International Journal of Epidemiology, 2020, 979–995
doi: 10.1093/ije/dyz237
Advance Access Publication Date: 22 February 2020
Original article

Methods

A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator

Karla Hemming [1]* Jessica Kasza [2] Richard Hooper,[3] Andrew Forbes[2] and Monica Taljaard[4,5]

[1]Institute of Applied Health Research, University of Birmingham, Birmingham, UK, [2]Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, VIC, Australia, [3]Pragmatic Clinical Trials Unit, Centre for Primary Care and Public Health, Queen Mary University of London, London, UK, [4]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada and [5]School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada

*Corresponding author. Institute of Applied Health Research, University of Birmingham, Birmingham B15 2TT, UK. E-mail: k.hemming@bham.ac.uk

Editorial decision 10 October 2019; Accepted 11 November 2019

Abstract

It has long been recognized that sample size calculations for cluster randomized trials require consideration of the correlation between multiple observations within the same cluster. When measurements are taken at anything other than a single point in time, these correlations depend not only on the cluster but also on the time separation between measurements and additionally, on whether different participants (cross-sectional designs) or the same participants (cohort designs) are repeatedly measured. This is particularly relevant in trials with multiple periods of measurement, such as the cluster cross-over and stepped-wedge designs, but also to some degree in parallel designs. Several papers describing sample size methodology for these designs have been published, but this methodology might not be accessible to all researchers. In this article we provide a tutorial on sample size calculation for cluster randomized designs with particular emphasis on designs with multiple periods of measurement and provide a web-based tool, the Shiny CRT Calculator, to allow researchers to easily conduct these sample size calculations. We consider both cross-sectional and cohort designs and allow for a variety of assumed within-cluster correlation structures. We consider cluster heterogeneity in treatment effects (for designs where treatment is crossed with cluster), as well as individually randomized group-treatment trials with differential clustering between arms, for example designs where clustering arises from interventions being delivered in groups. The calculator will compute power or precision, as a function of cluster size or number of clusters, for a wide variety of designs and correlation structures. We illustrate

the methodology and the flexibility of the Shiny CRT Calculator using a range of examples.

Key Messages

- Cluster randomized trials are increasingly being designed with variations from the conventional two-arm design, with many including multiple periods of measurement.
- Alongside this rapid increase in the use of the designs there has also been a rapid development in the methodology of sample size calculations for these trials.
- In cluster randomized trials with multiple time periods, such as the multiple-period cluster cross-over and the stepped-wedge designs, an exchangeability correlation structure is unlikely to be appropriate.
- This paper provides a tutorial on sample size calculations for cluster randomized trials, with particular emphasis on designs with multiple periods of measurement with allowance for time-dependent correlation structures, and introduces an online calculator for practical implementation.

Introduction

In this tutorial we provide an overview of different types of cluster randomized trial designs, with particular emphasis on designs with multiple periods. We outline a number of different within-cluster correlation structures that are either commonly used or plausible structures for these designs. We outline how empirical estimates for these within-cluster correlations might be obtained. We then summarize what is known about the statistical efficiency of these competing designs, before introducing the Shiny CRT Calculator to allow users to not only determine required sample sizes in their own settings, but also to compare statistical efficiency for different designs on a case-by-case basis. We include designs with differential clustering across arms (for example group therapy trials) and consider designs in which treatment effect heterogeneity across clusters is expected.

Our objectives are to help researchers navigate the literature on sample size calculations for multiple-period cluster randomized trials. By additionally providing an easy-to-use online calculator that nonetheless allows for much complexity, our intention is to promote more rigorously designed and powered cluster randomized trials.

We present several examples to illustrate this methodology and the Shiny CRT Calculator. Our objectives are to illustrate how researchers can: (i) determine sample size or power for a range of cluster randomized designs; (ii) identify cluster sizes under which all observations make a non-negligible contribution to the power; (iii) compare power achievable across different designs; (iv) incorporate more comprehensive correlation structures than the simple exchangeable structure; and (v) explore sensitivity to key correlation parameters, which are often estimated with uncertainty. We also illustrate how to modify these calculations when the study will evaluate interactions with treatment effect and cluster; and we illustrate how to identify optimal designs (with respect to minimizing the total sample size) for individually randomized arms with clustering in one arm only. The examples are based on real studies, but results may differ from those studies as some details may have been changed for illustration.

