, R)$. That is, $\boldsymbol{\epsilon}$ accounts for the variation in the measurements not explained by the fixed and random effects.

Note that, in particular, it is the prescribed $\boldsymbol{u}$, G, and Z that would likely differ for cross-sectional and cohort designs.

We assume the final element of $\boldsymbol{\beta}$, β_p, is our parameter of interest: the direct effect of the experimental intervention relative to the control. We denote this element for brevity by τ and test the following one-sided hypotheses

$$H_0 : \tau \leq 0, \quad H_1 : \tau > 0$$

Moreover, we assume it is desired to control the type-I error rate of this test to some level α when $\tau = 0$ and to have power to reject H_0 of at least $1 - \beta$ when $\tau = \delta$. Note that the determination of SW-CRT designs for two-sided hypotheses is also easily achievable using our methods. In addition, note that by the above we are not considering treatment by period interactions.

We specify a set of integers $\mathscr{T} = \{t_1, \ldots, t_{|\mathscr{T}|}\}$, with $t_1 \geq 1$ and $t_{|\mathscr{T}|} = T$. With this set, we employ interim analyses after time periods $t \in \mathscr{T}$. For example, $\mathscr{T} = \{2, 3, 5\}$ would imply interim analyses were to be conducted in a trial after time periods 2, 3, and 5. Note that we always schedule an analysis at the end of the trial, that is, after time period T. Furthermore, it is important to ensure that after time period t_1 at least one cluster has received the experimental intervention in some time period, otherwise estimating τ is impossible.

With the above, $\boldsymbol{y} \sim N(D\boldsymbol{\beta}, \Sigma)$, where $\Sigma = ZGZ^T + R$, and after each period $t \in \mathscr{T}$, we acquire an estimate $\hat{\boldsymbol{\beta}}_t = (\hat{\beta}_{1t}, \ldots, \hat{\beta}_{pt})^T$ for $\boldsymbol{\beta}$ through the standard maximum likelihood (ML) estimator of a linear mixed model, the generalised least squares estimate

$$\hat{\boldsymbol{\beta}}_t = (D_t^T \Sigma_t^{-1} D_t)^{-1} D_t^T \Sigma_t^{-1} \boldsymbol{y}_t$$

Here, the subscript t indices indicate that we are considering response data accrued in the first t time periods and the associated implied design matrices.

Following the notation of Jennison and Turnbull,[11] we acquire $\hat{\tau}_t = \hat{\boldsymbol{\beta}}_{pt}$, our estimate for τ. This leads to the following standardised Wald test statistic

$$Z_t = \frac{\hat{\tau}_t}{\sqrt{\mathrm{var}(\hat{\tau}_t)}} = \hat{\tau}_t I_t^{1/2}$$

where

$$I_t = \{(D_t^T \Sigma_t^{-1} D_t)^{-1}_{[p,p]}\}^{-1}$$

is the information for τ after time period t.

Note that all of $\boldsymbol{\beta}$ may not be estimable at each analysis $t \in \mathscr{T}$. However, by our requirement above that at least one cluster has received the experimental intervention in one of the time periods $1, \ldots, t_1$, it will always be possible to estimate τ. In this case, the matrix inverses in the formulae for $\hat{\boldsymbol{\beta}}_t$ and $I_t^{1/2}$ above should be interpreted as generalised inverses.[11]

While group sequential methodology is typically associated with designs with an independent increment structure, the important results hold for the more general scenario utilising linear mixed models considered here. In particular, we have that[11]

$$\mathbb{E}(Z_t) = \tau I_t^{1/2}, \qquad t \in \mathscr{T}$$

$$\text{cov}(Z_{t_i}, Z_{t_j}) = (I_{t_i}/I_{t_j})^{1/2}, \qquad t_i, t_j \in \mathscr{T}, t_i \leq t_j$$

Finally, given futility and efficacy bounds, $f_{t_1}, \ldots, f_{t_{|\mathscr{T}|}}$ and $e_{t_1}, \ldots, e_{t_{|\mathscr{T}|}}$, respectively, $f_t \leq e_t$ for all t, the following stopping rules are employed

- For $t \in \{1, \ldots, T-1\}$
  - If $t \notin \mathscr{T}$ continue through to the end of time period $t+1$, since stopping is only permitted at our pre-specified times;
  - If $t \in \mathscr{T}$
    * if $Z_t \leq f_t$ stop the trial and accept H_0;
    * if $Z_t > e_t$ stop the trial and reject H_0;
    * if $f_t < Z_t \leq e_t$ continue through to the end of period $t+1$.

- For $t = T$
  - if $Z_t \leq f_t$ accept H_0;
  - if $Z_t > e_t$ reject H_0.

We denote by $\omega_R \in \mathscr{T}$ the interim analysis at which the trial is stopped and by ψ_R the reason for stopping. That is, $\psi_R = 1$ if H_0 is rejected and is 0 otherwise. Before a trial, ω_R and ψ_R are random variables. We can, however, compute the probability $\omega_R = \omega$ and $\psi_R = \psi$ for any true treatment effect τ through the following integral

$$\mathbb{P}(\omega_R = \omega, \psi_R = \psi | \tau) = \int_{\text{l}(1,\omega,\psi)}^{\text{u}(1,\omega,\psi)} \cdots \int_{\text{l}(t_{|\mathscr{T}|},\omega,\psi)}^{\text{u}(t_{|\mathscr{T}|},\omega,\psi)} \phi\left\{\boldsymbol{x}, \boldsymbol{r}(\tau, |\mathscr{T}|) \circ \boldsymbol{I}^{1/2}, \Lambda\right\} \text{d}x_{|\mathscr{T}|} \ldots \text{d}x_1$$

