Article

CLINICAL TRIALS

Clinical Trials
2019, Vol. 16(3) 225–236

DOI: 10.1177/1740774519829053
journals.sagepub.com/home/ctj


Sample size calculation for stepped-wedge cluster-randomized trials with more than two levels of clustering

Steven Teerenstra[1], Monica Taljaard[2,3], Anja Haenen[4,5], Anita Huis[5], Femke Atsma[5], Laura Rodwell[1] and Marlies Hulscher[5]

Abstract

Background/Aims: Power and sample size calculation formulas for stepped-wedge trials with two levels (subjects within clusters) are available. However, stepped-wedge trials with more than two levels are possible. An example is the CHANGE trial which randomizes nursing homes (level 4) consisting of nursing home wards (level 3) in which nurses (level 2) are observed with respect to their hand hygiene compliance during hand hygiene opportunities (level 1) in the care of patients. We provide power and sample size methods for such trials and illustrate these in the setting of the CHANGE trial.

Methods: We extend the original sample size methodology derived for stepped-wedge trials based on a random intercepts model, to accommodate more than two levels of clustering. We derive expressions that can be used to determine power and sample size for p levels of clustering in terms of the variances at each level or, alternatively, in terms of intracluster correlation coefficients. We consider different scenarios, depending on whether the same units in a particular level are repeatedly measured as a cohort sample or whether different units are measured cross-sectionally.

Results: A simple variance inflation factor is obtained that can be used to calculate power and sample size for continuous and by approximation for binary and rate outcomes. It is the product of (1) variance inflation due to the multilevel structure and (2) variance inflation due to the stepped-wedge manner of assigning interventions over time. Standard and non-standard designs (i.e. so-called "hybrid designs" and designs with more, less, or no data collection when the clusters are all in the control or are all in the intervention condition) are covered.

Conclusions: The formulas derived enable power and sample size calculations for multilevel stepped-wedge trials. For the two-, three-, and four-level case of the standard stepped wedge, we provide programs to facilitate these calculations.

Keywords

Stepped-wedge trials, hybrid (stepped wedge) design, power, sample size, multilevel, variance inflation factor

Introduction

Hussey and Hughes[1] and Girling and Hemming[2] derived a power formula for the standard stepped-wedge cluster-randomized design (see Figure 1) with two levels of clustering (i.e. subjects within clusters), where cross-sectional samples are taken at the lowest (subject) level, that is, different subjects are measured in every period. In this article, we derive and demonstrate power and sample size calculations for stepped-wedge cluster trials with more than two levels, in which the lowest level is cross-sectional. One such example is the CHANGE trial (ClinicalTrial.gov NCT02817282), which aims to improve nurses' level of compliance with hand hygiene guidelines. This trial has four levels of clustering, with nurses (level 2) in wards (level 3) of

[1]Section Biostatistics, Department for Health Evidence, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands
[2]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada
[3]School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada
[4]Centre for Infectious Diseases, Epidemiology and Surveillance, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands
[5]Scientific Center for Quality of Healthcare, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands

Corresponding author:
Steven Teerenstra, Section Biostatistics, Department for Health Evidence, Radboud Institute for Health Sciences, Radboud University Medical Center, Internal Postal Code 133, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands.
Email: Steven.Teerenstra@Radboudumc.nl

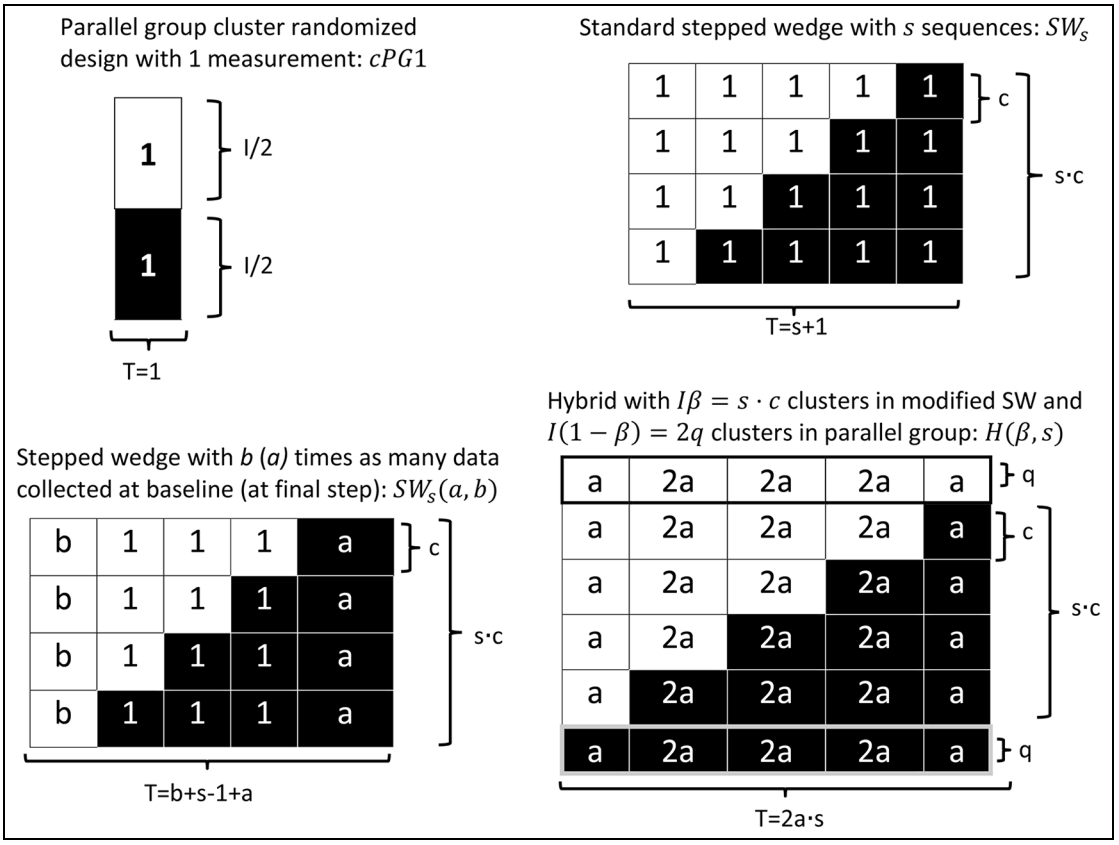

Figure 1. Cluster-randomized parallel group design and different stepped wedge like designs with $s = 4$ sequences. Each row corresponds to a sequence in the design with the number of clusters in that sequence at the right side of the row. The background color of a cell indicates the treatment (white for control and black for intervention) and the number within a cell gives the number of repeated measurements. The total number of measurements is indicated below the design. Further details are provided in the Supplementary Files (SF3, 4, and5).

several nursing homes (level 4). Nurses are followed in sessions where different opportunities for hand hygiene arise and observations (level 1) on compliance to the guideline are made.

If clusters consist of more than two levels, different scenarios are possible. For example, in the CHANGE trial that has four levels, the following scenarios are possible (Figure 2):

Level 4 repeated: the same nursing homes are repeatedly measured (i.e. as a cohort) but in every measurement period, different wards are measured, implying also that different nurses and hygiene observations are made (i.e. cross-sectional measurement at the lower levels).

Levels 4 and 3 repeated: the same wards within nursing homes are repeatedly measured, but in every measurement period, different nurses and hygiene observations are made cross-sectionally over time.