Overview of different types of cluster randomized trial designs

In the two-arm parallel cluster randomized trial (CRT), clusters (for example, wards, hospitals, communities) are randomly allocated to one of two 'arms' that would typically represent an intervention condition and a control condition, see Figure 1a. [1–3] Because observations from participants in the same cluster are usually correlated, the design is statistically less efficient than a comparable individually randomized trial. That is, for an equivalent number of observations, a cluster randomized trial will usually provide less power to detect the target effect size. One common variation on the standard parallel-arm design with a single measurement is to take a measurement at the start of the study (a baseline assessment, see Figure 1b) in addition to a follow-up assessment, which we refer to as a CRT with a baseline measure.[4] Baseline and follow-up measurements might be taken on the same participants (a cohort design) or on different participants (a cross-sectional design). This design can be statistically more efficient than a two-arm parallel CRT, depending on the

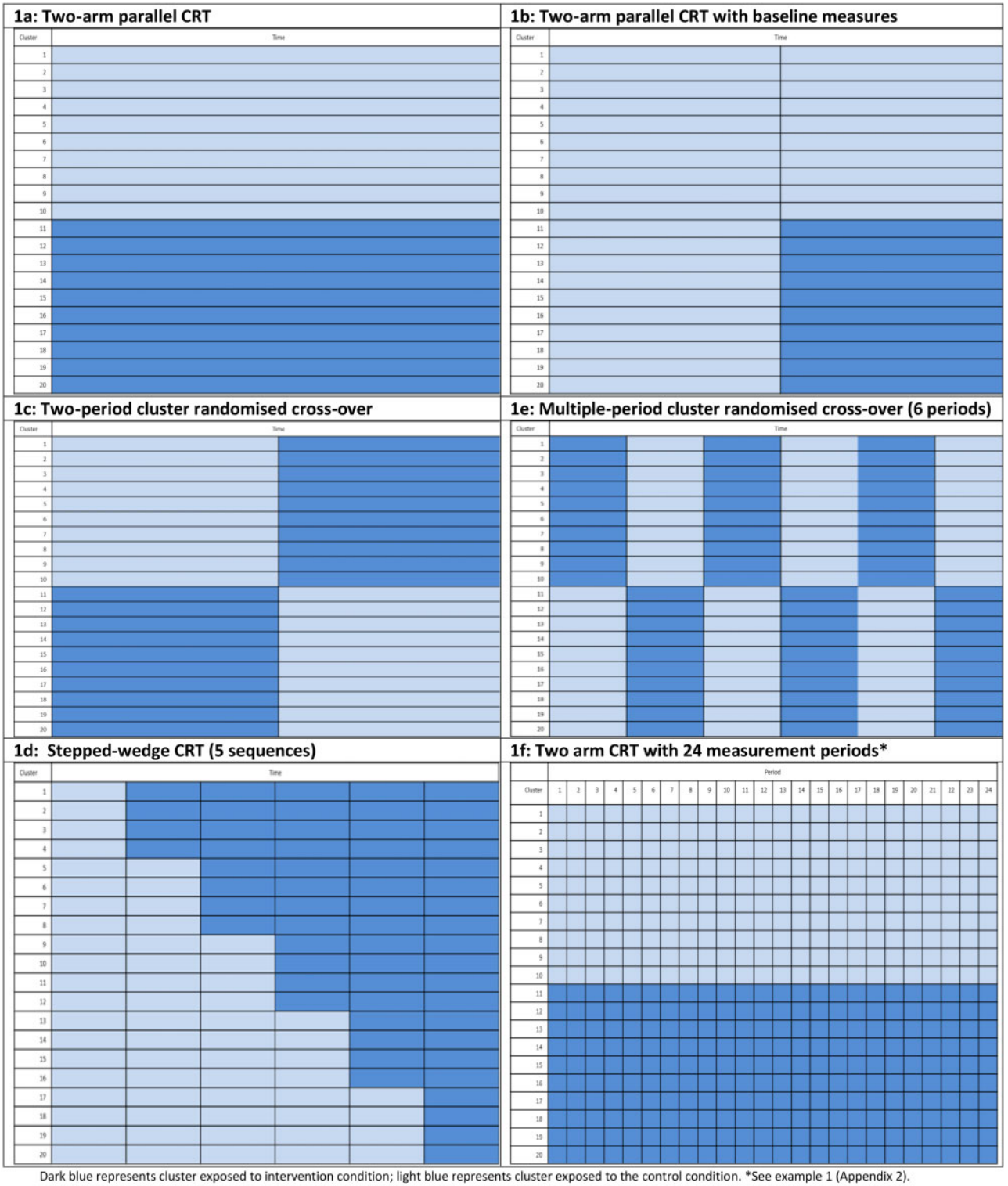

Figure 1. Schematic representation of different multi-period cluster randomsied trials.

