

# Relative efficiency of unequal cluster-sizes in stepped wedge and other trial designs under longitudinal or cross-sectional sampling

Girling, Alan

DOI:  
[10.1002/sim.7943](https://doi.org/10.1002/sim.7943)

License:  
Creative Commons: Attribution-NonCommercial (CC BY-NC)

Document Version  
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):  
Girling, A 2018, 'Relative efficiency of unequal cluster-sizes in stepped wedge and other trial designs under longitudinal or cross-sectional sampling', *Statistics in Medicine*. <https://doi.org/10.1002/sim.7943>

[Link to publication on Research at Birmingham portal](#)

#### Publisher Rights Statement:

This is the peer reviewed version of the following article: Alan Girling, Relative efficiency of unequal cluster sizes in stepped wedge and other trial designs under longitudinal or cross-sectional sampling, *Statistics of Medicine* which has been published in final form at [10.1002/sim.7943](https://doi.org/10.1002/sim.7943). This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions

#### General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

#### Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact [UBIRA@lists.bham.ac.uk](mailto:UBIRA@lists.bham.ac.uk) providing details and we will remove access to the work immediately and investigate.

# Relative efficiency of unequal cluster sizes in stepped wedge and other trial designs under longitudinal or cross-sectional sampling

Alan J. Girling 

Institute of Applied Health Research,  
University of Birmingham, Birmingham  
B15 2TT, UK

## Correspondence

Alan J. Girling, Institute of Applied Health Research, University of Birmingham, Birmingham B15 2TT, UK.  
Email: A.J.Girling@bham.ac.uk

## Funding information

National Institute for Health Research (NIHR); Collaborations for Leadership in Applied Health Research and Care for West Midlands (CLAHRC WM); HiSLAC

**Background:** A cluster trial with unequal cluster sizes often has lower precision than one with equal clusters, with a corresponding inflation of the design effect. For parallel group trials, adjustments to the design effect are available under sampling models with a single intracluster correlation. Design effects for equal clusters under more complex scenarios have appeared recently (including stepped wedge trials under cross-sectional or longitudinal sampling). We investigate the impact of unequal cluster size in these more general settings.

**Results:** Assuming a linear mixed model with an exchangeable correlation structure that incorporates cluster and subject autocorrelation, we compute the relative efficiency (RE) of a trial with clusters of unequal size under a size-stratified randomization scheme, as compared to an equal cluster trial with the same total number of observations. If there are no within-cluster time effects, the RE exceeds that for a parallel trial. In general, the RE is a weighted average of the RE for a parallel trial and the RE for a crossover trial in the same clusters. Existing approximations for parallel designs are extended to the general setting. Increasing the cluster size by the factor  $(1 + CV^2)$ , where  $CV$  is the coefficient of variation of cluster size, leads to conservative sample sizes, as in a popular method for parallel trials.

**Conclusion:** Methods to assess experimental precision for single-period parallel trials with unequal cluster sizes can be extended to stepped wedge and other complete layouts under longitudinal or cross-sectional sampling. In practice, the loss of precision due to unequal cluster sizes is unlikely to exceed 12%.

## KEYWORDS

cluster randomized trial, design effect, stepped wedge, unequal cluster sizes

## 1 | BACKGROUND AND SCOPE

In a cluster trial, the presence of unequally sized clusters can affect the precision of the treatment effect estimate. Suggested modifications to the equal cluster design effects (DEs) to account for this have focussed mostly on the parallel group trial under a simplified statistical model with a single intracluster correlation coefficient (ICC).<sup>1–4</sup> In this case the,

This is an open access article under the terms of the Creative Commons Attribution NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2018 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

precision is reduced compared to a trial in which the same total number of observations is equally distributed across the clusters.

Alternatives to the parallel group design include cluster crossover designs and stepped wedge designs in both of which the experimental condition (ie, treatment/control status) is varied during the course of the study. For such designs, it is natural to consider more complex statistical models in which subject and cluster effects vary over time.<sup>5-7</sup> In a recent paper,<sup>8</sup> Hooper et al have presented DEs for “complete” cluster designs, including parallel, crossover, and stepped wedge layouts, under either cross-sectional or longitudinal sampling and with allowance for within-cluster variation over time. This work encompasses a wide range of possibilities, in particular, the model of Hussey and Hughes<sup>9</sup> is subsumed as a special case, but the results apply only to designs in which the clusters are equal in size. Here, we consider the impact of variable cluster size on these DEs. We give expressions for the relative efficiency (RE) of the treatment effect estimate relative to that for the equal cluster design with the same total number of observations, which can then be used to modify the equal cluster DE.

The RE is derived under a stratified cluster randomization scheme for which the distribution of cluster sizes is the same within each design sequence (study arm). This is the same condition under which standard modifications to the parallel group DE have been derived.<sup>4</sup>

Throughout this work, “cluster size” refers to the total number of subjects (or observations) in a cluster and is allowed to vary between but not within clusters, so that a constant rate of sampling within each individual cluster is assumed. This corresponds to a (common) practical situation in which the clusters in a study can be identified with institutions of different sizes (eg, hospitals, care homes, general practices) with different capacities for recruitment. We do not consider the implications of uneven sampling within individual clusters.

The statistical model is described in Section 2. The main results on the RE of designs with unequal cluster sizes are presented in Section 3. Section 4 describes some practical implications with calculations for two recent trials. Conclusions and limitations are discussed in Section 5.

## 2 | STATISTICAL MODEL

We introduce unequally sized clusters into an existing linear mixed effects model<sup>8,10</sup> for closed-cohort trials, in which the same subjects are observed at each timepoint under an exchangeable correlation structure. Results for cross-sectional trials (where observations at different times are made on different subjects) can be obtained by reducing the subject-level variation to zero in this model.

We assume that a fixed number of clusters is observed over  $T$  epochs. There are two experimental conditions indexed by the binary array  $X_{lj}$ ,  $l = 1, \dots, L; j = 1, \dots, T$ , where  $X_{lj} = 0$  or 1 indicates “control” or “treated” conditions and  $L$  denotes the number of treatment sequences (or “arms”) in the trial. There are  $N (= rL)$  clusters divided by size into  $r$  strata of  $L$  clusters each and allocated to treatment arms according to a stratified scheme, which ensures that the size distribution is the same in each arm. In any given cluster, the number of subjects ( $m_k$ ) is a constant which depends only on the stratum  $k (= 1, \dots, r)$  to which the cluster belongs. The observations are generated by a linear mixed model of the form

$$Y_{ijkl} = \beta_j + \theta X_{lj} + c_{kl} + s_{ikl} + (ct)_{jkl} + (st)_{ijkl}. \quad (1)$$

Here,  $Y_{ijkl}$  is the outcome at time  $j$  ( $j = 1, \dots, T$ ) for the  $i$ th subject ( $i = 1, \dots, m_k$ ) in the cluster from stratum  $k$  in arm  $l$ ;  $\beta_1, \dots, \beta_T$  are fixed time effects and  $\theta$  is the treatment effect whose estimation is the purpose of the study. The other terms in (1) represent independent random effects:  $c$ , between clusters;  $s$ , between subjects;  $(ct)$ , cluster by time interaction; and  $(st)$ , subject by time interaction (including measurement error). These have variances  $\eta_C$ ,  $\eta_S$ ,  $\eta_{CT}$ , and  $\eta_{ST}$  respectively, with  $\eta_C + \eta_S + \eta_{CT} + \eta_{ST} = \sigma^2 = \text{var} Y_{ijkl}$ .

### 2.1 | Correlation structure

We follow an established notation<sup>8</sup> to characterize the correlation structure:

$$\begin{aligned} (\text{ICC}) \qquad \rho &= \frac{\eta_C + \eta_{CT}}{\eta_C + \eta_S + \eta_{CT} + \eta_{ST}} & (0 \leq \rho < 1) \\ (\text{cluster autocorrelation}) \qquad \pi &= \frac{\eta_C}{\eta_C + \eta_{CT}} & (0 \leq \pi \leq 1) \\ (\text{individual autocorrelation}) \qquad \tau &= \frac{\eta_S}{\eta_S + \eta_{ST}} & (0 \leq \tau < 1) \end{aligned}$$

(It is assumed throughout that  $\eta_{ST} > 0$ ).