where

- $\phi\{\boldsymbol{x}, \boldsymbol{\mu}, \Lambda\}$ is the probability density function of a multivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^T$ and covariance matrix Λ, $\dim(\Lambda) = k \times k$, evaluated at vector $\boldsymbol{x} = (x_1, \ldots, x_k)^T$;
- $\boldsymbol{r}(a, b) = (a, \ldots, a)^T$ is the vector formed from repeating a b times;
- $\circ$ denotes the Hadamard product of two vectors, that is, $(a_1, \ldots, a_n)^T \circ (b_1, \ldots, b_n)^T = (a_1 b_1, \ldots, a_n b_n)^T$;
- $\boldsymbol{I} = (I_{t_1}, \ldots, I_{t_{|\mathscr{T}|}})^T$ is the vector of information levels for the estimated treatment effects across the interim analyses, and its square root is taken in an element wise manner. That is, $\{(I_{t_1}, \ldots, I_{t_{|\mathscr{T}|}})^T\}^{1/2} = (I_{t_1}^{1/2}, \ldots, I_{t_{|\mathscr{T}|}}^{1/2})^T$;
- l and u are functions that tell us the lower and upper integration limits for the test statistic Z_t given values for t, ω, and ψ. For example, $\text{l}(1,2,1) = f_1$ and $\text{u}(1,2,1) = e_1$, while $\text{l}(2,2,1) = e_2$ and $\text{u}(2,2,1) = \infty$;
- Λ is the covariance matrix of the standardised test statistics at and across each interim analysis, that is

$$\Lambda = \begin{pmatrix} \text{cov}(Z_{t_1}, Z_{t_1}) & \ldots & \text{cov}(Z_{t_1}, Z_{t_{|\mathscr{T}|}}) \\ \vdots & \ddots & \vdots \\ \text{cov}(Z_{t_{|\mathscr{T}|}}, Z_{t_1}) & \ldots & \text{cov}(Z_{t_{|\mathscr{T}|}}, Z_{t_{|\mathscr{T}|}}) \end{pmatrix}$$

The probability that H_0 is rejected for any τ can then be computed as

$$\mathbb{P}(\text{Reject } H_0 | \tau) = \sum_{\omega \in \mathscr{T}} \mathbb{P}(\omega_R = \omega, \psi_R = 1 | \tau)$$

Moreover, we can determine the expected number of measurements that would be required by a design for any τ using the following formulae

$$\mathbb{E}(M | \tau) = \sum_{\omega \in \mathscr{T}} \sum_{\psi \in \{0,1\}} mC\omega \mathbb{P}(\omega_R = \omega, \psi_R = \psi | \tau)$$

Here, $mC\omega$ is the number of measurements that are required when the trial stop after time period ω. With the above, the operating characteristics of any specified SW-CRT with interim analyses can be determined. However, at the design stage, one needs to be able to ascertain values for C, m, T, the f_t, and e_t, to convey desired operating characteristics. This is achieved here using error spending methodology as discussed in the following section.

Error spending

Numerous procedures have today been proposed for the determination of group sequential trial designs. One of the earliest and most flexible such methods is the error spending approach.[12] In this case, functions f and e are used to determine the amount of type-I and type-II error 'spent' at a particular interim analysis. Here, as an example, we utilise this approach,

specifically employing a family of spending functions indexed by parameters γ_f and γ_e given by

$$\mathrm{e}(z) = \alpha z^{\gamma_e}$$
$$\mathrm{f}(z) = \beta z^{\gamma_f}$$

We define

$$\pi_{\{1,t\}} = \mathbb{P}(\omega_R = t, \psi_R = 1 \mid 0),$$
$$\pi_{\{2,t\}} = \mathbb{P}(\omega_R = t, \psi_R = 0 \mid \delta).$$

Thus, $\pi_{\{1,t\}}$ and $\pi_{\{2,t\}}$ are the probabilities of committing a type-I and type-II error, respectively, after time period t.

Then, for given choices of C, T and m, the values of the f_t and e_t are found iteratively as the solutions to

$$\pi_{\{1,t_1\}} = \mathrm{e}(I_{t_1}/I_{t_{|\mathscr{T}|}})$$
$$\pi_{\{2,t_1\}} = \mathrm{f}(I_{t_1}/I_{t_{|\mathscr{T}|}})$$

and

$$\pi_{\{1,t_i\}} = \mathrm{e}(I_{t_i}/I_{t_{|\mathscr{T}|}}) - \mathrm{e}(I_{t_{i-1}}/I_{t_{|\mathscr{T}|}})$$
$$\pi_{\{2,t_i\}} = \mathrm{f}(I_{t_i}/I_{t_{|\mathscr{T}|}}) - \mathrm{f}(I_{t_{i-1}}/I_{t_{|\mathscr{T}|}})$$

for $i \in \{2, \ldots, |\mathscr{T}| - 1\}$. Then, for convenience, we force $f_{|\mathscr{T}|} = e_{|\mathscr{T}|}$ so that a decision is made at the final analysis, and to prioritise the trial to have the desired type-I error rate, $e_{|\mathscr{T}|}$ is taken as the solution to

$$\pi_{\{1,t_{|\mathscr{T}|}\}} = \mathrm{e}(I_{t_{|\mathscr{T}|}}/I_{t_{|\mathscr{T}|}}) - \mathrm{e}(I_{t_{|\mathscr{T}|-1}}/I_{t_{|\mathscr{T}|}})$$

Note that one can prevent early stopping for futility or efficacy by setting $f_{t_1} = \ldots = f_{t_{|\mathscr{T}|-1}} = -\infty$ or $e_{t_1} = \ldots = e_{t_{|\mathscr{T}|-1}} = \infty$, and ignoring f or e, respectively, for $i \in \{1, \ldots, |\mathscr{T}| - 1\}$.