strength of the correlation between baseline and follow-up measurements.[5]

The two-period cluster cross-over design (CRXO), which is more statistically efficient than the CRT with a baseline measure, randomly allocates clusters to one of two different sequences: either the control condition followed by the intervention condition; or the intervention condition followed by the control condition, see

Figure 1c.[6,7] Although only feasible in some settings (e.g. where it is possible to remove a cluster-level intervention), the CRXO design is a very appealing design as it allows mitigation, at least in part, of some of the power loss due to cluster randomization.[8,9] The multi-period CRXO design is an extension of the two-period CRXO design, with multiple-cross overs within each sequence, see Figure 1d. [10]

Other types of cluster randomized designs have seen increasing interest in recent years. The stepped-wedge cluster randomized trial (SW-CRT), for example, allocates clusters to sequences that involve periods in the control condition followed by periods in the intervention condition.[11,12] This design often starts off with all clusters in the control condition; clusters are randomly allocated to cross over to the intervention condition, at different times, typically until all are exposed to the intervention condition (Figure 1e). Both the CRXO and the SW-CRT can be designed either to take repeated measurements on the same participants at each measurement occasion (closed cohort design); or on different participants at each measurement occasion (cross-sectional design).[12,13] Within the class of stepped-wedge designs, variants may include designs in which not all clusters are observed in all periods of the trial. This 'incompleteness' in the design may be for practical reasons (for example to allow implementation or transition periods[11]) or may be to improve efficiency (for example with the use of a dog-leg design[14]).

Treatment effects and time adjustment

In this tutorial we consider how to determine sample size or power for these study designs under the assumption of a time-averaged treatment effect and assume adjustment for fixed time effects.[15] This is somewhat different to inferences concerning trends in treatment effects over time or the sustainability of treatment effects, which have been covered in texts on longitudinal cluster randomized trials.[16] For all outcomes the results assume large sample theory; and for binary and count outcomes we approximate variances of differences using standard formulas (Supplementary Appendix 1, available as Supplementary data at *IJE* online) and allow for over-dispersion for count outcomes. Of note, for count outcomes the implied standard deviation in the treatment arm will decrease with increasing target effect size; and any over-dispersion is assumed the same across treatment conditions.

Within-cluster correlation structures

Sample sizes for the conventional two-arm parallel CRT are typically derived by inflating the sample size needed under individual randomization by a design effect (DE). In its simplest form this design effect is [1+(m-1) ICC] where m is the cluster size (assumed constant across all clusters) and the ICC represents the intra-cluster correlation coefficient (ICC), that is the correlation between two randomly chosen observations in the same cluster.[17,18] This implies an exchangeable or compound symmetry correlation structure, whereby the correlation between any two observations in the same cluster is assumed to be identical.

Correlation structures become more complex when the measurements are taken at more than a single point in time. There are various ways of parameterizing these correlations. One common method, used in multiple-period designs, is to distinguish the within-period ICC (which measures the correlation between any two observations from the same cluster in the same measurement period) from the between-period ICC (which measures the correlation between any two observations from the same cluster but in different measurement periods).[14,19] An equivalent parameterization, the one we use here, is in terms of the within-period ICC and the cluster auto-correlation (CAC) where the CAC is the ratio of the between-period ICC to the within-period ICC.[20] A correlation structure with a within-period and between-period ICC (equivalently a common between-period CAC), referred to henceforth as a two-period correlation structure, is intuitive in studies where there are two periods, such as a two-period cross-over design or a parallel design with a baseline measurement, but may be less intuitive in a design with more than two periods. However, the two-period correlation structure has also been advocated in study designs with more than two-periods, such as the SW-CRT and the multiple-period cluster randomized cross-over design.[14] Under this parameterization of the within-cluster correlation structure, design effects have been published for the CRT with any number of before and after measures, the two-period CRXO, the multiple-period CRXO and the SW-CRT design, among others.[14] These design effects are a function of the cluster size per period (not to be confused with the total cluster size), the number of sequences (in the SW-CRT) and the correlation between repeated measurements on the same clusters and possibly the same participants (Table 1). All design effects assume an equal number of clusters allocated to each arm or sequence.[21]