In a cross-sectional study, each subject contributes data only at a single timepoint so that a cluster of “size”  $m_k$  will contain a total of  $Tm_k$  subjects overall. A model for this scenario is obtained by setting the between subject effect,  $s$ , to zero in (1), whereupon  $\eta_S = 0$  and  $\tau = 0$ . The Hussey and Hughes model<sup>9</sup> omits both subject-level effects and cluster by period effects and relies on a single ICC to characterize the correlation structure. This corresponds to the particular case, where  $\tau = 0$  and  $\pi = 1$ .

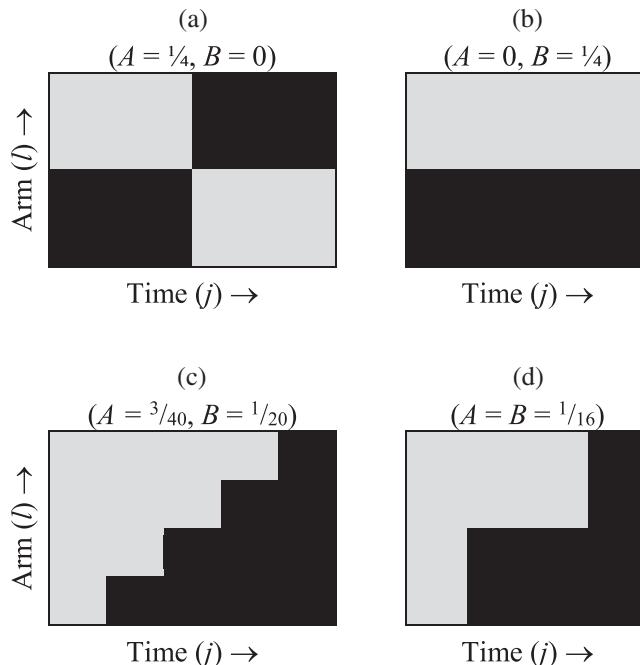
## 2.2 | The experimental layout

For the results given here, the experimental layout ( $X_{lj}$ ) is characterized by two coefficients:  $A = \frac{1}{LT} \sum_{j=1}^T \sum_{l=1}^L (X_{lj} - \bar{X}_{\bullet j} - \bar{X}_{l\bullet} + \bar{X}_{\bullet\bullet})^2$  and  $B = \frac{1}{L} \sum_{l=1}^L (\bar{X}_{l\bullet} - \bar{X}_{\bullet\bullet})^2$ . These coefficients are given elsewhere<sup>10</sup> for some standard designs, including those in Table 1.

The values of these coefficients depend only on the “geometry” of the layout, ie, the pattern of light and shade in Figure 1, and not on the numbers of clusters or time periods in the design. For instance, a parallel cluster trial over a single time period has exactly the same coefficient values as a parallel trial with observations taken over several time periods; an SW<sub>g</sub> design with a single cluster randomized at each step has the same coefficients as an SW<sub>g</sub> design with several clusters at each step provided this number is the same for all steps. Most existing work on the efficiency of unequal cluster designs applies only to two-arm, single-period parallel layouts, a special case of layout (b) in Figure 1 with  $T = 1$ .

**TABLE 1** Coefficients  $A, B$  for four standard design layouts (illustrated in Figure 1) taken from table 1 of Girling and Hemming,<sup>10</sup> with  $B \equiv b$  and  $A \equiv a - b$ , in their notation

Design Layout	<b><math>A</math></b>	<b><math>B</math></b>
(a) Crossover	$1/4$	0
(b) Parallel	0	$1/4$
(c) g-Step Wedge (SW <sub>g</sub> )	$\frac{1}{12} \left(1 - \frac{2}{g(g+1)}\right)$	$\frac{1}{12} \left(1 - \frac{2}{g+1}\right)$
(d) Delay-control design (DCD <sub>p,q,r</sub> )	$\frac{1}{4}q(1-q)$	$\frac{1}{4}q^2$



**FIGURE 1** Diagrammatic representation of four experimental layouts ( $X_{lj}; 1 \leq l \leq L, 1 \leq j \leq T$ ). Light-shading indicates  $X_{lj} = 0$ , dark-shading indicates  $X_{lj} = 1$ . The values of the coefficients  $A$  and  $B$  depend only on the patterns of light and shade in these figures and not on the dimensions  $L$  and  $T$ . (a) Crossover; (b) Parallel; (c) SW<sub>4</sub>; (d) DCD<sub>0.25, 0.5, 0.25</sub>

### 3 | RELATIVE EFFICIENCY OF UNEQUAL CLUSTER DESIGNS

Let  $\text{var}(\hat{\theta}; m_1, m_2, \dots, m_r)$  denote the sampling variance of the best linear unbiased estimate of  $\theta$  in model (1) with (possibly unequal) cluster sizes  $m_1, \dots, m_r$ . The RE of this design compared to the equal cluster design with the same total number of observations is defined as

$$\text{RE} = \frac{\text{var}(\hat{\theta}; m, m, \dots, m)}{\text{var}(\hat{\theta}; m_1, m_2, \dots, m_r)},$$

where  $m = \frac{1}{r} \sum_{k=1}^r m_k$ , the mean cluster size.

Our main result states that

$$\text{RE} = \frac{A\Psi(\lambda_0 m^{(p)}) + B\nu\Psi(\lambda_1 m^{(p)})}{A + B\nu}, \quad (2)$$

where

$$\begin{aligned} m^{(p)} &= m \frac{\rho}{1 - \rho} & (m^{(p)} \geq 0) \\ \lambda_0 &= \lambda_0(\pi, \tau) = \frac{1 - \pi}{1 - \tau} & (\lambda_0 \geq 0) \\ \lambda_1 &= \lambda_1(\pi, \tau) = \frac{1 + (T - 1)\pi}{1 + (T - 1)\tau} & (\lambda_1 \geq 0) \\ \nu &= \frac{1 - \tau + m^{(p)}(1 - \pi)}{1 + (T - 1)\tau + m^{(p)}(1 + (T - 1)\pi)} & (0 \leq \nu \leq 1) \end{aligned}$$

(The parameter  $\nu$  can be interpreted as 1 minus the cluster-mean correlation<sup>10</sup> for a cluster of size  $m$  – ie, with  $Tm$  observations overall.)

The function  $\Psi(\cdot)$  is defined as

$$\Psi(\alpha) = (1 + \alpha) \times \frac{1}{r} \sum_{k=1}^r \frac{m_k}{m + \alpha m_k}, \text{ for } \alpha \geq 0. \quad (3)$$

The functional form of  $\Psi$  is determined by the distribution of cluster sizes relative to the mean cluster size, ie, on the distribution of  $z_k = m_k/m$ . In the case of equal clusters (ie,  $z_k \equiv 1$ ),  $\Psi(\alpha) = 1$  for all  $\alpha$ , as expected. For clusters unequal in size,  $\Psi(0) = 1$ , and  $\Psi(\alpha) \leq 1$  for  $\alpha > 0$ , with a unique minimum in the range  $(0, \infty]$ , as shown in the appendix. In general,  $\Psi(\infty)$  equals the proportion of nonzero clusters in the design. In most practical applications,  $\Psi(\infty) = 1$  and the impact of cluster inequality is low when the scaled mean cluster size,  $m^{(p)} = mp/(1 - \rho)$ , is either very small or very large.