Levels 4 and 3 and 2 repeated: the same nurses within wards within nursing homes are repeatedly measured (cohort design at these levels); in each measurement period, different (i.e. cross-sectional) hygiene observations are made.

As illustrated in the range of possible scenarios above, the highest level (referred to as a cluster in this article) is always repeatedly measured and the lowest level cross-sectionally. Up to a certain level, all levels below this level are cross-sectionally measured, but levels above it as cohort.

Our method covers both "standard" stepped-wedge designs (i.e. designs where all clusters start in the control and end in the intervention condition) and non-standard designs (i.e. stepped-wedge designs with more, less, or no data collection before and/or after roll-out[3] and hybrid designs;[2] see Figure 1 with $s = 4$ stepped-wedge sequences).

Methods

In order to support the flow of arguments, technical derivations are provided in the Supplementary Files (SF) and notations given in Table 1. At time t, cluster i is either in the control condition ($X_{it} = 0$) or in the intervention condition ($X_{it} = 1$). For power calculations, we make the simplifying assumption that the differences between conditions, δ, is the same wherever and whenever the intervention is introduced and is maintained at this level. Hussey and Hughes[1] modeled the clustering of subjects within clusters by a random

intercept for cluster (level 2 random effect). For more than two levels, we extend this idea by incorporating random effects for each clustering level. For example, for four levels, the outcome Y_{itjkm} of "observation" (level 1 unit) $m = 1, \ldots, n_1$ of "subject" (level 2 unit) $k = 1, \ldots, n_2$ within "sub-cluster" (level 3 unit) $j = 1, \ldots, n_3$ within "cluster" (level 4 unit) $i = 1, \ldots, I$ in measurement/period $t = 1, \ldots, T$ is

$$\left.\begin{array}{c} Y_{itjkm} = \mu + u_{000i} + u_{00i(t)j} + u_{0i(t)jk} + \beta_t + \delta X_{it} + e_{itjkm}, \\ u_{000i}, u_{00i(t)j}, u_{0i(t)jk} \text{ random effects at levels } 4, 3, \text{ and } 2 \text{ with variances } \sigma_4^2, \sigma_3^2, \text{ and } \sigma_2^2, \text{respectively} \\ e_{itjkm} \text{ random effect (residual) at level } 1, \text{with variance } \sigma_1^2 \\ \{u_{000i}, u_{00(it)j}, u_{0(it)jk}, e_{itjkm}\} \text{mutually independent}; \\ u_{00i(t)j}, u_{00i(t')j} \text{ are equal (unequal) for } t \neq t' \text{ if level } 3 \text{ measured as cohort (cross}-\text{sectional}); \\ u_{0i(t)jk}, u_{0i(t')jk} \text{ are equal (unequal) for } t \neq t' \text{ if level } 2 \text{ measured as cohort (cross}-\text{sectional}) \end{array}\right\} \quad (1)$$

If an intermediate level is measured as cohort, the index (t) can be dropped. In this article, we assume that at every measurement time/period ($t = 1, 2, \ldots, T$)

1. All clusters ($i = 1, 2, \ldots, I$) are measured;
2. Each level-2 unit (e.g. nurse) has the same number n_1 of level-1 units (e.g. observations); each level-3 unit (e.g. nursing home) has the same number n_2 of level-2 units (e.g. nurses), and so on.
3. Randomization is always on the highest level.

In terms of the cluster averages $Y_{it\bullet}$ at each time point/period (so $Y_{it\bullet} = (\sum_{j,k,m} Y_{itjkm})/(n_1 n_2 n_3)$ for four levels), we have a repeated measurement design, and the above model implies equal covariance $\tau^2 = Cov(Y_{it\bullet}, Y_{it'\bullet})$ between averages of the same cluster over time, and equal variance $Var(Y_{it\bullet}) = \sigma^2 + \tau^2$ of the clusters across all time/period (SF1). The variance of the weighted least-squares estimator $\hat{\delta}$ for the intervention effect is (Hussey & Hughes, 2007)

$$\left.\begin{array}{c} var\left(\hat{\delta}\right) = \frac{I\sigma^2\left(\sigma^2 + T\tau^2\right)}{f(X)\sigma^2 + g(X)\tau^2} \\ f(X) = S \cdot I - C,\ g(X) = S^2 + S \cdot I \cdot T - R \cdot I - C \cdot T \\ S = \Sigma_{it} X_{it},\ C = \Sigma_t \left(\Sigma_i X_{it}\right)^2,\ R = \Sigma_i \left(\Sigma_t X_{it}\right)^2 \end{array}\right\} \quad (2)$$

where S is the *sum* of matrix elements, C is the sum of squared *column* sums, and R is the sum of squared *row* sums of $X = (X_{it})$.

In terms of the correlation $\rho = corr(Y_{it\bullet}, Y_{it'\bullet})$ between averages of the same cluster over time, we can reformulate this as (SF2)

$$var\left(\hat{\delta}\right) = \frac{I \cdot (1-\rho) \cdot [1 + (T-1)\rho]}{f(X) \cdot (1-\rho) + g(X) \cdot \rho} \cdot var(Y_{it\bullet}) \quad (3)$$

or in equivalent formulation by Girling and Hemming[2] (SF2)

$$\left.\begin{array}{c} var\left(\hat{\delta}\right) = \frac{(1-\rho)}{I \cdot T \cdot (a_D(X) - b_D(X) \cdot R)} \cdot var(Y_{it\bullet}) \\ a_D(X) = \frac{1}{I \cdot T} \cdot \Sigma_{it}\left(X_{ij} - X_{\bullet t}\right)^2, \\ b_D(X) = 1/I\, \Sigma_{it}\left(X_{i\bullet} - X_{\bullet\bullet}\right)^2, \\ R = \frac{T \cdot \rho}{1 + (T-1) \cdot \rho} \\ X_{\bullet t} = \Sigma_i X_{it}/I, X_{i\bullet} = \Sigma_t X_{it}/T,\ X_{\bullet\bullet} = \Sigma_{it} X_{it}/(I \cdot T). \end{array}\right\} \quad (4)$$

where a_D is the within-column variance of (X_{it}) and b_D is the between-row variance. Note that ρ is not an intracluster correlation coefficient, but it can be expressed in terms of intracluster correlations of the multilevel design (Table 2).

Taking f, g corresponding to a standard stepped-wedge design, we get

$$var\left(\hat{\delta}\right) = \frac{6}{I \cdot \left(s - \frac{1}{s}\right)} \cdot \sigma^2 \cdot \left[1 + \frac{\frac{s}{2} \cdot \tau^2}{\sigma^2 + \left(1 + \frac{s}{2}\right)\tau^2}\right] \quad (5a)$$

$$var\left(\hat{\delta}\right) = \frac{6 \cdot (1-\rho)}{I \cdot \left(s - \frac{1}{s}\right)} \cdot \frac{[1 + s\rho]}{\left[1 + \frac{s}{2}\rho\right]} \cdot var(Y_{it\bullet}) \quad (5b)$$

For two levels, equation (5b) reduces to the variance formula in the appendix of the article by Woertman et al.[4]

For f, g of the other designs, see SF4 and 5.

Impact of design and multilevel structure

The design (i.e. the specification of intervention/control condition for each cluster at each time) influences $var(\hat{\delta})$ via f, g or a_D, b_D, while the data generating model (1) influences $var(\hat{\delta})$ via ρ and $var(Y_{it\bullet})$ or, equivalently, σ^2 and τ^2. Specifically, the number of levels and the sample size at each level determine $var(Y_{it\bullet})$, while the specification of which levels are measured as a cohort and which levels cross-sectionally determines ρ (see Table 2).