Now, all that remains is to be able to identify the C, T, and m that provide the desired power. To do this, a choice must be made as to which two of these three parameters are pre-specified. A numerical search is then performed over the third parameter. That is, we search for the minimal value of this parameter that ensures

$$\mathbb{P}(\text{Reject } H_0|\delta) \geq 1 - \beta$$

For individually randomised trials, this search is usually done assuming the relevant parameter is continuous, with it then rounded up to the nearest allowable integer to ensure the desired power is met. The ability to do this here depends upon having an explicit closed-form expression for I_t for any C, T, and m. Such a formulae is available for a range of SW-CRT designs and analysis models.[2,8,13] For some scenarios, however, to determine the required values of these parameters, an algorithm for discrete optimisation would need to be utilised. With this though, we have then completely described a means for researchers to determine SW-CRT designs with interim analyses and desired operating characteristics. In addition, our formula for $\mathbb{P}(\text{Reject } H_0|\tau)$ and $\mathbb{E}(M|\tau)$ allow the performance of these sequential designs to be compared to both each other, and to the corresponding classical fixed design, across all possible values of τ.

Hussey and Hughes model

In this section, and for the majority of the remainder of the article, we focus on cross-sectional SW-CRTs since the majority of research into the design has been set in this domain. This means that our value of m can now be interpreted as the sample size required per cluster per time period and M by the total required number of patients.

In addition, for all considered examples, we utilise the following model which has been proposed for the analysis of cross-sectional SW-CRTs[2]

$$y_{ijk} = \mu + \pi_j + \tau X_{ij} + c_i + \epsilon_{ijk}$$

where

- y_{ijk} is the response of the kth individual ($k = 1, \ldots, m$) in the ith cluster ($i = 1, \ldots, C$) and jth time period ($j = 1, \ldots, T$);
- μ is an intercept term;
- π_j is the fixed effect for the jth period (with $\pi_1 = 0$ for identifiability purposes);
- τ is the fixed treatment effect on the experimental intervention relative to the control;
- X_{ij} is the binary treatment indicator for the ith cluster and jth time period. That is, $X_{ij} = 1$ if cluster i receives the intervention in time period j. We will denote by X the matrix formed from the X_{ij};
- c_i is the random effect for cluster i, with $c_i \sim N(0, \sigma_c^2)$;
- ϵ_{ijk} is the individual-level error such that $\epsilon_{ijk} \sim N(0, \sigma_e^2)$.

Note that specification of the matrices G and R from earlier therefore requires only values for σ_c^2 and σ_e^2 to be provided. In addition, one could extend the above model to allow a cohort design, or cluster by period interactions.[8,13]

Our reasons for focusing on this model are twofold. First, as a commonly studied and utilised model, it is a sensible choice to consider when determining and exploring the performance of example sequential SW-CRT designs. Additionally, in this case, if C and T are both pre-specified, the search over m can be done assuming it to be continuous since[2]

$$I_t = \frac{(\sigma^2 + t\sigma_c^2)(CU - W) + \sigma_c^2(U^2 - CV)}{C\sigma^2(\sigma^2 + t\sigma_c^2)}$$

where $\sigma^2 = \sigma_e^2/m$ and

$$U = \sum_{i=1}^{C}\sum_{j=1}^{t} X_{ij}$$

$$V = \sum_{i=1}^{C}\left(\sum_{j=1}^{t} X_{ij}\right)^2$$

$$W = \sum_{j=1}^{t}\left(\sum_{i=1}^{C} X_{ij}\right)^2$$

Software for determining designs in this scenario is available from https://sites.google.com/site/jmswason/supplementary-material.

To summarise the above, a design in this scenario can now be determined given values for C, T, σ_c^2, σ_e^2, α, β, δ, $\mathscr{T}$, a choice for whether to allow early stopping for futility, efficacy, or futility and efficacy, and then the specification of γ_e and/or γ_f as appropriate.

Unknown variance

In the above, we required all variance parameters to be fully specified. In practice, the key variance parameters of any analysis model will not be known precisely. Instead pre-trial estimates are provided, which we denote for the Hussey and Hughes model by $\tilde{\sigma}_c^2$ and $\tilde{\sigma}_e^2$. If there is little confidence in these assumed values, a sample size re-estimation design would be more appropriate to provide the desired power, and the use of interim analyses as presented here would be unwise.

Even in the case where there is strong confidence in their values, it would often still be preferred to utilise the values for the variance parameters estimated from a trial's accrued data in the formation of the test statistic at each interim analysis, rather than the specified pre-trial estimates $\tilde{\sigma}_c^2$ and $\tilde{\sigma}_e^2$. Specifically, we wish to take

$$Z_t = \hat{\tau}_t \hat{I}_t^{1/2}$$

where $\hat{I}_t$ is the observed information at analysis t. Use of these test statistics can lead to inflation of the type-I error rate above the nominal level if no adjustment to the stopping boundaries identified under assumed known variance is made. For this adjustment, a quantile substitution procedure was previously proposed.[14] To relax the requirement for the variance parameters to be specified, we here consider the performance of this methodology at controlling the type-I error rate to the desired α in our sequential SW-CRT designs. Explicitly, a SW-CRT design with interim analyses is determined, and then the boundaries f_t and e_t are altered to f_{t*} and e_{t*}, which are the solutions of the following equations

$$\int_{f_t}^{\infty} \phi\{x, 0, 1\}\mathrm{d}x = \int_{f_{t*}}^{\infty} \varphi\{x, \nu_t\}\mathrm{d}x$$

$$\int_{e_t}^{\infty} \phi\{x, 0, 1\}\mathrm{d}x = \int_{e_{t*}}^{\infty} \varphi\{x, \nu_t\}\mathrm{d}x$$

for $t \in \mathscr{T}$. Here, $\varphi\{x, \nu\}$ is the probability density function of a central t-distribution with variance 1, and degrees of freedom ν, evaluated at x. For the explored examples here, utilising the Hussey and Hughes model, we take ν_t to be the classical decomposition of degrees of freedom in balanced, multilevel analysis of variance (ANOVA) designs[15]

$$\nu_t = mCt - C - t$$

In this instance, it is also necessary to decide whether to utilise ML, or restricted error maximum likelihood (REML) estimation, when fitting the chosen linear mixed model at each interim analysis. Here, we consider the performance of both options.