The DE for the unequal cluster design is given by

$$\text{DE} = \text{DE}_0 \div \text{RE}, \quad (4)$$

where the equal cluster DE is

$$\text{DE}_0 = \frac{T\nu(1 + (m - 1)\rho)}{4(1 + (T - 1)\nu)(A + B\nu)}. \quad (5)$$

Alternatively,  $\text{DE}_0$  may be obtained using the more transparent approach given by Hooper et al.<sup>8</sup>

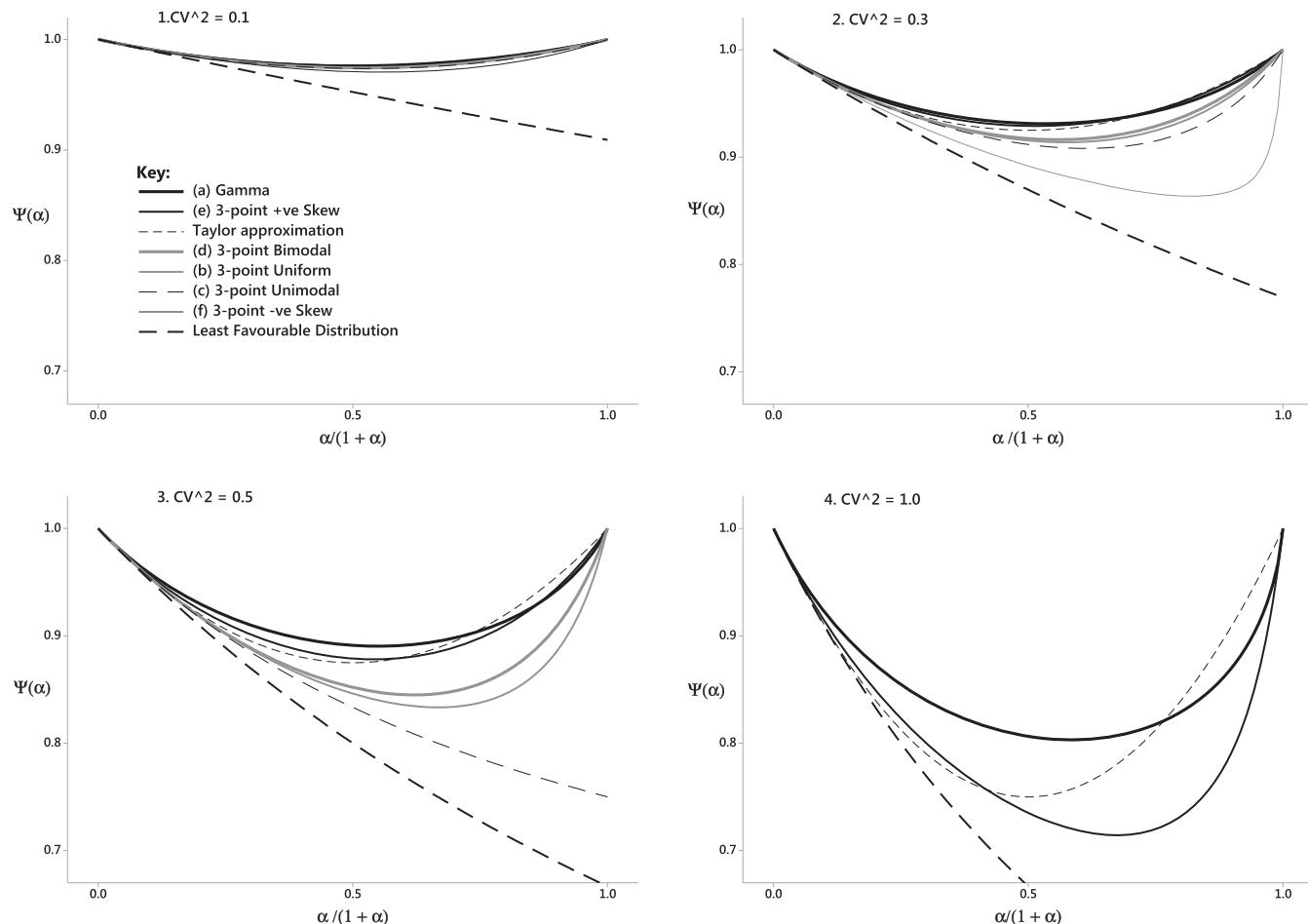
#### 3.1 | Relative efficiency for parallel and crossover layouts

Under a parallel layout, the coefficients from Table 1 are  $A = 0$ ,  $B = 1/4$ , and (2) reduces to

$$\text{RE}_{\text{Parallel}} = \Psi(\lambda_1 m^{(p)}) \equiv \frac{1 + (\lambda_1 m - 1)\rho}{mr} \sum_{k=1}^r \frac{m_k}{1 + (\lambda_1 m_k - 1)\rho}. \quad (6)$$

This equation generalizes the result<sup>4</sup> for single-period parallel designs, for which  $T = 1$  and  $\lambda_1 = 1$ .

For any given set of relative cluster sizes ( $z_k$ ),  $\text{RE}_{\text{Parallel}}$  depends only on a scaled version of the mean cluster size,  $m$ , namely,  $\lambda_1 m^{(p)} = m \times \frac{\lambda_1 \rho}{1 - \rho}$ . In principle, this means that the behavior of the RE as a function of the mean cluster size under a complex correlation structure can be read from a plot of  $\Psi(\alpha)$  with an appropriate scaling of the  $\alpha$ -axis. For example,



**FIGURE 2** The function  $\Psi(\alpha)$  for a selection of cluster size distributions, together with the Taylor approximation due to van Breukelen. The argument of  $\Psi(\alpha)$  is given on the transformed scale  $u = \alpha/(1 + \alpha)$ , which enables the complete range of  $\alpha$ -values to be plotted on a bounded interval, with  $u = 1$  when  $\alpha = \infty$ . The key is the same in all 4 panels and is ordered according to ordinate values in the middle of the range. The three-point specimen distributions (b), (c), (d), (e), and (f) are available only at lower CV-values (see Table 2): so (f) is missing from panel 3 ( $CV^2 = 0.5$ ) and only (e) appears in panel 4 ( $CV^2 = 1$ )

in Figure 2, the value  $\alpha = 1/3$  (corresponding to 0.25 on the horizontal scale) gives the RE when  $\lambda_1 m^{(\rho)} = 1/3$ , ie, when the mean cluster size  $m$  is equal to  $\frac{1-\rho}{3\lambda_1\rho}$ .

Similarly, in a crossover layout,  $A = 1/4$ ,  $B = 0$  and

$$RE_{\text{Crossover}} = \Psi(\lambda_0 m^{(\rho)}) .$$

This has the same form as  $RE_{\text{Parallel}}$  with  $\lambda_1$  replaced by  $\lambda_0$ . If  $\lambda_1$  and  $\lambda_0$  are both positive (ie,  $\pi < 1$ ), then  $RE_{\text{Parallel}}$  and  $RE_{\text{Crossover}}$  run through exactly the same range of values as  $m^{(\rho)}$  varies. Effectively  $RE_{\text{Crossover}}$  is just the RE for a parallel study with different values  $(\pi', \tau')$  for the two autocorrelation parameters chosen so that  $\lambda_1(\pi', \tau') = \lambda_0(\pi, \tau)$ . In the (admittedly unusual) case that  $\pi = \tau < 1$ , then  $\lambda_1 = \lambda_0 = 1$  and the RE is independent of  $T$  and is the same for both parallel and crossover layouts, and indeed for any complete layout whatsoever (from Equation (2)). An important special case arises when  $\pi = 1$ . This corresponds to a complete absence of cluster-level time effects, as in the Hussey and Hughes model. Here,  $\lambda_0 = 0$ ,  $RE_{\text{Crossover}} \equiv 1$ , and there is no loss of precision associated with unequal cluster sizes.

The general behavior of the function  $\Psi$  is illustrated in Figure 2 by sampling the relative cluster sizes  $z_k (=m_k/m)$  from several different probability distributions with mean 1 and specified coefficient of variation (CV = standard deviation  $\div$  mean). These include (a) a Gamma distribution with shape parameter  $= 1/CV^2$ , and five three-point distributions, labeled (b) to (f) in Table 2, based on those suggested by van Breukelen et al<sup>4</sup> for a similar purpose. Plots of the expected value of  $\Psi(\alpha)$  are shown in Figure 2 for  $CV^2 = 0.1, 0.3, 0.5$ , and  $1.0$ .