For the multi-period CRXO, the SW-CRT design and any other design with multiple periods, correlation structures that allow for a within-period ICC and a common between-period CAC may not be realistic: the CAC may depend on the length of time between periods.[22] Intuitively, the between-period ICC for two temporally adjacent periods is generally expected to be larger than for two periods further apart in time. Thus, instead of one

Table 1. Design effects for various cluster trials

Design	Design effect for clustering (DE_c)	Design effect for repeated measures on same cluster (DE_R)	$var^{-1}(trt)$
Parallel CRT (s = 2; t = 1)	$[1+(m-1)\rho]$	1	$\frac{k*m}{2*DE_C*DE_R}$
Parallel CRT with baseline (s = 2; t = 2)	$[1+(m-1)\rho]$	$1-r^2$	$\frac{k*m}{2*DE_C*DE_R}$
Two-period cluster crossover (s = 2; t = 2)	$[1+(m-1)\rho]$	$\frac{1-r}{2}$	$\frac{k*m}{2*DE_C*DE_R}$
Stepped-wedge (s = w; t = w + 1)	$[1+(m-1)\rho]$	$\frac{3w(1-r)(1+wr)}{(w^2-1)(2+wr)}$	$\frac{w*k*m}{4*DE_C*DE_R}$
Multi-period cluster cross-over (s = 2; t = v)	$[1+(m-1)\rho]$	$\frac{(1-r)}{v}$	$\frac{k*m}{2*DE_C*DE_R}$
Differential clustering across arms (parallel two arms)(s = 2; t = 1)	$TSS*[\frac{[1+(m_c-1)\rho_c]}{m_c k_c}+\frac{[1+(m_t-1)\rho_t]}{m_t k_t}]$	1	$\frac{k*m}{2*DE_C*DE_R}$

Power is $\Phi^{-1}\left(-z_{1-\frac{\alpha}{2}}+\frac{ses}{\sqrt{var(trt)}}\right)$ where Φ is the cumulative distribution of the standard normal; s number sequences (or arms); t number time-periods; k number of clusters per sequence (arm); m cluster size per period. TSS is total sample size in study, thus s*t*k*m; $r=\frac{m\rho CAC}{DE_c}$ for cross-sectional designs and $r=\frac{m\rho CAC+(1-\rho)IAC}{DE_c}$ for closed cohort designs; ρ *and* α are within-period ICC and significance level (ICC: intra-cluster correlation; CAC: cluster auto-correlation); ses is standardized effect size (Supplementary Appendix 1, available as Supplementary data at *IJE* online). For designs with differential clustering, $m_c$is the cluster size in the control arm; m_t is the cluster size in the treatment arm; k_c is the number of clusters in the control arm; $k_t$is the number of clusters in the treatment arm; ρ_c and ρ_t represent the correlation within the control and treatment arms respectively; and $TSS = m_c k_c + m_t k_t$.

between-period ICC, there could potentially be one for each pair of periods within a cluster. Several possible extended correlation structures can be defined: the most general requiring only that the entire within-cluster correlation matrix be positive definite. The simplest of these correlation structures, and the one that we consider here, supposes that the CAC decays exponentially as a function of the distance between two periods. We call this a discrete time decay correlation structure. For example, the CAC for two periods 'j' periods apart in a study would be the CAC for temporally adjacent periods to the power of j. The power of studies for this discrete time decay correlation structure can be determined by numerically inverting matrices.[22] These more complicated correlation structures might also be of importance for parallel CRTs where the observations are measured at anything other than a single cross-section (see Example 1 below; and, Supplementary Appendix 2, available as Supplementary data at *IJE* online for an illustration of this).[22]

Finally, where the study involves repeated measurements on the same participants over time, the individual auto-correlation coefficient (IAC) measures the strength of the correlation between two observations on the same participant (irrespective of how far apart the measurements are taken).[23]