Thus, in total, the performance of each of four possible analysis procedures was explored: ML or REML estimation, with or without boundary adjustment through quantile substitution. To estimate empirical rejection rates, 100,000 trials were simulated for each considered parameter set.

Note that for simplicity, when generating data π_j was set to 0 for $j = 1, \ldots, T$, and μ_0 was set to 0. Since the analysis is asymptotically invariant under additive period effects, incorporating non-zero period effects would not be expected to greatly affect the results.

Example SW-CRT design scenarios

A SW-CRT on the effect of training doctors in communication skills on women's satisfaction with doctor–woman relationship during labour and delivery was recently conducted.[16] The trial included four hospitals ($C = 4$), with balanced stepping across five time periods ($T = 5$). The final analysis estimated $\hat{\tau} = -0.13$ with a 95% confidence interval of $(-0.29, 0.04)$ and estimated the between cluster and residual variances to be $\sigma_c^2 = 0.02$ and $\sigma_e^2 = 0.51$ respectively. Taking these variance parameters as true, a conventional SW-CRT design would have required $m = 70$ patients per cluster per time period for the trial's desired type-I and type-II error rates of $\alpha = 0.05$ and $\beta = 0.1$, respectively, powering for a clinically relevant difference of $\delta = 0.2$. Thus, for Scenario 1, we take $C = 4$, $T = 5$, $\alpha = 0.05$, $\beta = 0.1$, $\delta = 0.2$, $\sigma_c^2 = 0.02$, and $\sigma_e^2 = 0.51$.

We motivate our second example design scenario (Scenario 2) based on the average design characteristics of completed SW-CRTs according to a recently completed review.[17] Explicitly, we set the number of

Table 1. The performance of several sequential SW-CRT designs (Designs 1–6), along with that of the corresponding classical SW-CRT design (Design 7), is summarised, for Scenarios 1 and 2.

Scenario 1											
Design	$\mathscr{T}$	Stopping	γ_e	γ_f	m	$\mathbb{E}(M\|0)$	$\mathbb{P}$(Reject $H_0\|0$)	$\mathbb{E}(M\|\delta)$	$\mathbb{P}$(Reject $H_0\|\delta$)	min M	max M
Design 1	{2,3,4,5}	E&F	0.5	0.5	104	1043.49	0.05	1113.17	0.90	832	2080
Design 2	{3,5}	F	NA	1	75	1031.73	0.05	1464.44	0.90	900	1500
Design 3	{3,4,5}	E	1	NA	97	1912.03	0.05	1288.63	0.93	1164	1940
Design 4	{2,3,4,5}	E&F	1.5	1	84	946.52	0.05	1040.49	0.90	672	1680
Design 5	{3,5}	F	NA	1.5	73	1032.61	0.05	1433.30	0.90	876	1460
Design 6	{3,4,5}	E	0.5	NA	104	2044.88	0.05	1353.52	0.95	1248	2080
Design 7	{5}	NA	NA	NA	70	1400.00	0.05	1400.00	0.90	1400	1400
Scenario 2											
Design 1	{2,4,7,9}	E&F	0.5	0.5	11	878.21	0.05	1063.59	0.81	440	1980
Design 2	{5,9}	F	NA	1	9	856.89	0.05	1037.91	0.82	900	1620
Design 3	{3,6,9}	E	1	NA	8	1416.43	0.05	1031.39	0.81	480	1440
Design 4	{2,4,7,9}	E&F	1.5	1	9	1583.44	0.05	1084.97	0.82	360	1620
Design 5	{5,9}	F	NA	1.5	8	924.10	0.05	1365.12	0.82	800	1440
Design 6	{3,6,9}	E	0.5	NA	8	952.90	0.05	1382.72	0.83	480	1440
Design 7	{9}	NA	NA	NA	7	1260.00	0.05	1260.00	0.81	1260	1260

E&F: efficacy and futility; E: efficacy; F: futility; NA: not applicable.
All rounding is to two decimal places.

clusters to be 20 ($C = 20$), and the number of time periods to be nine ($T = 9$), to correspond to the median values used in-practice. We suppose $\alpha = 0.05$, $\beta = 0.2$, and choose $\sigma_e^2 = 1$, $\sigma_c^2 = 1/9$ to imply a more moderate value for the intra-cluster correlation of 0.1 compared to Scenario 1. Prescribing near-balanced stepping, we specify that three clusters switch to the intervention in the second through fifth, and two clusters in each of the remaining, time periods. Finally, to ensure a total sample size approximately equal to the median value of completed SW-CRTs, we choose $\delta = 0.24$. Specifically, this implies $m = 7$ patients are needed per cluster per time period to meet the above operating characteristics.

For both scenarios, we then consider the effect of different choices for the remaining design parameters: $\mathscr{T}$, γ_e, and/or γ_f.

Results

Example sequential SW-CRT designs

The performance of several example sequential SW-CRT designs with differing choices for $\mathscr{T}$, and the allowed reasons for early stopping, is summarised in Table 1 for Scenarios 1 and 2. It is clear that the incorporation of early stopping can substantially reduce the expected sample size under H_0 (up to 32% in Scenario 1 using Design 4, and 32% in Scenario 2 using Design 2) and H_1 (up to 26% in Scenario 1 again using Design 2, and 18% in Scenario 2 using Design 3), with no cost to the type-I or type-II error rates.

However, as would be expected, the maximal sample size that could be required by the sequential designs is larger than that of the corresponding fixed sample design. Furthermore, the sample size required by the sequential designs can be subject to substantial variability. In Figure 1, this variability is displayed for the sequential designs with early stopping for efficacy and futility ($\gamma_e = \gamma_f = 0.5$), in Scenario 1 ($\mathscr{T} = \{2, 3, 4, 5\}$) and Scenario 2 ($\mathscr{T} = \{2, 4, 7, 9\}$), when $\tau = 0, \delta$. We observe that while the expected sample sizes may always be lower than that of the corresponding fixed sample design, there is always a non-negligible probability that a larger sample size could be expected (up to 38% when $\tau = \delta$ for the Scenario 2 design).