**TABLE 2** Some specimen distributions to illustrate the impact of varying CV on the relative efficiency. Each distribution has mean 1 and is fully specified once the CV is given. The final column shows the greatest value of the  $CV^2$  for which the distribution is available (given that  $z_i \geq 0$  necessarily). The Least Favorable Distribution minimizes  $\Psi(\alpha)$  for all  $\alpha \geq 0$  over all such distributions with nonnegative support

	Support	Probabilities	Range	CV <sup>2</sup>	Max CV <sup>2</sup>
Three-point symmetrical distributions					
(b) Uniform ( $p = 1/3$ )	{ $a, 1, 2 - a$ }	{ $p, 1 - 2p, p$ }	$2(1 - a)$	$2p(1 - a)^2$	0.6667
(c) Unimodal ( $p = 1/4$ )					0.5000
(d) Bimodal ( $p = 2/5$ )					0.8000
Three-point skew distributions					
(e) Positive skew	{ $1 - S/3, 1 + S/6, 1 + 2S/3$ }	{ $1/2, 1/3, 1/6$ }	S	$5S^2/36$	1.2500
(f) Negative skew	{ $1 - 2S/3, 1 - S/6, 1 + S/3$ }	{ $1/6, 1/3, 1/2$ }	S	$5S^2/36$	0.3125
Least Favorable Distribution (LFD)	{ $0, 1 + c^2$ }	{ $\frac{c^2}{1+c^2}, \frac{1}{1+c^2}$ }	$1 + c^2$	$c^2$	$\infty$

Other things being equal,  $\Psi(\alpha)$ , and hence the RE, is smaller for higher values of the CV. The Gamma and three-point positive skew distributions give similar results for  $\Psi(\alpha)$  over values of the CV most likely to be encountered in practice (ie,  $CV^2 = 0.1, 0.3, 0.5$  in panels 1, 2, 3). The other three-point distributions are associated with more extreme (ie, lower) values as  $\alpha$  increases. Notice that the three-point unimodal distribution with  $CV^2 = 0.5$  entails a finite probability for clusters of size zero so that  $\Psi(\alpha)$  does not return to 1 as  $\alpha \rightarrow \infty$ . In general,  $\Psi(\infty) =$  the proportion of nonzero clusters in the design.

For any cluster size distribution with  $CV = c$ ,

$$\Psi(\alpha) \geq \Psi_{\text{LFD}}(\alpha) = \frac{1 + \alpha}{1 + \alpha + \alpha c^2} > \frac{1}{1 + c^2}, \text{ for } \alpha \geq 0. \quad (7)$$

The expression after the “=” is the actual value of  $\Psi(\alpha)$  under a particular two-point distribution for relative cluster sizes. This “Least Favorable Distribution” (LFD) is given by  $\Pr\{z_k = 1 + c^2\} = \frac{1}{1+c^2} = 1 - \Pr\{z_k = 0\}$  and included in Figure 2. Although the LFD is unlikely to be a realistic model in practice (because it involves clusters of size 0), it provides an absolute bound on the possible loss of efficiency under any distribution whatsoever.

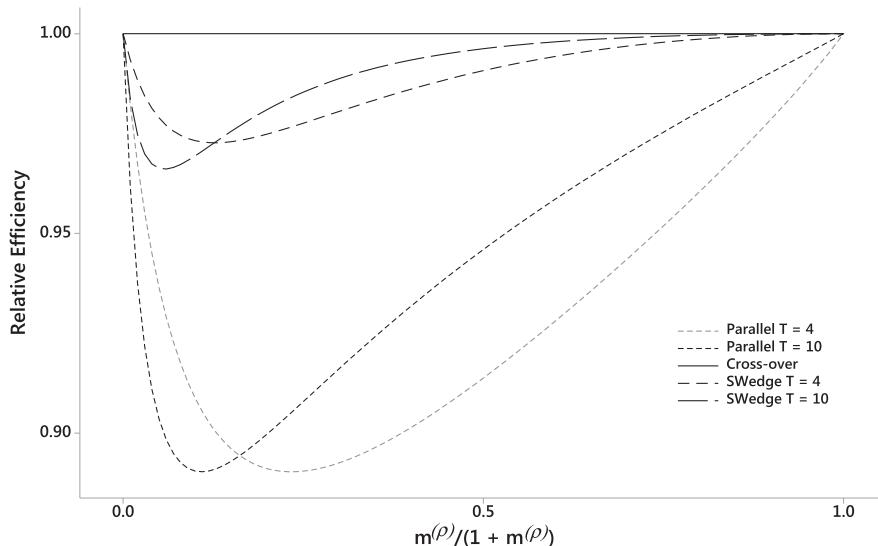
### 3.2 | General layouts, including stepped wedge and delay-control designs

For a layout with both  $A > 0$  and  $B > 0$  the RE in (2) is a weighted average of  $\text{RE}_{\text{Parallel}}$  and  $\text{RE}_{\text{Crossover}}$  with weights in the ratio  $A:Bv$ . As noted above,  $\text{RE}_{\text{Crossover}}$  is itself the RE for a parallel design under a modified correlation structure. In general, the weights depend on the experimental layout (through  $A$  and  $B$ ), the mean cluster size ( $m$ ), and the correlations  $\rho$ ,  $\pi$ , and  $\tau$  but not on the distribution of the relative cluster sizes  $z_k$ . The CV of the cluster size distribution and the shape of that distribution influence the RE only through their impact on the form of the function  $\Psi(\cdot)$ .

In the absence of cluster-specific time effects (ie,  $\pi = 1$ ,  $\Psi(\lambda_0 m^{(\rho)}) = 1$ ), the RE cannot be less than that under a parallel layout. This no longer holds when  $\pi < 1$ . Nevertheless, for any layout with a specified distribution of relative cluster sizes, the RE is bounded below by the (unique) minimum value of  $\Psi(\alpha)$  over the range  $0 < \alpha \leq \infty$ . This bound is also the least RE attainable under any parallel layout. In this sense, the loss of efficiency can be no worse than the worst-case associated with a parallel group design.

For a stepped wedge design with  $g = T - 1$  steps, the RE depends on the number of measurement periods,  $T$ . However, as  $T$  increases,  $v \rightarrow 0$ ,  $\frac{A}{B} = 1 + \frac{2}{T-1} \rightarrow 1$  and the ratio of weights  $A:Bv \rightarrow \infty$ . It follows that the RE for a stepped wedge design approaches that of the crossover design when  $T$  is large.

Following the work of Girling and Hemming,<sup>10</sup> we use the term delay-control design (DCD<sub>p,q,r</sub>) for a design in which a parallel study runs for a proportion  $q$  of the study duration, preceded by a baseline control period (proportion  $p$ ), and followed by a postimplementation period (proportion  $r$ ) in which the treatment is present in all clusters ( $p + q + r = 1$ ). Instances include parallel designs with baseline observations<sup>6</sup> for which  $r = 0$ , and “delayed start designs”<sup>11</sup> for which  $p = 0$ . In any case, the coefficients  $A$  and  $B$  depend only on the proportion of time ( $q$ ) spent under the parallel regime (see Table 1) with  $\frac{A}{B} = q^{-1} - 1$ . If  $q$  is fixed, this implies that the RE again approaches the crossover result as  $T$  increases, although less rapidly than for a stepped wedge design.



**FIGURE 3** Relative efficiency (of unequal clusters) for a crossover layout (solid line) and some parallel (dotted lines) and stepped wedge (broken lines) layouts for the Hussey and Hughes model ( $\pi = 1$ ,  $\tau = 0$ ) under a Gamma distribution for cluster size with  $CV^2 = 0.5$ . The horizontal axis represents a transformed version of the average cluster size ( $m$ ) and ICC ( $\rho$ ), as described in the text

### 3.3 | Numerical results: $CV^2 = 0.5$

We present some detailed results for cluster size distributions with  $CV^2 = 0.5$ , a value at the higher end of those commonly encountered in cluster studies.<sup>12</sup>