As illustrated for the CHANGE trial in the section "Introduction," various scenarios can arise because up to a certain level, all units of lower levels are measured cross-sectionally, and from that level upward, all levels have their units measured repeatedly as cohort. Relevant formulas for each possible scenario with two, three, and four levels are provided in Table 2. Derivation and implementation of these formulas in

SAS® and Excel® programs are in the SF, which also contains the results for more than four levels.

Variance inflation due to the multilevel structure

The factor $var(Y_{it\bullet})$ in equations (3) and (4) is calculated the same as in cluster-randomized trials with a *parallel group* post-test (i.e. with one measurement) design. For two levels, $var(Y_{it\bullet}) = [1 + (n-1)ICC] \cdot 4\sigma^2_{tot}/N_{tot}$ where $ICC = \rho_{12}$ is the intracluster correlation of subjects within clusters and $[1 + (n-1)ICC]$ is the variance inflation factor (*VIF*), also known as design effect[5] (SF1.2). For more than two levels, variance inflation factors due to the multiple levels of clustering can also be used, and there are several ways to define these. One is to define separate variance inflation factors for the correlation of level 1 units in level 2 units, for the correlation of level 2 units in level 3 units and so on;[6,7] another is to define separate variance inflation factors based on the correlation of level 1 units in the same level 2 units, the correlation of level 1 units in the same level 3 units, but different level 2 units, and so on.[8,9] Both types of intracluster correlations and variance inflation factors can be expressed in terms of the other (SF1.1). Here, we use only the first mentioned type. Then, the variance inflation for p levels is

$$VIF_p = \left[1 + (n_1 - 1)_{12}\right] \cdot [1 + (n_2 - 1)\tilde{\rho}_{23}] \ldots \left[1 + \left(n_{p-1} - 1\right)\tilde{\rho}_{p-1,p}\right] \quad (6)$$

To clarify the meaning of this in the CHANGE trial setting, the intracluster correlation ρ_{12} is the true (population) correlation between any pair of observations within the same nurse; the intracluster correlation ρ_{23} is the correlation between true outcomes of two nurses within the same ward; and so on. Because we only have a sample of n_1 observations per nurse, the true outcome of the nurses can only be approximated by taking the average of the observations per nurse, and therefore, the correlation between the outcomes of two nurses within the same home is attenuated to $\tilde{\rho}_{23}$. The same holds for the other correlations. More on the estimation, interpretation and the attenuation of these intracluster correlations can be found in the article by Teerenstra et al.[6]

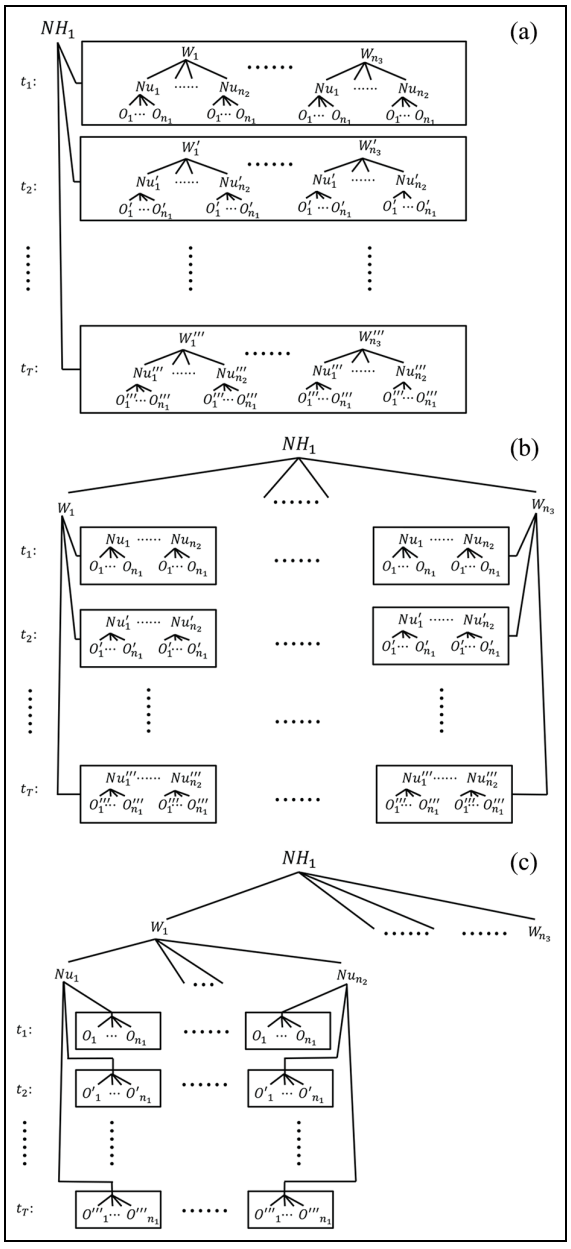

Figure 2. Scenarios in four-level stepped-wedge design (CHANGE trial setting): (a) only nursing homes *NH* (level 4) followed as cohort, (b) wards *W* (level 3) within nursing homes (level 4) followed as cohort, and (c) nurses *Nu* (level 2) within wards (level 3) in nursing homes (level 4) followed as cohort. The boxed parts of the multilevel data are measured cross-sectionally. In particular, the observations *O* (level 1) are always measured cross-sectionally.

Variance inflation factor for stepped-wedge designs

Using equation (3) or (4) and the research by Girling and Hemming[2] and Thompson et al.,[3] we provide variance inflation factors for the p-level "standard" cluster-randomized stepped-wedge design with s sequences, the stepped wedge with more/fewer/no observations before and/or after roll-out, and the hybrid design (SF8). We formulate these compared to a p-level cluster-randomized parallel group design with one measurement ($cPG1$) design

Table 1. Notations in this article illustrated in the CHANGE trial setting.

Parameter	Meaning (in the four-level CHANGE trial)
$Y_{it\bullet}$	The average of outcome Y in cluster i at time t, that is, the dot means averaging over all sub-units
δ	Treatment effect
β_t	Time effect at measurement time/period t
X_{it}	Design matrix: $X_{it} = 1$ if cluster i has intervention at time t, and $X_{it} = 0$ if it is in control condition
σ^2_{tot}	Total variance of level 1 units unconditional, that is, regardless of the cluster they belong to
ρ_{12}	True (population) value of correlation of level-1 units (observations) within a level-2 unit (nurse)
ρ_{23}	True (population) value of correlation of level-2 units (nurse) within a level-3 unit (ward)
$\tilde{\rho}_{23}$	Sample estimated value of correlation of level-2 units (nurse) within a level-3 unit (ward)
ρ_{34}	True (population) value of correlation of level-3 units (ward) within a level-4 unit (nursing home)
$\tilde{\rho}_{34}$	Sample estimated value of correlation of level-3 units (ward) within a level-4 unit (nursing home)
n_1	Number of level-1 units (observations) per level-2 unit (nurse)
n_2	Number of level-2 units (nurses) per level-3 unit (ward)
n_3	Number of level-3 units (wards) per level-4 unit (nursing home)
s	Number of sequences in a stepped wedge (also if part of a larger design)
c	Number clusters in a sequence of a stepped-wedge design
T	Number of measurement times/periods (including the baseline)
I	Total number of clusters (nursing homes)
τ^2	$Cov(Y_{it\bullet}, Y_{is\bullet})$: covariance between averages of the same cluster at different times t and s
$\tau^2 + \sigma^2$	$Var(Y_{it\cdot})$: variance of a cluster average at a time t
σ^2_1	Variance at level 1, that is, variance of level-1 units (observations) within their level-2 unit (nurse)
σ^2_2	Variance at level 2, that is, variance of level-2 units (nurses) within their level-3 unit (ward)
σ^2_3	Variance at level 3, that is, variance of level-3 units (wards) within their level-4 unit (nursing home)
σ^2_4	Variance at level 4, that is, variance between level-4 units (nursing homes)
VIF_p	Variance inflation factor due to the multilevel structure of the data having p levels
ρ	$Corr(Y_{it\bullet}, Y_{is\bullet})$ correlation between averages of the same cluster at different times t and s