Considerations on $\mathscr{T}$, γ_e, γ_f, and the allowed reasons for early stopping

In Figure 2, we demonstrate the effect of different choices for $\mathscr{T}$ in Scenarios 1 and 2 with all other parameters fixed (early stopping for efficacy and futility with $\gamma_e = \gamma_f = 0.5$). It can be seen that, as is the case for individually randomised group sequential trials, increasing the number of interim analyses typically reduces the expected sample size of our sequential SW-CRT designs. However, this usually comes at a cost of an increased maximal sample size (Table 2). Moreover, for a fixed number of interim analyses, placing them after earlier time periods typically leads to smaller minimal sample sizes, but larger expected sample sizes when $\tau = 0$ or $\tau = \delta$. These patterns are, however, only a rough trend. The requirement for m to be an integer means that they may not always be present.

Table 2. The performance of several sequential SW-CRT designs with different possible choices for $\mathcal{T}$ in Scenarios 1 and 2, along with that of the corresponding classical SW-CRT design ($\mathcal{T} = \{5\}$ in Scenario 1 and $\mathcal{T} = \{9\}$ in Scenario 2), is summarised.

$\mathcal{T}$	m	$\mathbb{E}(M\|0)$	$\mathbb{E}(M\|\delta)$	min M	max M
Scenario 1					
{2,3,4,5}	104	1043.49	1113.17	832	2080
{2,3,5}	100	1051.78	1139.21	800	2000
{3,4,5}	93	1153.99	1175.46	1116	1860
{2,5}	90	1188.84	1296.69	720	1800
{3,5}	90	1148.57	1184.27	1080	1800
{4,5}	79	1268.06	1270.79	1264	1580
{5}	70	1400.00	1400.00	1400	1400
Scenario 2					
{2,4,7,9}	11	878.21	1063.58	440	1980
{2,3,6,9}	11	891.44	1091.02	440	1980
{3,6,9}	10	859.24	1017.45	600	1800
{2,4,9}	10	902.58	1131.62	400	1800
{5,9}	9	965.12	1042.02	900	1620
{3,9}	9	979.53	1180.94	540	1620
{9}	7	1260.0	1260.00	1260	1260

Early stopping is allowed for efficacy and futility, with $\gamma_e = \gamma_f = 0.5$. All rounding is to two decimal places. All designs have a type-I error rate of 0.05, a type-II error rate of 0.1 in Scenario 1, and a type-II error rate of 0.2 in Scenario 2, as desired.

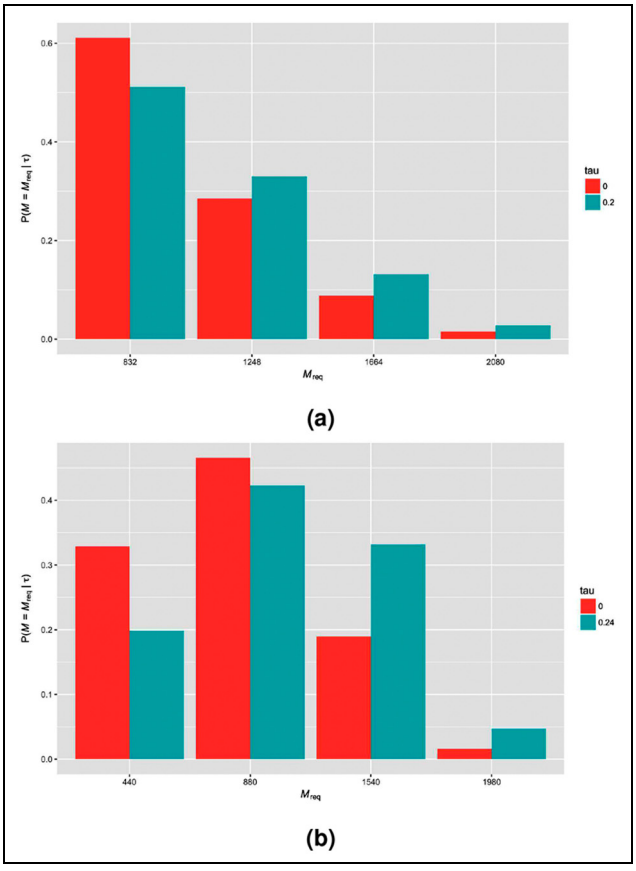

Figure 1. The probability distribution of the sample size required by example sequential designs (early stopping for efficacy and futility with $\gamma_e = \gamma_f = 0.5$) for (a) Scenario 1 ($\mathcal{T} = \{2, 3, 4, 5\}$) and (b) Scenario 2 ($\mathcal{T} = \{2, 4, 7, 9\}$) is shown when $\tau = 0, \delta$ ($\delta = 0.2$ for Scenario 1, $\delta = 0.24$ for Scenario 2).