#### 3.3.1 | The Hussey and Hughes model ( $\pi = 1$ , $\tau = 0$ )

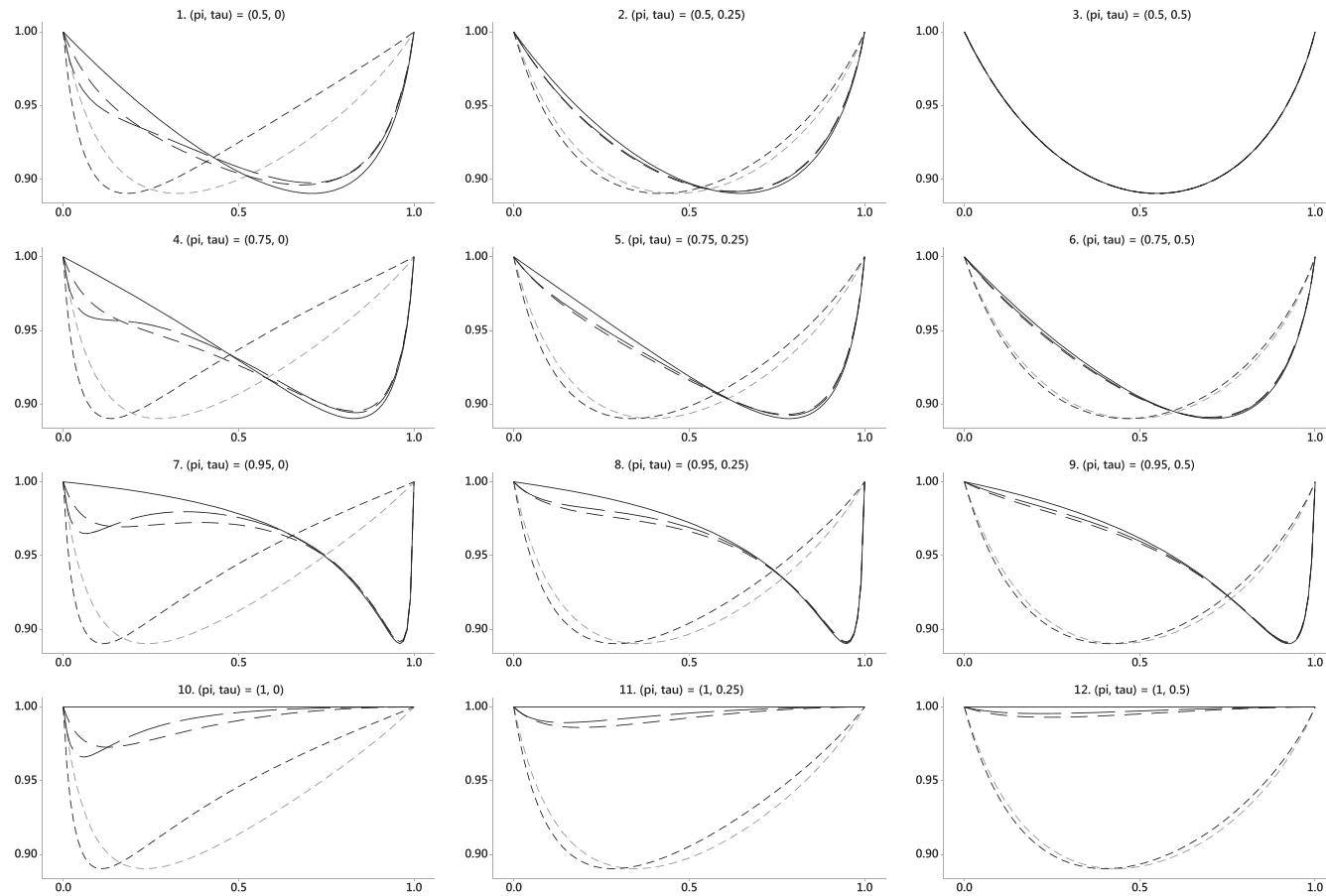
Under this model,  $\lambda_0 = 0$ ,  $\lambda_1 = T$ ,  $v = \frac{1}{1+Tm^{(\rho)}}$  and the RE can be written as  $\frac{A+Bv\Psi(v^{-1}-1)}{A+Bv}$ . Here, the parallel design (with  $A = 0$ ) has the worst (ie, least) RE among all complete layouts; the best RE ( $\equiv 1$ ) is achieved by the crossover design under which there is no loss of efficiency at all. The behavior of the RE as a function of  $m^{(\rho)}/(1 + m^{(\rho)})$  for crossover, parallel, and stepped wedge layouts (with  $T = 4$  and  $10$ ) is shown in Figure 3, for Gamma-distributed cluster sizes with  $CV^2 = 0.5$ . In general, the efficiency loss for a stepped wedge design is much less than that for the corresponding parallel design.

#### 3.3.2 | General correlation structure

Results for some representative values of  $\pi$  ( $= 1, 0.95, 0.75, 0.50$ ) and  $\tau$  ( $= 0, 0.25, 0.5$ ) are shown in Figure 4, under a Gamma distribution for cluster size with  $CV^2 = 0.5$ . In panel 3 of Figure 4,  $\pi = \tau$  and the RE is the same for all layouts and independent of the actual value of  $\pi$  (and  $\tau$ ). In the remaining panels, the autocorrelation parameters have been chosen so that  $0 \leq \tau < \pi \leq 1$ . A value of  $\pi$  close to 1 corresponds to slowly varying cluster effects consistent with a small perturbation from traditional ICC models. In addition, the value of  $\tau$  is attenuated by the inclusion of measurement error in the subject x time variance component,  $\eta_{ST}$ . Nonetheless, the assumption that  $\tau \leq \pi$  may not be appropriate in settings where the cluster effects are highly volatile.

The results are based on model (1), which applies explicitly to a longitudinal study where the same subjects are observed at every observation time. By reducing the subject variance  $\eta_S$  to 0, a model for cross-sectional studies is obtained in which effectively different subjects are measured at each timepoint. Thus, REs for cross-sectional studies appear as special cases by setting  $\tau = 0$  in (2), giving rise to the results in the first column of Figure 4 (panels 1, 4, 7, 10).

In general, the RE for any layout is bounded below by the minimum value of the function  $\Psi(\alpha)$ , which is  $\approx 0.89$  under the Gamma distribution considered here. Its actual value must lie between the RE for a cross-sectional layout and that for a parallel layout with the same number of observation times  $T$ . Although the function  $\Psi(\alpha)$  has a unique minimum, this property does not necessarily extend to the behavior of the RE as a function of  $m^{(\rho)}$ , as in panel 7 where a pair of minima is visible for the stepped wedge layout with  $T = 10$ . Nevertheless, when  $T$  is large, the RE for a stepped wedge layout approximates to that of the crossover design with the same correlation parameters. Once again, if  $\pi = 1$ , the loss of



**FIGURE 4** Relative efficiency and a transformed version of average cluster size for a crossover layout (solid line) and some parallel (dotted lines) and stepped wedge (broken lines) layouts for  $\pi = (0.5, 0.75, 0.95, 1.0)$  and  $\tau = (0, 0.25, 0.5)$ . The axes and line types are the same as those in Figure 3, which is reproduced for completeness in panel 10 here. Panels in the first column (1, 4, 7, 10) correspond to cross-sectional studies with individual autocorrelation = 0. The cluster size distribution is Gamma with  $CV^2 = 0.5$

efficiency for a stepped wedge design is necessarily less than for the corresponding parallel design (panels 10, 11, 12), but this is not so if temporal variation within clusters is present ( $\pi < 1$ , panels 1 to 9).

Corresponding results for parallel designs with baseline observations are presented in an online appendix in Supp Figure 1 (for  $T = 4$ ) and Supp Figure 2 ( $T = 10$ ). Two exemplars of such designs are shown, namely, where the baseline runs for exactly half of the study period, (ie,  $DCD_{0.5,0.5,0}$ ) as in the work of Teerenstra et al<sup>6</sup> and a design with a single baseline time epoch ( $DCD_{1/T,1-1/T,0}$ ).

#### 4 | PRACTICAL CONSEQUENCES

If the cluster sizes are known, Equations (2) and (4) can be used directly to obtain an explicit form for the DE that generalizes the known result (eg, equation 15 in the work of Rutherford et al<sup>13</sup>) for single-period parallel designs.

Where only the mean and standard deviation of the cluster sizes are known, several methods have been proposed to approximate the unequal cluster DE for single-period parallel studies.<sup>3,14</sup> Because each method entails a particular approximation to the function  $\Psi(\alpha)$  in equation (6) they can all be extended to any of the designs considered here.