$$VIF_{rm:cPG1} = \left\{ \begin{array}{c} VIF_{SW_s:cPG1} = \frac{3}{2} \cdot \frac{(1-\rho)\cdot(1+s\cdot\rho)}{\left(s-\frac{1}{s}\right)\cdot\left(1+\frac{s}{2}\rho\right)} \\ VIF_{SW_s(a,b):cPG1} = \frac{3}{2} \cdot \frac{(1-\rho)\cdot(1+[a+b-2+s]\cdot\rho)}{\left(s-\frac{1}{s}\right)\cdot\left(1+[a+b-2+\frac{s}{2}]\cdot\rho\right)} \\ VIF_{H(\beta,s):cPG1} = \frac{(1-\rho)}{T} \cdot \frac{1}{1-\frac{\beta^2}{3}\left(1+\frac{2}{s^2}\right)+R\cdot\left(1-\frac{\beta}{3}\left[2+\frac{1}{s^2}\right]\right)} \end{array} \right\} \quad (7)$$

and thus, the variance inflation factor compared to a parallel group *individually randomized* design with one measurement (using a t-test) is then

$$VIF = VIF_{rm:cPG1} \cdot VIF_p \quad (8)$$

where VIF_p is the variance inflation factor due to multilevel structure as explained above.

From equation (8), we can see that the total variance inflation comes from two aspects of the design: the manner of assigning intervention over the measurement times and the multilevel structure at each measurement time.

Sample size and power calculation

As sample size formulas and programs for a parallel group individually randomized designs with one measurement (i.e. post-test design) are readily available, sample size calculation for the stepped-wedge trial with p levels can easily be performed by first calculating the *total* sample size $N_{tot,PG1}$ (to detect a prespecified effect δ with prespecified power of $(1-\beta)\cdot 100\%$ at a significance level α). Note that most programs and formulas give the number of subjects per arm, so for the total sample size, this needs to be doubled. After that, we multiply this total sample size by the variance inflation factors to account for the multilevel stepped-wedge design. For a "standard" stepped-wedge design, the total sample size at each measurement time $N_{tot,t}$ (i.e. the total required number of level-1 units across all clusters and arms at each measurement time/period) is

$$N_{tot,t} = VIF \cdot N_{tot,PG1} = VIF_{rm} \cdot VIF_p \cdot N_{tot,PG1} \quad (9)$$

and dividing this by the number of level-1 units per cluster at each measurement yields the total required number of clusters (I). Dividing this total number of clusters by the number of steps gives the number of clusters per sequence $c = I/s$ (in the hybrid design after accounting for the fraction β). The parameters ρ and VIF_p needed to calculate VIF follow from Table 2 for three-level and four-level designs or from the arguments used in the SF for p-levels designs.

Instead of calculating the total sample size (or number of clusters needed), power for a range of feasible configurations (i.e. number of clusters, sample size at different levels, and intracluster correlations) could be

Table 2. Formulas for standard stepped-wedge trials with two, three, or four levels.

Stepped-wedge scenarios	Conversion formulas	
Two levels	$\sigma_2^2 = \rho_{12} \cdot \sigma_{tot}^2$ $\sigma_1^2 = (1 - \rho_{12}) \cdot \sigma_{tot}^2$	$var(Y_{it\bullet}) = \frac{\sigma_{tot}^2}{n_1} VIF_2, \quad \sigma^2{}_{tot} = \sigma_2^2 + \sigma_1^2$ $VIF_2 = \left[1 + (n_1 - 1)_{12}\right], \quad \rho_{12} = \frac{\sigma_2^2}{\sigma_2^2 + \sigma_1^2}$
	Covariance τ^2 and variance $\tau^2 + \sigma^2$ of cluster-time averages	Correlation $\rho = corr(Y_{it\bullet}, Y_{is\bullet})$ and variance $var(Y_{it\bullet})$ of cluster-time averages
Level 2 (cluster) repeatedly measured Level 1 (e.g. subject) cross-sectionally measured	$\tau^2 = \sigma_2^2$ $\sigma^2 = \frac{\sigma_1^2}{n_1}$	$\rho = \frac{n_1 \rho_{12}}{\left[1 + (n_1 - 1)_{12}\right]}$
Three levels	$\sigma_3^2 = \rho_{23}\rho_{12} \cdot \sigma_{tot}^2$ $\sigma_2^2 = (1 - \rho_{23})\rho_{12} \cdot \sigma_{tot}^2$ $\sigma_1^2 = (1 - \rho_{12}) \cdot \sigma_{tot}^2$	$var(Y_{it\bullet}) = \frac{\sigma_{tot}^2}{n_1 n_2} VIF_3, \sigma^2{}_{tot} = \sigma_3^2 + \sigma_2^2 + \sigma_1^2$ $VIF_3 = \left[1 + (n_1 - 1)_{12}\right][1 + (n_2 - 1)\tilde{\rho}_{23}]$ $\rho_{12} = \frac{\sigma_3^2 + \sigma_2^2}{\sigma_3^2 + \sigma_2^2 + \sigma_1^2}$ $\tilde{\rho}_{23} = \tilde{\rho}_{23}(n_1) = \frac{\sigma_3^2}{\left[\sigma_3^2 + \sigma_2^2 + \frac{\sigma_1^2}{n_1}\right]} = \rho_{23} \frac{n_1 \rho_{12}}{[1 + (n_1 - 1)\rho_{12}]}$ $\rho_{23} = \frac{\sigma_3^2}{\sigma_3^2 + \sigma_2^2}$
	Covariance τ^2 and variance $\tau^2 + \sigma^2$ of cluster-time averages	Correlation $\rho = corr(Y_{it\bullet}, Y_{is\bullet})$ and variance $var(Y_{it\bullet})$ of cluster-time averages
Level 3 (cluster) repeatedly measured Levels 2 and 1 cross-sectionally (e.g. subjects and sub-clusters or observations and subjects)	$\tau^2 = \sigma_3^2$ $\sigma^2 = \frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1 n_2}$	$\rho = \frac{\rho_{12}\rho_{23} n_1 n_2}{\left[1 + (n_1 - 1)_{12}\right][1 + (n_2 - 1)\tilde{\rho}_{23}]}$
Level 3 (cluster) and Level 2 (subject) repeatedly measured; Level 1 (observation) measured cross-sectionally	$\tau^2 = \sigma_3^2 + \frac{\sigma_2^2}{n_2}$ $\sigma^2 = \frac{\sigma_1^2}{n_1 n_2}$	$\rho = \frac{\rho_{12} n_1 [1 + (n_2 - 1)\rho_{23}]}{\left[1 + (n_1 - 1)_{12}\right][1 + (n_2 - 1)\tilde{\rho}_{23}]}$