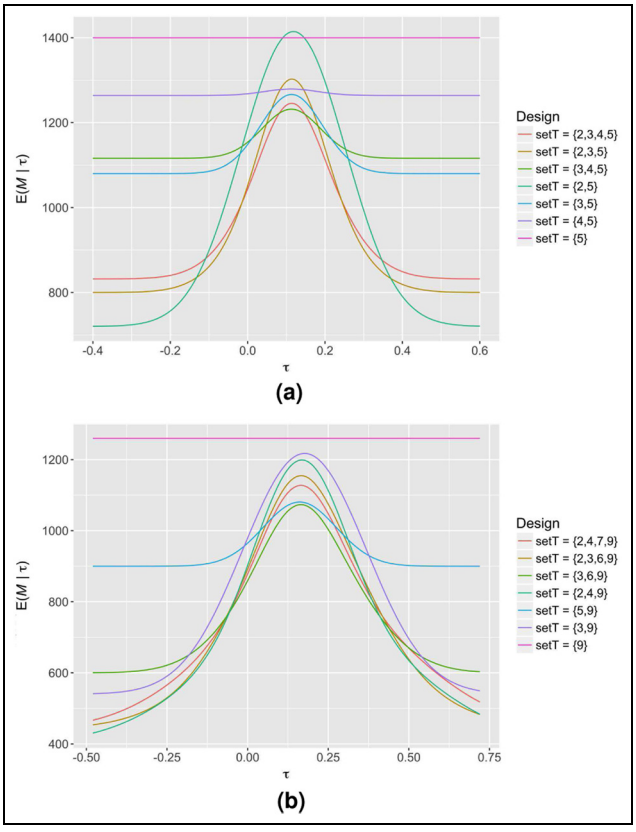

Figure 2. The expected sample size curves of several sequential SW-CRT designs with different possible choices for $\mathcal{T}$ in (a) Scenario 1 and (b) Scenario 2, along with that of the corresponding classical SW-CRT design ($\mathcal{T} = \{5\}$ in Scenario 1 and $\mathcal{T} = \{9\}$ in Scenario 2), are displayed. Early stopping is allowed for efficacy and futility, with $\gamma_e = \gamma_f = 0.5$.

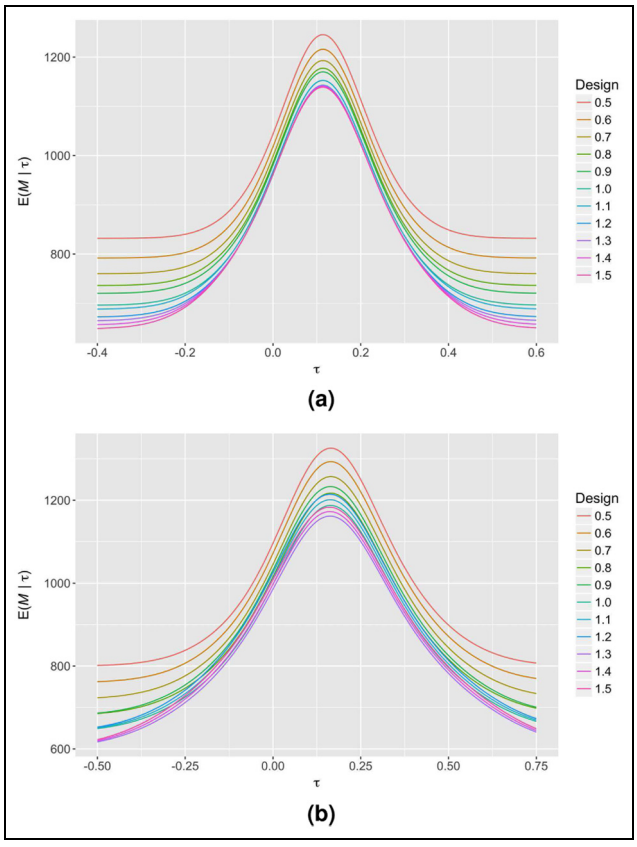

Figure 3. The expected sample size curves of several sequential SW-CRT designs with different possible choices for $\gamma_e = \gamma_f$ are displayed for (a) Scenario 1 and (b) Scenario 2. Early stopping is allowed for efficacy and futility, with $\mathscr{T} = \{,,,\}$ in Scenario 1 and $\mathscr{T} = \{,,,\}$ in Scenario 2.

Similarly, Figure 3 displays the impact of altering $\gamma_e = \gamma_f$ in Scenarios 1 and 2 when there is early stopping for efficacy and futility, and all other parameters are fixed ($\mathscr{T} = \{2,3,4,5\}$ in Scenario 1, and $\mathscr{T} = \{2,4,7,9\}$ in Scenario 2). Typically, the fact that the majority of information in a SW-CRT is accrued towards the completion of the trial means that increasing the value of $\gamma_e = \gamma_f$ results in designs with lower expected sample sizes, since it is preferable to spend the type-I and type-II error later in the trial. Once more, however, this is not guaranteed to be the observed pattern, as illustrated by Scenario 2.

Finally, in Figure 4, we observe the effect of the choice of allowed reasons for early stopping (Scenario 1 with $\mathscr{T} = \{2,3,4,5\}$ and $\gamma_e = \gamma_f = 0.5$). Incorporating early stopping for both efficacy and futility provides good performance across all possible values for τ. However, this design carries the largest possible maximal sample size (2080, relative to 1760 for efficacy stopping only, and 1740 for futility stopping only). Allowing early stopping for only futility (efficacy) results in the largest possible reduction to the expected sample size under H_0 (H_1), but comes at the biggest cost to the expected sample size under H_1 (H_0).

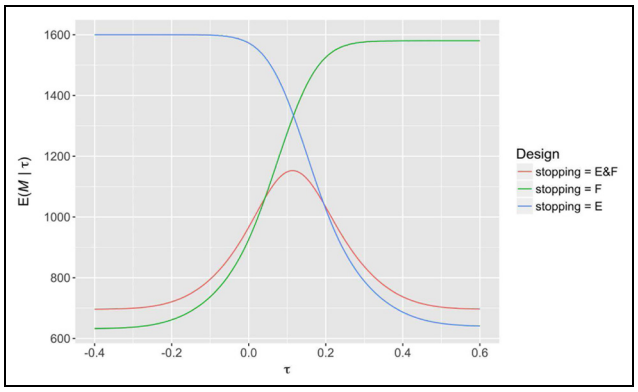

Figure 4. The expected sample size curves of several sequential SW-CRT designs with different possible allowed reasons for early stopping are displayed for Scenario 1. Each design has $\mathscr{T} = \{,,,\}$, with $\gamma_e = \gamma_f = 0.5$ when required.