The method due to van Breukelen et al<sup>4</sup> uses a Taylor approximation of the form

$$\Psi(\alpha) \approx \Psi_{BC}(\alpha) = 1 - \frac{\alpha}{(1 + \alpha)^2} c^2,$$

(where  $c = CV$  of cluster size) which is illustrated in Figure 2. It works well over a wide range of parameter values particularly for the positively skewed distributions of cluster size likely to be encountered in practice. Since, in practice,  $c^2$

seldom exceeds 0.5, it has been suggested that the RE will be no less than  $\min_{\alpha} \Psi_{BC}(\alpha) = 88\%$  for most parallel studies.<sup>12</sup> The same remark is equally applicable in the general setting.

Another popular method<sup>13</sup> for single-period parallel studies replaces  $m$  with  $m(1 + c^2)$  in the equal cluster DE, ie,  $1 + (m - 1)\rho$ , to give  $DE = 1 + \{(1 + c^2)m - 1\}\rho$ . This amounts to using the LFD in place of the actual distribution of cluster sizes and entails the approximation

$$\Psi(\alpha) \approx \Psi_{LFD}(\alpha) = \frac{1 + \alpha}{1 + (1 + c^2)\alpha}.$$

The LFD method necessarily underestimates the RE of the unequal cluster design, but it can be useful for exploring the potential impact of unequal cluster sizes. For example, it shows that  $\Psi(\alpha) > 0.95$  for all  $\alpha$  whenever  $c < 0.23$ , giving an RE in (2) of at least 95%. Thus, the suggestion<sup>2</sup> that cluster inequality can be ignored in parallel studies unless the CV exceeds 0.23 is equally applicable to any design layout.

In practice, conservative sample size calculations based on the LFD can be made without recourse to detailed computation of relative efficiencies. For a given CV =  $c$ , the LFD entails a mixture of null clusters (of size 0) and nonnull clusters of size  $m(1 + c^2)$ , in the ratio  $c^2 : 1$ . Since the null clusters contribute nothing to the experimental precision, the precision under the LFD is equal to that of a design in which the observations are divided equally between the nonnull clusters, which would constitute a fraction  $1/(1 + c^2)$  of the clusters in the study. It follows that a conservative estimate for the total number of observations required under any layout can be made by applying equal cluster methods<sup>8</sup> with an assumed cluster size =  $m(1 + c^2)$ .

## 4.1 | Two examples

In the examples below, power is estimated from the standard formula  $\Phi(|\theta_0| \times (\text{Precision})^{1/2} - 1.96)$ , where  $\theta_0$  is the hypothesized value of the treatment effect and  $\Phi$  is the cumulative standard normal distribution function.

The EPOCH trial<sup>15</sup> aims to evaluate an intervention to improve mortality from emergency laparotomy in 90 hospitals. The design was based on six replicates of an SW<sub>15</sub> design with cross-sectional data collected in 16 consecutive five-week periods. The average number of subjects per period over all hospitals was assumed to be 18, with variation between hospitals characterized by a CV<sup>2</sup> of 0.5 (from a preliminary subsample of hospitals). The study was powered to detect a reduction in 30-day mortality from 25% to 22% (ie,  $\theta_0 = -3\%$ ). We illustrate the calculations using an approximate linear mixed model for binomial data<sup>9</sup> with ICC  $\rho = 0.0075$ ,  $\pi = 1$  and  $\tau = 0$ . (The details differ slightly from the published protocol, which incorporates allowance for transition periods between treatment conditions and an ad hoc adjustment for geographical clustering of hospitals.) For SW<sub>15</sub>, the design coefficients from Table 1 are  $A = 0.0826$  and  $B = 0.0729$ . Hence, for equal clusters of size  $m = 18$ , we have  $v = 0.3148$  and  $DE_0 = 2.3508$  (from Equation (5)). The precision of the estimated change in percentage mortality in an individually randomized trial with the same total number ( $6 \times 15 \times 16 \times 18 = 25\,920$ ) of observations is 3.4560. Thus, the equal cluster precision is given by  $3.4560 \div 2.3508 = 1.4710$ , which translates to a power of 95.3%.

From Section 3.3.1, the RE due to unequal cluster sizes is

$$RE = \frac{A + Bv\Psi(v^{-1} - 1)}{A + Bv} = 0.783 + 0.217 \times \Psi(2.177).$$

Simulating a Gamma distribution of cluster sizes, we find  $\Psi(2.177) = 0.896$  ( $\pm 0.002$ ), and RE = 0.977; whereas,  $\Psi_{BC}(2.177) = 0.892$ , RE = 0.976. The resulting DE (=  $DE_0 \div RE$ ) is similar in both cases with power = 94.9%.

The LFD (“worst-case”) adjustment for CV<sup>2</sup> = 0.5 can be achieved by assuming equal clusters in a design with  $m = 18(1 + 0.5) = 27$  subjects per period in each of  $90 \div (1 + 0.5) = 60$  hospitals, ie, with the same total number of observations. To do this,  $v$  is recalculated:  $v = 0.2345$ , and the equal cluster DE = 2.4885, yielding a precision of  $3.4560 \div 2.4885 = 1.3888$ . The corresponding RE =  $2.3508 \div 2.4885 = 0.945$ , with power 94.2%.

Hooper et al<sup>8</sup> present calculations for a longitudinal SW<sub>3</sub> trial<sup>16</sup> incorporating subject and cluster autocorrelations under model (1). They assume  $\rho = 0.33$ ,  $\pi = 0.9$ ,  $\tau = 0.7$ ,  $\sigma = 5$ , with the study powered to detect an improvement in outcome of two units on a 10-point ordinal scale. With  $m = 10$  subjects per cluster, they conclude that a minimum of 93 subjects is required, which (after rounding up to whole numbers of clusters) translates to four clusters in each arm, ie, 120 subjects in total. The resulting precision of the effect estimate is 2.5673, giving a power of 89.3%. In the actual trial, between 8 and 12 subjects were recruited in each cluster, consistent with a cluster size CV of the order of 0.1. Here, allowance for unequal

cluster size would have had a negligible effect on precision and power. The “worst-case” is computed by assuming  $10 \times (1 + 0.1^2) = 10.1$  subjects per cluster in an SW<sub>3</sub> study with  $4 \div (1 + 0.1^2) = 3.96$  clusters per arm. This has precision of 2.5512 and estimated power of 89.1%.

## 5 | CONCLUSIONS AND LIMITATIONS

We have obtained the RE for a design with unequal cluster sizes under an exchangeable correlation structure that encompasses both cross-sectional and longitudinal studies, with and without cluster-level time effects. The RE can be used to adjust existing DEs for stepped wedge and cluster crossover studies to take account of cluster inequality. In general, the RE must lie between the RE for a parallel layout and the RE for a crossover layout with the same set of observation times. In the absence of cluster-level time effects, as in the Hussey and Hughes model, there is no loss of efficiency in a crossover study, implying that the RE for an arbitrary layout is never less than that for a parallel design. Even when cluster-level time effects are present, it remains true that the loss of efficiency under the worst-case configuration of parameters for any layout (eg, stepped wedge, crossover) is never more than that in the worst-case for a parallel design. A consequence is that results for parallel layouts that rely on worst-case scenarios will often carry over to more general designs. For example, the statements “when the coefficient of variation is <0.23, the effect of adjustment for variable cluster size on sample size is negligible”<sup>2</sup> and “the loss of efficiency due to variation in cluster sizes rarely exceeds 10%”<sup>4</sup> apply equally well to stepped wedge and other designs.

When the CV of cluster sizes is known, an upper bound on the unequal cluster DE (or sample size) can be obtained by inflating the average cluster size in the corresponding calculation with equal clusters. For parallel studies, this has been a popular method for dealing with clusters of unequal size, and it applies equally well for the more general designs considered here. It can be implemented using existing methodology for equal clusters.

We treat only rectangular (ie, “complete”) design layouts under linear mixed models with an exchangeable correlation structure for subject and cluster effects, but these are precisely the conditions under which explicit general formulae for the equal cluster DEs are known. Thus, our work complements these formulae<sup>8</sup> but does not consider more general auto-correlation structures. For example, the exponential correlation model discussed by Kasza et al<sup>7</sup> leads to unmanageable algebraic forms even for the equal cluster DEs. This would likely rule out a simple expression for the RE, although the impact of unequally sized clusters could be investigated numerically.