(continued)

Table 2. Continued

Stepped-wedge scenarios	Conversion formulas	
Four levels	$\sigma_4^2 = \rho_{34}\rho_{23}\rho_{12} \cdot \sigma_{tot}^2$ $\sigma_3^2 = (1-\rho_{34})\rho_{23}\rho_{12} \cdot \sigma_{tot}^2$ $\sigma_2^2 = (1-\rho_{23})\rho_{12} \cdot \sigma_{tot}^2$ $\sigma_1^2 = (1-\rho_{12}) \cdot \sigma_{tot}^2$	$var(Y_{it\bullet}) = \frac{\sigma_{tot}^2}{n_1 n_2 n_3} VIF_4, \sigma_{tot}^2 = \sigma_4^2 + \sigma_3^2 + \sigma_2^2 + \sigma_1^2$ $VIF_4 = \left[1 + (n_1 - 1)_{12}\right][1 + (n_2 - 1)\tilde{\rho}_{23}][1 + (n_3 - 1)\tilde{\rho}_{34}]$ $\rho_{12} = \frac{\sigma_4^2 + \sigma_3^2 + \sigma_2^2}{\sigma_4^2 + \sigma_3^2 + \sigma_2^2 + \sigma_1^2}$ $\tilde{\rho}_{23} = \tilde{\rho}_{23}(n_1) = \frac{\sigma_4^2 + \sigma_3^2}{\sigma_4^2 + \sigma_3^2 + \sigma_2^2 + \sigma_1^2/n_1} = \rho_{23}\frac{n_1\rho_{12}}{[1 + (n_1 - 1)\rho_{12}]}, \rho_{23} = \frac{\sigma_4^2 + \sigma_3^2}{\sigma_4^2 + \sigma_3^2 + \sigma_2^2}$ $\tilde{\rho}_{34} = \tilde{\rho}_{34}(n_2, n_1) = \frac{\sigma_4^2}{\left[\sigma_4^2 + \sigma_3^2 + \frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1 n_2}\right]} = \rho_{34}\frac{n_2\tilde{\rho}_{23}}{[1 + (n_2 - 1)\tilde{\rho}_{23}]}, \rho_{34} = \frac{\sigma_4^2}{\sigma_4^2 + \sigma_3^2}$
	Covariance τ^2 and variance $\tau^2 + \sigma^2$ of cluster-time averages	Correlation $\rho = corr(Y_{it\bullet}, Y_{is\bullet})$ and variance $var(Y_{it\bullet})$ of cluster-time averages
Level 4 (cluster) repeatedly measured; levels 3 and 2 and 1 sampled cross-sectionally (e.g. sub-clusters and subjects and observations)	$\tau^2 = \sigma_4^2$ $\sigma^2 = \frac{\sigma_3^2}{n_3} + \frac{\sigma_2^2}{n_3 n_2} + \frac{\sigma_1^2}{n_3 n_2 n_1}$	$\rho = \frac{\rho_{12}\rho_{23}\rho_{34} n_1 n_2 n_3}{\left[1 + (n_1 - 1)_{12}\right][1 + (n_2 - 1)\tilde{\rho}_{23}][1 + (n_3 - 1)\tilde{\rho}_{34}]}$
Levels 4 (cluster) and 3 repeatedly measured; levels 2 and 1 sampled cross-sectionally	$\tau^2 = \sigma_4^2 + \frac{\sigma_3^2}{n_3}$ $\sigma^2 = \frac{\sigma_2^2}{n_3 n_2} + \frac{\sigma_1^2}{n_3 n_2 n_1}$	$\rho = \frac{\rho_{12}\rho_{23} n_1 n_2 [1 + (n_3 - 1)\rho_{34}]}{\left[1 + (n_1 - 1)_{12}\right][1 + (n_2 - 1)\tilde{\rho}_{23}][1 + (n_3 - 1)\tilde{\rho}_{34}]}$
Levels 4, 3, and 2 units repeatedly measured; level 1 sampled cross-sectionally	$\tau^2 = \sigma_4^2 + \frac{\sigma_3^2}{n_3} + \frac{\sigma_2^2}{n_3 n_2}$ $\sigma^2 = \frac{\sigma_1^2}{n_3 n_2 n_1}$	$\rho = 1 - \frac{1 - \rho_{12}}{\left[1 + (n_1 - 1)_{12}\right][1 + (n_2 - 1)\tilde{\rho}_{23}][1 + (n_3 - 1)\tilde{\rho}_{34}]}$

See Table 1 for definition of parameters and see the section "Variance inflation due to the multilevel structure" for their explanation.

calculated to see which configuration, if any, provides sufficient power. This can be done using the usual power formula

$$Power(\delta) = \Phi\left(\frac{\delta}{\sqrt{var\left(\hat{\delta}\right)}} - z_{1-\alpha/2}\right)$$

where Φ is the cumulative distribution function of the standard normal distribution and $z_{1-\alpha/2}$ is its $100\% \cdot (1-\alpha/2)$ percentile.

To calculate $var(\hat{\delta})$, equation (5a) with σ^2 and τ^2 can be applied or equation (5b) with ρ and $var(Y_{it\bullet})$ using the appropriate formulas for $\sigma^2, \tau^2, \rho, var(Y_{it\bullet})$ in Table 2. The latter comes down to using the variance inflation factors, that is, $var(\hat{\delta}) = VIF_{rm} \cdot VIF_p \cdot 4\sigma^2_{tot}/N_{tot,t}$ where $N_{tot,t}$ is the total number of level-1 units in the trial *at each measurement time/period*. For the standard stepped wedge, we can rewrite this to

$$\begin{aligned} var(\hat{\delta}) = {} & 4\sigma^2_{tot} \\ & \frac{3}{2} \cdot \frac{(1-\rho) \cdot [1 + s\rho]}{\left(s - \frac{1}{s}\right) \cdot \left[1 + \frac{s}{2}\rho\right]} \cdot \frac{1}{I} \cdot \frac{\left[1 + (n_1 - 1)_{12}\right]}{n_1} \\ & \cdot \frac{[1 + (n_2 - 1)\tilde{\rho}_{23}]}{n_2} \cdots \frac{\left[1 + \left(n_{p-1} - 1\right)\tilde{\rho}_{p-1,p}\right]}{n_{p-1}} \end{aligned} \quad (10)$$

in order to investigate the impact of various design parameters on the power. Figure 3 shows $VIF_{SW:cPG1}(s, \rho)$ for increasing values of ρ for various values of s, the number of sequences.

For a small number of clusters, the sample size and power formulas hold only approximately. For continuous, normally distributed outcomes, this is because of the low degrees of freedom, while for binary/rate outcomes, this is because formulas (2) and (4) depend on approximating the statistical distribution of cluster averages by a normal distribution using the central limit theorem. Therefore, we recommend the use of simulation studies to check power and also type I error for designs with a small number of clusters. However, the formulas in this article can be used to see whether feasible designs (i.e. in terms of number of clusters and/or number of measurements) would be worth such further investigation.

Binary and incidence outcomes

As the argumentation underlying the formulas relies on approximating the statistical distribution of cluster averages by the normal distribution using the central limit theorem, the formulas can be used for binary and incidence outcomes as well, provided the number of clusters is sufficiently large. We now discuss what value for σ^2_{tot} could be taken for non-small and small samples.