Quantile substitution

In Table 3, the empirical rejection rate of the sequential SW-CRT designs for Scenario 1 with $\mathscr{T} = \{2,3,4,5\}$ and $\mathscr{T} = \{3,5\}$, taking $\gamma_e = \gamma_f = 0.5$, are explored for each of our four considered analysis procedures, for $\tau = 0$ and $\tau = \delta$, and finally for three possible combinations of $\tilde{\sigma}_c^2$ and $\tilde{\sigma}_e^2$. Namely, these are, to reflect a situation where estimates provided are close to their true values, $(\tilde{\sigma}_c^2, \tilde{\sigma}_e^2) = 0.9(\sigma_c^2, \sigma_e^2)$, $(\tilde{\sigma}_c^2, \tilde{\sigma}_e^2) = (\sigma_c^2, \sigma_e^2)$, and $(\tilde{\sigma}_c^2, \tilde{\sigma}_e^2) = 1.1(\sigma_c^2, \sigma_e^2)$. For comparison, the empirical rejection rates of the corresponding fixed sample SW-CRT designs are also shown.

It is clear that when ML estimation is utilised in the sequential designs there can be substantial inflation in the empirical type-I error rate (up to 0.0812 for $(\tilde{\sigma}_c^2, \tilde{\sigma}_e^2) = 0.9(\sigma_c^2, \sigma_e^2)$ without quantile substitution when $\mathscr{T} = \{2,3,4,5\}$). However, when REML estimation is used, there is generally much better control (with a maximum of only 0.0645 for $(\tilde{\sigma}_c^2, \tilde{\sigma}_e^2) = 0.9(\sigma_c^2, \sigma_e^2)$ with quantile substitution when $\mathscr{T} = \{2,3,4,5\}$). In general, it appears the sample size is large enough that quantile substitution makes little difference to the empirical type-I error rate. However, the small number of clusters makes REML estimation particularly important.

The small number of clusters also results in inflation of the type-I error rate in the fixed sample designs. It is clear that this inflation is typically less than that for the corresponding sequential design and analysis procedure, though the difference is smaller when REML estimation and quantile substitution is utilised. Moreover, the inflation is smaller for the sequential designs with $\mathscr{T} = \{4,5\}$ compared to those with $\mathscr{T} = \{2,3,4,5\}$, because the later timing of the analyses helps alleviate the issues caused by the small value of C.

Discussion

In this article, we demonstrated how established group sequential trial methodology can be adapted to

Table 3. The empirical rejection rate using the four considered analysis procedures (ML or REML estimation, with or without boundary adjustment (BA) through quantile substitution) is displayed, for several possible values of the assumed variance parameters, true treatment effect, and the designs with $\mathscr{T} = \{2, 3, 4, 5\}$, $\mathscr{T} = \{4, 5\}$, and $\mathscr{T} = \{5\}$.

$(\tilde{\sigma}_c^2, \tilde{\sigma}_e^2)$	τ	Estimation	BA	Empirical rejection rate		
				$\mathscr{T} = \{2, 3, 4, 5\}$	$\mathscr{T} = \{4, 5\}$	$\mathscr{T} = \{5\}$
$0.9(\sigma_c^2, \sigma_e^2)$	0	ML	No	0.0812	0.0712	0.0625
	0	ML	Yes	0.0808	0.0699	0.0605
	0	REML	No	0.0640	0.0595	0.0541
	0	REML	Yes	0.0645	0.0595	0.0550
	δ	ML	No	0.8748	0.8800	0.8809
	δ	ML	Yes	0.8760	0.8843	0.8845
	δ	REML	No	0.8818	0.8840	0.8825
	δ	REML	Yes	0.8789	0.8844	0.8839
(σ_c^2, σ_e^2)	0	ML	No	0.0777	0.0674	0.0600
	0	ML	Yes	0.0781	0.0675	0.0596
	0	REML	No	0.0627	0.0560	0.0536
	0	REML	Yes	0.0624	0.0575	0.0531
	δ	ML	No	0.9014	0.9089	0.9097
	δ	ML	Yes	0.9017	0.9090	0.9102
	δ	REML	No	0.9080	0.9106	0.9076
	δ	REML	Yes	0.9075	0.9099	0.9079
$1.1(\sigma_c^2, \sigma_e^2)$	0	ML	No	0.0755	0.0666	0.0582
	0	ML	Yes	0.0769	0.0667	0.0591
	0	REML	No	0.0600	0.0564	0.0515
	0	REML	Yes	0.0608	0.0573	0.0521
	δ	ML	No	0.9219	0.9298	0.9309
	δ	ML	Yes	0.9228	0.9290	0.9289
	δ	REML	No	0.9270	0.9306	0.9310
	δ	REML	Yes	0.9273	0.9304	0.9312

ML: maximum likelihood; REML: restricted error maximum likelihood.
All rounding is to four decimal places.

determine SW-CRT with interim analyses. It was clear from our examples that the incorporation of interim analyses into the SW-CRT design could bring substantial reductions in the expected sample size under both H_0 and H_1. Researchers should therefore certainly consider incorporating interim analyses into any future SW-CRT they conduct. However, it is important to note several practical considerations about the employment of interim analyses.

Although the inherent time period structure of SW-CRTs lends itself well to sequential methods, this does rely upon the efficient collection and storage of data for analysis. Putting measures in place to prevent operational issues would therefore be essential. In reality, a small delay between time periods may be necessary to allow for an interim analysis to be conducted. Without this, the clusters will have already begun data accrual for the following time period, which would bring a loss of efficiency to the required number of measurements. In addition, the more interim analyses a trialist includes theoretically reduces the expected sample size; however, this too comes with a larger burden in terms of the cost of analysis. In practice, trading off some loss in efficiency to reduce this burden may be wise.