The RE computed here refers to the sampling behavior of an (asymptotically) efficient estimator of the treatment effect, such as would be obtained by maximum likelihood. For parallel studies, this is equivalent (asymptotically) to a cluster-level estimator using inverse-variance weighting of the cluster means. In practice, alternative analyses are sometimes specified based on equally weighted or size-weighted cluster means. When the clusters are equal in size, all three weightings are equivalent; but the equivalence no longer holds for unequally sized clusters, and the alternative estimators are technically inefficient. It follows that, for these estimators, the RE due to cluster inequality is necessarily less than that computed here. Indeed, this property has been used to derive lower bounds for the RE in parallel studies.<sup>4</sup> Some alternative estimators have been proposed for stepped wedge designs, eg, the vertical estimators of Matthews and Forbes,<sup>17</sup> but these are not efficient estimators even when the clusters are of equal size. Thus, the current work seems to have no direct implications for the impact of cluster inequality on the precision of such estimators.

This paper concerns the consequences of inequality between the clusters within the arms of a trial, not inequality between different arms, which, in principle, can be randomized away. Indeed, the formulae for the RE are strictly valid only for size-stratified randomization schemes under which the precision of the trial is the same irrespective of the outcome of the cluster randomization. This limitation is shared by most, but not all,<sup>18</sup> published adjustments to the DEs for parallel group designs where it is explicitly (or implicitly) assumed that the distribution of cluster sizes is identical in the two arms of the trial. Such adjustments are asymptotically valid (and commonly used) even under unrestricted randomization schemes if the number of clusters is large. A similar asymptotic argument applies for more general designs, including the stepped wedge, although it may be a less reliable guide to finite sample behavior when balance over several treatment sequences (arms) is required.

In practice, stepped wedge designs are sometimes used for very small numbers of clusters, perhaps with only one or two clusters allocated to each sequence. In such cases, the current analysis is incomplete: the RE given here provides a sensible average adjustment, but a full treatment of the problem may be more complex. The realized precision of the trial can vary with the particular outcome of the randomization. Further work is needed to explore the implications for trial design.

## ACKNOWLEDGEMENTS

Financial support was acknowledged from the National Institute for Health Research (NIHR) Collaborations for Leadership in Applied Health Research and Care for West Midlands (CLAHRC WM) and the HiSLAC study (NIHR Health Services and Delivery Research Programme ref 12/128/07). The views and opinions expressed are those of the author and do not necessarily reflect those of the NIHR.

## ORCID

Alan J. Girling  <http://orcid.org/0000-0003-2022-045X>

## REFERENCES

1. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. London, UK: Arnold; 2000.
2. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol*. 2006;35:1292-1300.
3. Kerry SM, Bland JM. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Statist Med*. 2001;20:377-390.
4. van Breukelen GJ, Candel MJ, Berger MP. Relative efficiency of unequal *versus* equal cluster sizes in cluster randomized and multicentre trials. *Statist Med*. 2007;26:2589-2603.
5. Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. *Statist Med*. 2008;27:5578-5585.
6. Teerenstra S, Eldridge S, Graff M, de Hoop E, Borm GF. A simple sample size formula for analysis of covariance in cluster randomized trials. *Statist Med*. 2012;31:2169-2178.
7. Kasza J, Hemming K, Hooper R, Matthews J, Forbes AB. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Stat Methods Med Res* 2017. <https://doi.org/10.1177/0962280217734981>
8. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Statist Med*. 2016;35:4718-4728.
9. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials*. 2007;28:182-191.
10. Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Statist Med*. 2016;35:2149-2166.
11. Zhan Z, de Bock GH, van den Heuvel ER. Statistical methods for unidirectional switch designs: past, present, and future. *Stat Methods Med Res* 2017. <https://doi.org/10.1177/0962280216689280>
12. van Breukelen GJ, Candel MJ. Comments on 'Efficiency loss because of varying cluster size in cluster randomized trials is smaller than literature suggests'. *Statist Med*. 2012;31:397-400.
13. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. *Int J Epidemiol*. 2015;44:1051-1067.
14. Manatunga AK, Hudgens MG, Chen S. Sample size estimation in cluster randomized studies with varying cluster size. *Biom J*. 2001;43:75-86.
15. Pearse R. Enhanced peri-operative care for high-risk patients (EPOCH) trial: a stepped wedge cluster randomised trial of a quality improvement intervention for patients undergoing emergency laparotomy. <http://www.epochtrial.org/docs/EPOCH%20Protocol%20revision%20V2%200%20280414.pdf>. Accessed January 30, 2018.
16. Tirlea L, Truby H, Haines TP. Investigation of the effectiveness of the "girls on the go!" program for building self-esteem in young women: trial protocol. *SpringerPlus*. 2013;2:683.
17. Matthews JNS, Forbes AB. Stepped wedge designs: insights from a design of experiments perspective. *Statist Med*. 2017;36:3772-3790.
18. Candel MJM, van Breukelen GJP. Repairing the efficiency loss due to varying cluster sizes in two-level two-armed randomized trials with heterogeneous clustering. *Statist Med*. 2016;35:2000-2015.

## SUPPORTING INFORMATION

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

**How to cite this article:** Girling AJ. Relative efficiency of unequal cluster sizes in stepped wedge and other trial designs under longitudinal or cross-sectional sampling. *Statistics in Medicine*. 2018;1-13.  
<https://doi.org/10.1002/sim.7943>

## APPENDIX A

## RELATIVE EFFICIENCY OF STRATIFIED UNEQUAL CLUSTER DESIGNS

An outline of the argument leading to Equation (2) is given here. The following notation is used: for any  $n > 0$ ,  $1_n$  is the  $(n \times 1)$  vector of 1s;  $I_n$  is the  $(n \times n)$  identity matrix;  $J_n = \frac{1}{n}1_n 1_n^T$ ; and  $\Pi_n = I_n - J_n$ . The matrices  $J_n$  and  $\Pi_n$  are mutually orthogonal projection matrices (ie,  $J^2 = J$ ,  $\Pi^2 = \Pi$ ,  $J\Pi = \Pi J = 0$ ). For any two matrices A and B,  $A \oplus B$  denotes the direct sum and  $A \otimes B$  the direct (Kronecker) product.

1. Since the subjects within a cluster are exchangeable, the BLUE for  $\theta$  in model (1) will depend only on the means of the observations in each cluster at each timepoint, ie, the “cell means,”  $\bar{Y}_{\bullet jkl}$ . The array of cell means may be vectorized as  $y = \left( \left( \text{vec} \bar{Y}_{\bullet j1l} \right)^T \middle| \left( \text{vec} \bar{Y}_{\bullet j2l} \right)^T \middle| \dots \left( \text{vec} \bar{Y}_{\bullet jrl} \right)^T \right)^T$ , a column vector consisting of  $r$  blocks, one for each stratum, with entries within blocks ordered by time within clusters. With this notation, model (1) may be written in the form

$$y = Z\beta + \theta x + \varepsilon. \quad (\text{A1})$$

Here,  $Z = 1_r \otimes 1_L \otimes I_T$ ;  $x = 1_r \otimes x_0$  with  $x_0 = \text{vec}(X^T)$ ;  $\beta$  is the  $(T \times 1)$  vector of time effects; and  $\varepsilon$  is a random  $(rLT \times 1)$  vector with zero mean and variance-covariance matrix of the form  $\mathbf{V} = V(m_1) \oplus V(m_2) \oplus \dots \oplus V(m_r)$ , where  $V(m_k)$  is the variance-covariance matrix of the  $T$  cell means from a cluster of size  $m_k$ , namely,

$$V(m_k) = \left( \eta_C + \frac{\eta_S}{m_k} \right) 1_T 1_T^T + \left( \eta_{CT} + \frac{\eta_{ST}}{m_k} \right) I_T.$$

2. From the theory of weighted least squares estimation, the precision of the BLUE of  $\theta$  in model (A1) is  $(\text{var } \hat{\theta})^{-1} = x^T P x$ , with

$$P = \mathbf{V}^{-1} - \mathbf{V}^{-1} Z (Z^T \mathbf{V}^{-1} Z)^{-1} Z^T \mathbf{V}^{-1}. \quad (\text{A2})$$