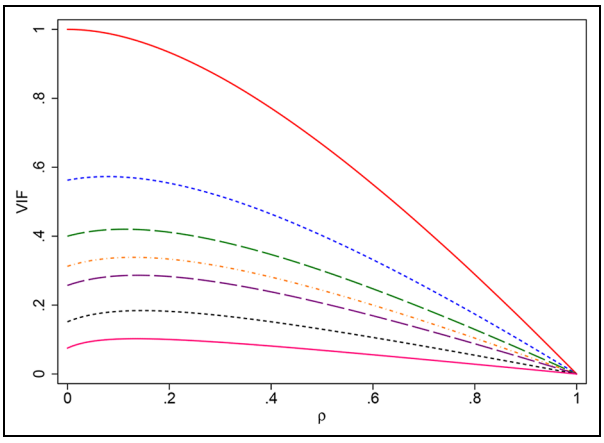

Figure 3. Variance inflation factor for the standard stepped wedge as a function of the correlation ρ between cluster averages over time. From top to bottom, the curves for the number of sequences $s = 2, 3, 4, 5, 6, 10, 20$ are shown.

If we take a two-level design and a binary outcome as an example, we can model the trial hierarchically as follows. Each subject j in cluster i has a binary outcome B_{ij} that is 1 with probability p_i, when cluster i is in the control condition, and with probability $p_i + \delta$, when cluster i is in the invention condition. The probabilities p_i vary over the clusters according to some distribution with mean μ and variance s^2_c. Then, the within-cluster variance in cluster i is $p_i(1-p_i)$ in the control condition. Over all clusters in the control condition, the expected total variance, that is, the variance of a level 1 unit regardless (unconditional) of the cluster it comes from, is $\sigma^2_{tot} = \mu(1-\mu)$, which can be decomposed into an expected within-cluster variance of $\sigma^2_1 = E[p_i(1-p_i)] = \mu(1-\mu) - s^2_c$ and between-cluster variance of $\sigma^2_2 = var(p_i) = s^2_c = \rho_{12} \cdot \sigma^2_1/(1-\rho_{12})$ (SF9). Because these expectations are averages that hold when the number of clusters is sufficiently large, it may make sense to take the following small-sample strategy. If we think that cluster-specific probabilities p_i will in practice mostly be between p_{min} and p_{max}, we take within that range the value p_{close} that is closest to 0.5, and set $\sigma^2_1 = p_{close}(1-p_{close})$, because that is the maximum value of the within-cluster variances $p_i(1-p_i)$ in the clusters in control condition. Noting that $\sigma^2_1 = (1-\rho_{12}) \cdot \sigma^2_{tot}$, we set the total variance to $\sigma^2_{tot,\,control} = p_{close}(1-p_{close})/(1-\rho_{12})$. The same reasoning could be applied when clusters are in the intervention condition, and thus, the largest (or average) of the two could be taken as σ^2_{tot}. This result also holds when there are more than two levels.

For a rate (incidence) outcome, the count (or rate) outcome of subject j in cluster i is R_{ij} that has expected value (average) λ_i, and these λ_i have mean λ and variance s^2_c. For a cluster in the control condition, the expected total variance, that is, the variance of a level-1

unit unconditional of the cluster it comes from, is $\sigma^2_{tot} = \lambda + s^2_c = \lambda/(1-\rho_{12})$ with $\sigma^2_1 = E[\lambda_i] = \lambda$ the expected (i.e. average over the clusters) within-cluster variance and $\sigma^2_2 = var(R_i) = s^2_c = \rho_{12}\cdot\sigma^2_1/(1-\rho_{12})$ the between-cluster variance. A conservative small sample strategy could then be to take $\sigma^2_1 = \lambda_{max}$, and thus set $\sigma^2_{tot,\,control} = \lambda_{max}/(1-\rho_{12})$, if we think that cluster-specific rate λ_i will in practice mostly fall between λ_{min} and λ_{max}. A similar reasoning applies when a cluster is in the intervention condition and the average or maximum of these two could be taken as σ^2_{tot}.

To illustrate sample size versus power calculations, for different endpoints, and small versus large sample considerations, we present two examples in the setting of the CHANGE trial. These were not the final calculations for this trial but similar to those performed.

Example 1: binary outcome in four-level standard stepped wedge

As a first example, we calculate power for hand hygiene compliance (a binary outcome) in a four-level standard stepped wedge using the following assumptions. The duration of the trial only allows four sequences ($s = 4$). The target effect size is an improvement from 20% to 35% ($\delta = 0.15$). It is assumed that the correlation among measurements within a nurse would be rather high ($\rho_{12} = 0.6$), while the correlation among nurses within a ward would be smaller ($\rho_{23} = 0.05$) and that of wards within a nursing home even smaller ($\rho_{34} = 0.01$). Based on feasibility, around five observations ($n_1 = 5$) per nurse, 15 nurses ($n_2 = 15$) per ward, maximally five wards per nursing home ($n_3 = 5$), and four nursing homes ($I = n_4 = 4$) would be possible. Given the small number of clusters (four nursing homes), it could make sense to take a conservative approach for the total variance σ^2_{tot} as was discussed above. If the level-1 probabilities are closest to 0.5 at $p_{closest} = 0.40$ (instead of 0.35) in the control condition and at $p_{closest} = 0.25$ (instead of 0.20) in the experimental condition, respectively, we take the average of the corresponding variances $\sigma^2_1 = (0.40\cdot 0.60 + 0.25\cdot 0.75)/2 = 0.21375$ and given that $(1-\rho_{12})\sigma^2_{tot} = \sigma^2_1$, the total variance is then $\sigma^2_{tot} = 0.21375/(1-0.6) = 0.534375$. If different nurses are sampled in each measurement time/period, level-2 and -1 units (nurses and measurements) are not repeated, and using the formulas in Table 2 (second scenario of the four-level standard stepped wedge)

$$\tau^2 = \sigma^2_4 + \frac{\sigma^2_3}{n_3} = \left(\rho_{34}\rho_{23}\rho_{12} + \frac{(1-\rho_{34})\rho_{23}\rho_{12}}{n_3}\right)\sigma^2_{tot} = \left(0.0003 + \frac{0.0297}{5}\right)\cdot 0.534375 \cong 33.345\cdot 10^{-4}$$

and

$$\sigma^2 = \frac{\sigma^2_2}{n_3 n_2} + \frac{\sigma^2_1}{n_3 n_2 n_1} = \left(\frac{(1-\rho_{23})\rho_{12}}{n_3 n_2} + \frac{1-\rho_{12}}{n_3 n_2 n_1}\right)\cdot\sigma^2_{tot} = \left(\frac{0.57}{5\cdot 15} + \frac{0.4}{5\cdot 15\cdot 5}\right)\cdot 0.534375 \cong 46.313\cdot 10^{-4}$$

so that

$$var(\hat{\delta}) = \frac{6}{I\cdot\left(s-\frac{1}{s}\right)}\cdot\sigma^2\cdot\left[1 + \frac{\frac{s}{2}\cdot\tau^2}{\sigma^2 + \left(1+\frac{s}{2}\right)\tau^2}\right] \cong \frac{6}{4\cdot\left(4-\frac{1}{4}\right)}\cdot 46.313\cdot 10^{-4}\cdot\left[1 + \frac{\frac{4}{2}\cdot 33.345\cdot 10^{-4}}{46.313\cdot 10^{-4} + \left(1+\frac{4}{2}\right)\cdot 33.345\cdot 10^{-4}}\right] \cong 26.967\cdot 10^{-4}$$

and $Power\,(\delta) = \Phi(\delta/\sqrt{\hat{\delta}} - z_{1-\alpha/2}) = \Phi(0.15/\sqrt{26.967\cdot 10^{-4}} - 1.96) = \Phi(0.928515) = 0.8234$. Figure 4 gives an impression of the sensitivity when one of the sample sizes or intracluster correlations is varied while the others are kept constant.