Furthermore, the increase in sample size required per cluster per period in the sequential designs may mean the length of each time period needs to be increased. This would specifically be true when the length of a time period is chosen based on the supposed achievable recruitment rate. In this instance, however, the possibility to stop the trial early in the sequential designs means the average length of a trial could often still be reduced.

Moreover, the methodology presented here requires data to be unblinded at each interim analysis. Although many SW-CRTs are performed in an unblinded manner,[18] it would be important to ensure even then that the results of the data analysis at interim are kept hidden from all but those on the Data Monitoring Committee.

There is also much to consider in terms of the choice of allowed reasons for early stopping. It was previously noted that stopping early for futility would be unlikely in a SW-CRT because of the often held a priori belief the intervention will be effective.[10] However, a recent literature review established that 31% of SW-CRTs completed to date did not find a significant effect of their intervention on any primary outcome measure.[17] For this reason, incorporation of futility stopping does in fact seem warranted. Nonetheless, there are additional factors to consider. Primarily, the plan to eventually implement the intervention in all clusters, as is often the case in SW-CRTs, could be decided upon as

an incentive for cluster participation in the trial. If this is the case, one must be careful to acknowledge to enrolled clusters that they may in fact not receive the intervention if the trial is stopped early for futility. Furthermore, some SW-CRTs are planned roll-outs of a programme, in which case there may not be a desire to stop the roll-out for futility if the study is part of a larger programme implementation. If this is the case, it may be likely that a SW-CRT design with early stopping would not be appropriate.

Moreover, the stopping of a trial for efficacy would typically imply the immediate deployment of an intervention to all clusters will then follow. However, with SW-CRTs often used when there are logistic constraints, this may not be possible. It could be that an intervention is rolled out as quickly as is possible, but this fact should be considered before early stopping for efficacy is included in a design. Finally, in some instances, there may be a desire to study the development of an intervention within the clusters over time. Stopping a trial early for efficacy or futility may prevent this possibility. In this case, it could be wise to only include stopping for futility to guard solely against harmful interventions.

There are several methodological considerations that should be recognised. First, the approach used to sequential SW-CRT design here assumes the trial's nuisance parameters to be known. We demonstrated that REML estimation can help deal with this problem in the case where there is only small uncertainty in their values, and the number of clusters is small. As was noted, a sample size re-estimation procedure would be required if this was not the case. Moreover, even in this instance, there was still some inflation to the empirical type-I error rate. This is common, however, to both the classical fixed sample design and our proposed sequential designs. Nonetheless, smaller inflation was observed in a sequential design with fewer interim analyses, placed later into a trial. Therefore, similar to the burden introduced from introducing additional analyses discussed above, this should be factored in when choosing an appropriate sequential design.

Additionally, as with any trial design scenario, if the model assumed at the design stage does not hold, the trial's operating characteristics will not be reliable. For a sequential design, depending on the violation, the degree to which the type-I and type-II error rates depart from their planned values could be larger than that of a fixed sample SW-CRT design. It would be important therefore when choosing an appropriate sequential SW-CRT, as for classical SW-CRT designs, to assess the sensitivity of the design to deviations in the underlying distribution of the data.

Finally, we have here only considered the design of SW-CRTs with interim analyses. It is well known that if naive estimators are used after a sequential trial, then acquired treatment effects will be biased. The development of methodology to account for this would be required. Fortunately, there is a breadth of literature on this for individually randomised trials upon which such work could be based (see, for example, Bretz et al.[19]).

In conclusion, although there are several factors that must be considered by a trialist before deciding to incorporate early stopping in to a SW-CRT design, they should certainly consider whether the methodology is appropriate. With the inclusion of interim analyses, they can more suitably guard against much investment being spent on an inferior intervention or indeed help hasten the roll-out of an efficacious treatment.

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

This work was supported by the Wellcome Trust (grant number 099770/Z/12/Z to M.J.G.), the National Institute for Health Research Cambridge Biomedical Research Centre (grant number MC_UP_1302/4 to J.M.S.W.), and the Medical Research Council (grant number MC_UP_1302/6 to A.P.M.).

References

1. Cook TD and Campbell DT. *Quasi-experimentation: design & analysis issues for field settings*. Chicago, IL: Rand McNally, 1979.
2. Hussey MA and Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007; 28(2): 182–191.
3. 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.
4. Hemming K, Haines TP, Chilton PJ, et al. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* 2015; 350: h391.
5. Baio G, Copas A, Ambler G, et al. Sample size calculation for a stepped wedge trial. *Trials* 2015; 16(1): 354.
6. Hemming K and Girling A. A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. *Stata J* 2014; 14(2): 363–380.
7. Lawrie J, Carlin JB and Forbes AB. Optimal stepped wedge designs. *Stat Probabil Lett* 2015; 99: 210–214.
8. 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.
9. 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.
10. 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(1): 93.

11. Jennison C and Turnbull BW. *Group sequential methods with applications to clinical trials*. Boca Raton, FL: Chapman & Hall/CRC, 2000.
12. Lan KKG and DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; 70(3): 659–663.
13. 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.
14. Whitehead J, Valdes-Marquez E and Lissmats A. A simple two-stage design for quantitative responses with application to a study in diabetic neuropathic pain. *Pharm Stat* 2009; 8: 125–135.
15. Pinheiro JC and Bates D. *Mixed-effects models in S and SPLUS* (statistics and computing). New York: Springer, 2009.
16. Bashour HN, Kanaan M, Kharouf MH, et al. The effect of training doctors in communication skills on women's satisfaction with doctor-woman relationship during labour and delivery: a stepped wedge cluster randomised trial in Damascus. *BMJ Open* 2013; 3(8): e002674.
17. 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.
18. Mdege ND, Man MS, Taylor Nee, Brown 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.
19. Bretz F, Koenig F, Brannath W, et al. Adaptive designs for confirmatory clinical trials. *Stat Med* 2009; 28(8): 1181–1217.