3. To compute the matrix  $P$ , we must find the inverse of  $\mathbf{V}$ . To do this, note that  $\mathbf{V}$  can be expressed as

$$\mathbf{V} = \text{Diag}(\zeta)^{-1} \otimes I_L \otimes \Pi_T + \text{Diag}(\xi)^{-1} \otimes I_L \otimes J_T,$$

where the  $(r \times 1)$  vectors  $\zeta$  and  $\xi$  have components  $\zeta_k = \zeta(m_k) = \left( \eta_{CT} + \frac{\eta_{ST}}{m_k} \right)^{-1}$  and  $\xi_k = \xi(m_k) = \left( T\eta_C + \frac{T\eta_S}{m_k} + \eta_{CT} + \frac{\eta_{ST}}{m_k} \right)^{-1}$ . It follows from the properties of  $J_T$  and  $\Pi_T$  that

$$\mathbf{V}^{-1} = \text{Diag}(\zeta) \otimes I_L \otimes \Pi_T + \text{Diag}(\xi) \otimes I_L \otimes J_T. \quad (\text{A3})$$

4. From (A3),

$$Z^T \mathbf{V}^{-1} Z = (1_r^T \text{Diag}(\zeta) 1_r) \cdot (1_L^T 1_L) \cdot \Pi_T + (1_r^T \text{Diag}(\zeta) 1_r) \cdot (1_L^T 1_L) \cdot J_T = rL \left( \bar{\zeta} \Pi_T + \bar{\xi} J_T \right),$$

with  $\bar{\zeta} = \frac{1}{r} \sum_1^r \zeta_k$  and  $\bar{\xi} = \frac{1}{r} \sum_1^r \xi_k$ . Moreover,  $(Z^T \mathbf{V}^{-1} Z)^{-1} = \frac{1}{rL} \left( \bar{\zeta}^{-1} \Pi_T + \bar{\xi}^{-1} J_T \right)$ . On substituting this inverse into (A2) and using (A3), we have

$$P = \text{Diag}(\zeta) \otimes I_L \otimes \Pi_T - \frac{1}{rL\bar{\zeta}} \zeta \zeta^T \otimes 1_L 1_L^T \otimes \Pi_T + \text{Diag}(\xi) \otimes I_L \otimes J_T - \frac{1}{rL\bar{\xi}} \xi \xi^T \otimes 1_L 1_L^T \otimes J_T.$$

5. From the fact that  $1_r^T \text{Diag}(\omega) 1_r = 1_r^T \omega = \frac{1}{1_r^T \omega} 1_r^T \omega \omega^T 1_r$  for any  $(r \times 1)$  vector,  $\omega$  ( $= \zeta$  or  $\xi$  here), it follows that

$$x^T P x = (1_r^T \otimes x_0^T) P (1_r \otimes x_0) = r x_0^T (\Pi_L \otimes \Pi_T) x_0 \times \bar{\zeta} + r x_0^T (\Pi_L \otimes J_T) x_0 \times \bar{\xi} = rLT \left( A \bar{\zeta} + B \bar{\xi} \right), \quad (\text{A4})$$

where  $A$  [ $= (LT)^{-1} x_0^T (\Pi_L \otimes \Pi_T) x_0$ ] and  $B$  [ $= (LT)^{-1} x_0^T (\Pi_L \otimes J_T) x_0$ ] are the coefficients that characterize the design layout, as defined in the main text of the paper (see Table 1).

6. The expression (A4) gives the precision of the  $\theta$ -estimate for a (stratified) design in which  $m_k$  subjects are observed in the clusters in stratum  $k$ . In the equal cluster case,  $m_k = m$  for each  $k$  so that  $\bar{\zeta}$  and  $\bar{\xi}$  are replaced by  $\zeta(m)$  and  $\xi(m)$ , respectively. Thus, the relative efficiency (unequal vs equal clusters) is

$$\text{RE} = \frac{rLT(A\bar{\zeta} + B\bar{\xi})}{rLT(A\zeta(m) + B\xi(m))} = \frac{A\frac{\bar{\zeta}}{\zeta(m)} + B\frac{\bar{\xi}(m)}{\zeta(m)} \cdot \frac{\bar{\xi}}{\xi(m)}}{A + B\frac{\xi(m)}{\zeta(m)}}. \quad (\text{A5})$$

Some elementary algebra shows that (A5) is equivalent to Equation (2), ie,  $\frac{\bar{\zeta}}{\zeta(m)} = \Psi(\lambda_0 m^{(p)})$ ,  $\frac{\bar{\xi}}{\xi(m)} = \Psi(\lambda_1 m^{(p)})$ , and  $\frac{\xi(m)}{\zeta(m)} = v$ .

7. Using (A4), the precision in the equal cluster case can be expressed as

$$rLT(A\zeta(m) + B\xi(m)) = \frac{rLTm}{4\sigma^2} \cdot \frac{4[1 + (T - 1)v](A + Bv)}{Tv(1 + (m - 1)\rho)}. \quad (\text{A6})$$

The first factor on the right-hand side is the precision for a randomized controlled trial with the same total number of observations, ( $= rLTm$ ). Thus, the second factor is the inverse of the equal cluster DE given in (5).

## APPENDIX B

### PROPERTIES OF $\Psi(\alpha)$

In this appendix, we use a more general definition of the function  $\Psi(\alpha)$  than in Equation (3), in terms of an expected value over a general probability distribution. Namely,

$$\Psi(\alpha) = E\left[\frac{(1+\alpha)Z}{1+\alpha Z}\right], \text{ where } Z \text{ is a nonnegative random variable with } E(Z) = 1.$$

1.  $\Psi(0) = 1$ , from the definition.
2.  $\Psi(\infty) = \Pr\{Z > 0\}$ . This follows because  $\Psi(\infty) = \lim_{\alpha \rightarrow \infty} E\left[\frac{(1+\alpha)Z}{1+\alpha Z}\right] = E\left[\lim_{\alpha \rightarrow \infty} \frac{(1+\alpha)Z}{1+\alpha Z}\right]$ .
3.  $\Psi(\alpha) \leq 1$ , from Jensen's inequality since  $\frac{(1+\alpha)Z}{1+\alpha Z}$  is a convex function of  $Z$ .
4. Either  $\Psi(\alpha)$  is constant ( $= 1$ ) or it has a unique minimum in the range  $0 < \alpha \leq \infty$ . Putting  $u = \frac{\alpha}{1+\alpha}$ , we have  $\frac{\partial^2}{\partial u^2} \Psi(\alpha) = E\left[\frac{Z(Z-1)^2}{(1-u+uZ)^2}\right] \geq 0$  so that  $\Psi(\alpha)$  is convex when considered as a function of  $u$ . The result follows because  $u$  is strictly increasing in  $\alpha$ .
5. (Least Favorable Distribution) Suppose that  $Z$  has discrete support  $S$  with probability distribution  $\{p_z; z \in S\}$ . Let  $\{q_z; z \in S\}$  be the “size-biased” distribution defined by  $q_z = zp_z$  for  $z \in S$ . This is a proper distribution because  $\sum_{z \in S} q_z = \sum_{z \in S} zp_z = E(Z) = 1$ . Now, apply Jensen's inequality to the convex function (of  $Z$ ):  $\frac{1+\alpha}{1+\alpha Z}$ , over the size-biased distribution  $\{q_z\}$ . This yields  $E_q\left[\frac{1+\alpha}{1+\alpha Z}\right] \geq \frac{1+\alpha}{1+\alpha E_q(Z)}$ , where  $E_q$  denotes an expectation over the  $q$ -distribution, ie,  $E_q f(Z) = E[Zf(Z)]$  for any function  $f()$ .

Hence,  $\Psi(\alpha) = E\left[\frac{(1+\alpha)Z}{1+\alpha Z}\right] = E_q\left[\frac{1+\alpha}{1+\alpha Z}\right] \geq \frac{1+\alpha}{1+\alpha E(Z^2)} = \frac{1+\alpha}{1+\alpha(1+c^2)}$ , where  $c$  is the CV of  $Z$ . It is straightforward to show that the bound is attained under the LFD defined in Section 3.1.