Example 2: rate outcome in a three-level standard stepped wedge

As second example, we use the variance inflation factor to calculate sample size for infection incidence (a rate). These rates are measured on patients within wards in nursing homes; hence, a 3-level design. We would expect the correlation of infection rates within wards to be high ($\rho_{12} = 0.7$), while infections in one ward would not automatically increase infections in another ward within the same nursing home, so a low correlation of ward-infection rates within a nursing home ($\rho_{23} = 0.01$). The effect of interest is a decrease from 11 to 5 infections per 1000 resident days ($\delta = 6\cdot 10^{-3}$). Anticipating a large number of clusters, we do not take $\sigma^2_1 = \lambda_{max}$ the maximum of the cluster-specific rates per condition but the average of the cluster-specific rate λ for each condition. Thus, $\sigma^2_1 = (\lambda_{ctl} + \lambda_{exp})/2 = 8\cdot 10^{-3}$ and $\sigma^2_{tot} = \sigma^2_1/(1-\rho_{12}) \cong 26.67\cdot 10^{-3}$. The total sample size in an equal size parallel group *individually* randomized design needed to detect this difference with 0.8 power at a significance level of 0.05 is

$$N_{tot,PG1} = 2\cdot 2\cdot\left(z_{1-\frac{0.05}{2}} + z_{0.8}\right)^2\cdot\sigma^2_{tot}/\delta^2 \cong 2\cdot 2\cdot 7.85\cdot 26.67\cdot 10^{-3}/\left(6\cdot 10^{-3}\right)^2 = 23,262$$

With $n_1 = 10$ patients per ward and $n_2 = 4$ wards per nursing home, the variance inflation due to clustering is $VIF_3 = [1 + (n_1 - 1)_{12}][1 + (n_2 - 1)\cdot\{\rho_{23}\cdot n_1$

$\cdot\rho_{12}/(1+(n_1-1)\rho_{12})\}] = [7.3][1+3\cdot\{0.01\cdot 10\cdot 0.7/(7.3)\}] = 7.51$. If we assume that only patients are cross-sectionally measured, we are in the second three-level scenario (Table 2) and $\rho = \rho_{12}\cdot n_1\cdot[1+(n_2-1)\rho_{23}]/VIF_3 = 0.7\cdot 10\cdot[1.03]/[7.51] \cong 0.96$. Thus, the variance inflation due to the stepped-wedge design is

$$VIF_{SW:cPG1} = \frac{3}{2}\cdot\frac{(1-\rho)\cdot[1+s\rho]}{\left(s-\frac{1}{s}\right)\cdot\left[1+\frac{s}{2}\rho\right]} = \frac{3}{2}\cdot\frac{(1-0.96)\cdot[1+4\cdot 0.96]}{\left(4-\frac{1}{4}\right)\cdot\left[1+\frac{4}{2}\cdot 0.96\right]} \cong 0.026$$

and the total variance inflation is $VIF_{SW} = 0.026\cdot 7.51 \cong 0.20$. Then, the total sample size needed *per measurement time/period* is $N_{tot,PG1}\cdot VIF_{SW} = 23,262\cdot 0.20 \cong 4652$ and the number of nursing homes (clusters) needed $N_{tot,PG1}\cdot VIF_{SW}/n_1n_2 \cong 4652/(10\cdot 4) \cong 116$, so four groups of 29 clusters should suffice.

Programs (SAS® and MS Excel®) to facilitate calculations are provided via https://github.com/steventeer-enstra/multilevel-stepped-wedge and in the SF (SAS® program only).

Discussion

Power and sample size formulas for stepped-wedge designs are typically restricted to two or three levels.[7,9] In this article, these formulas were extended to designs with more levels and it was demonstrated that they can either be expressed in terms of variance components or intracluster correlations. The latter expression clearly shows the separate effect of the multilevel structure within time and the stepped-wedge structure over time, similar to what has been shown for other designs but with two levels.[10,11]

From the formulas, it can be seen that the different design parameters have the following impact on power and sample size:

(I): Increasing the *number of clusters I* increases power (SF7.1).
(s): Increasing the *number of sequences s* increases power,[1,4,9] except for the case of the hybrid design and when the total cluster size over all measurements is constant (SF7.1).
(n_i): Increasing the *sample size at any level* increases power (SF7.2, Figure 4). We can achieve any desired power by sufficiently increasing the sample size at any of the levels that are measured as cohort and also by increasing the sample size of the first "cross-sectional" level that is below those levels (SF7.3). In particular, this also applies to the two-level stepped-wedge design, so by increasing the number of cross-sectionally measured subjects, we can reach any power level. This is in contrast with the parallel arm cluster-randomized trial that can plateau (potentially below 80%) if the number of subjects is increased indefinitely.[12] As a consequence, a lack of power due to a limited number of clusters can be compensated by increasing the sample size at particular lower levels. As one can see in Figure 4, not only the sample size at level 3, but also at level 2 can increase the power to 1, but power plateaus below 0.9 when increasing the sample size at level 1. This behavior can most easily be understood in a two-level stepped-wedge trial. As the random effect of a cluster is assumed not to vary over time, the within-cluster comparison is actually a comparison of all subjects before switching to the intervention and after, because the random effect of cluster drops out of the equation. This means that the within-cluster comparisons can become arbitrarily precise with increasing level 1 sample size and this drives the power to 1.
($\rho_{u,u+1}$): Unlike in parallel group cluster-randomized trials, an increase in the *intracluster correlation coefficients* does not necessarily mean a decrease in power, but actually may increase power in some situations as can be observed in Figure 4. This is because increasing an intracluster correlation $\rho_{u,u+1}$ influences the power both via the variance inflation factor due to the multilevel structure, VIF_p, and via the stepped-wedge design variance inflation factor, $VIF_{SW:cPG1}$. The first factor, VIF_p, will linearly increase with $\rho_{u,u+1}$ (Formula (6)). However, $VIF_{SW:cPG1}$ will generally first increase and then decrease when an intracluster correlations $\rho_{u,u+1}$ increases. This is because with increasing $\rho_{u,u+1}$, the correlation ρ between averages of the same cluster at different times/periods will increase as well (SF7.4), but $VIF_{SW(s)}(\rho)$ will first increase with increasing ρ until some turning point and then decrease as is illustrated in Figure 3. Intuitively, this decrease can be understood because the standard stepped wedge depends on between- and within-cluster comparisons. The between-cluster comparisons will become less precise when the correlation $\rho_{u,u+1}$ increases, but the precision will be dominated by the within-cluster comparisons for larger $\rho_{u,u+1}$. In the within-cluster comparisons, the random effects for clustering drop out, and so increasing $\rho_{u,u+1}$ will mean that the units at level u before and after the switch will be better correlated, so the within-cluster comparison will be more precise. All in all, an increasing intracluster correlation $\rho_{u,u+1}$ can thus give different patterns for the variance inflation and power. For example, when the increasing behavior of VIF_p dominates for small $\rho_{u,u+1}$, while for larger $\rho_{u,u+1}$ the decreasing behavior of $VIF_{SW(s)}$ dominates, then we would see power first decrease and then increase as a function of $\rho_{u,u+1}$. Another typical behavior is that power decreases with increasing $\rho_{u,u+1}$, because the increasing behavior of VIF_p dominates that of $VIF_{SW(s)}$ for all values of $\rho_{u,u+1}$. Both behaviors can be seen in Figure 4.

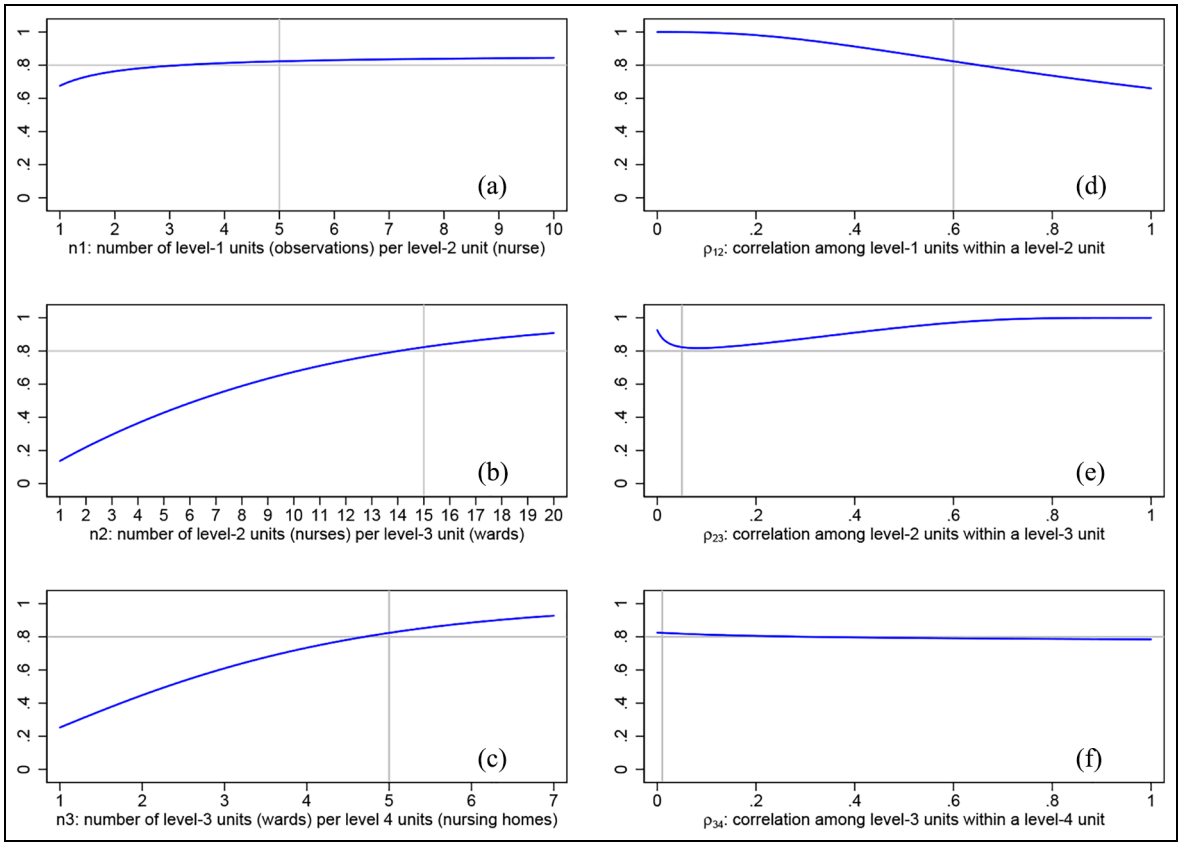

Figure 4. Impact of cluster size and intracluster correlations at different levels in a "standard" stepped wedge. Power of the 4 level "standard" stepped-wedge trial of Example 1 when varying either one sample size (part a-c) or one intracluster correlation (part d-f) at the specified level while keeping the other sample sizes and intracluster correlations constant. The vertical reference lines indicate the values of sample size and intracluster correlation as in Example 1 ($\rho_{12} = 0.6, \rho_{23} = 0.05, \rho_{34} = 0.01, n_1 = 5, n_2 = 15, n_3 = 5, n_4 = 1 = 4$).

Both increasing sample size and intracluster correlations coefficients can have unexpected power properties due to the random effects canceling out. Therefore, one may question how realistic it is to assume that the random effects (of a cluster) are not varying over time. This assumption implies that the correlation of two subjects within a cluster is the same whether they are measured at the same time t or at different times. It also implies that the correlation ρ of cluster means at different times only depends on intracluster correlations $\rho_{u,u+1}$, that is, correlations at a fixed time (Table 2). For some outcomes in type-2 diabetes, Martin et al.[13] found this not to be the case in a two-level setting. More empirical research is needed to see whether and when an assumption of constant correlation over time is reasonable; if this is not the case, then power will be lower than what is calculated from our formulas.[11,14]

The variance components or intracluster correlation coefficients needed for the calculations should preferably be estimated from studies with similar outcomes and context. These studies should have the same number of levels, but do not need to be stepped wedge, prospective, or randomized. In the absence of such studies, content-matter specialists could provide plausible values, and they could do so either in terms of variance components or intracluster correlations. Given the uncertainties in these educated guesses, we recommend that a range of plausible values for each of these parameters be considered.

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 in part by the Netherlands Organisation for Health Research and Development (ZonMw; grant no. R522002009).

Supplemental material

Supplemental material for this article is available online.

Trial described

CHANGE trial (ClinicalTrial.gov NCT02817282).

References

1. Hussey M and Hughes J. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007; 28: 182–191.
2. Girling AJ and Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Stat Med* 2016; 35: 2149–2166.
3. 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* 2016; 14: 639–647.
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: 752–758.
5. Donner A and Klar N. *Design and analysis of cluster randomization trials in health research*. London: Arnold, 2000.
6. Teerenstra S, Moerbeek M, van Achterberg, et al. Sample size calculations for 3-level cluster randomized trials. *Clin Trials 2008*; 5: 486–495.
7. 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: 181–196.
8. Teerenstra S, Lu B, Preisser J, et al. Sample size considerations for GEE analyses of three-level cluster randomized trials. *Biometrics* 2010; 66: 1230–1237.
9. Heo M, Kim N, Rinke ML, et al. Sample size determinations for stepped-wedge clinical trials from a three-level data hierarchy perspective. *Stat Methods Med Res* 2018; 27: 480–489.
10. Teerenstra S, Eldridge S, Graff M, et al. A simple sample size formula for analysis of covariance in cluster randomized trials. *Stat Med* 2012; 31: 2169–2178.
11. 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: 4718–4728.
12. Guittet L, Giraudeau B and Ravaud P. A priori postulated and real power in cluster randomized trials: mind the gap. *BMC Med Res Methodol* 2005; 5: 25.
13. Martin J, Girling A, Nirantharakumar K, et al. Intracluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care. *Trials* 2016; 17: 402.
14. Kasza J, Hemming K, Hooper R, et al. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Stat Methods Med Res*. Epub ahead of print 1 January 2017. DOI: 10.1177/0962280217734981.