Empirical estimates for within-cluster correlations

To implement any sample size calculation for a cluster randomized trial, researchers need some a priori estimates for within-cluster correlations. When trial outcomes will be obtained from routinely collected data and historical data are available at the design stage, then tailored estimates may be obtained, for example by fitting general linear mixed models to the data. To match assumptions made under the derivation of the design effects outlined above (i.e. two-period correlation structure), linear mixed models with fixed period effects, random cluster effects and random cluster by period effects can be fitted to provide estimates of both the within-period and the between-period ICCs. For binary outcomes, ICCs on the proportions scale are required for sample size calculations and so this means that linear mixed models should be fitted and not logistic models.[24] As there is still uncertainty on how to obtain estimates for between-period ICCs (or equivalently CACs) in the case of binary proportions, until further work has been done, we suggest fitting linear mixed models to the binary data, as has been suggested by others.[25] Estimates of correlation structures, which allow the CAC to decay exponentially as a function of time between periods, can be obtained by fitting models with fixed period effects and random cluster-by-period effects, where particular structures (e.g. discrete time decay) can be assumed for the correlation matrix of the cluster by period random effects. Currently SAS appears to be the only standard statistical software package that allows for specification of correlation structures at all levels of the cluster hierarchy.[22] More details on the estimation of these correlations are provided in Example 2 below.

When only small datasets are available, for example from pilot studies, it has been shown that the uncertainty associated with estimated correlations will often be so great as to render these estimates uninformative.[26] In such situations, recommendations are to use patterns observed

from empirical studies of correlations.[27,28] Important determinants of ICCs are known to be: outcome type (ICCs for process outcomes, such as an outcome which records whether or not a patient has their blood pressure measured, are typically higher than for clinical outcomes, such as a patient's blood pressure); setting (ICCs in secondary care are typically higher than ICCs in primary care); and possibly cluster size and prevalence (ICCs from smaller clusters and more prevalent outcomes tend to be higher).[27,29] There is much less published information available on determinants of between-period ICCs or CACs. Limited published information suggests values for CACs (in the case of a constant between-period ICC) might be anywhere between 0.3 and 0.9, with some indication that CACs might decrease the longer the duration of the study.[30,31] Given the uncertainty in these estimates, sensitivity analyses are particularly important to inform sample size and power calculations.

For multiple-period trials, like those we are considering here, correlation structures may depend not only on the definition of the cluster but also on the duration of the periods. This means that empirical correlations should be estimated not only from data sources representing similar types of clusters, but also similar period lengths. Likewise, when estimating correlation and decay structures, the time periods used in the estimation of these parameters should be of a similar duration as in the planned trial.

Incomplete designs

Sometimes designs might incorporate a 'transition period' to allow the intervention to become embedded into practice. In these periods, clusters can be considered neither fully exposed nor fully unexposed to the intervention. If observations from these periods are not intended to be used in the analysis, these transition periods should be allowed for in the power calculation.[11] There may be other reasons for incomplete designs in which not all cluster periods contribute data. For example, incomplete designs may deliberately include only cluster periods contributing more statistical information, to reduce the data collection burden.[14,32] Cluster periods that contribute the most statistical information, will depend on the assumed correlation structure but seem to be those immediately before and after the cluster switches to the intervention condition, and cluster-periods at the corners of the design.[33] Some trials have adopted staircase designs, where clusters are only measured in periods immediately before and after the treatment switch.[34] Closed-form expressions for design effects do not exist for many of these incomplete designs, and so the power of incomplete designs is determined using numerical methods for inverting matrices.[11] A related concept are designs in which there is an unequal number of clusters allocated to each sequence. These designs have non-uniform allocation ratios (see Example 2 for an illustration of this).

Statistical efficiency

Statistical efficiency of any given cluster randomized design can be considered from different perspectives. In its simplest form, the most statistically efficient design can be viewed as the design that achieves the greatest power for a fixed sample size. However, the total sample size in a cluster trial is a function of the number of clusters, the number of participants per cluster and the number of repeated measurements. Exactly what is meant by 'efficiency'—i.e. what is being optimized and under what constraints—will depend on the specific circumstances. For example, in situations where there are financial or ethical considerations associated with the inclusion of individual participants, there may be a desire to minimize the total number of participants. In other situations there may be high costs associated with the enrolment of clusters, and there may be desire to minimize the number of clusters. Alternatively, in situations where there is a high burden associated with data collection, there may be a desire to minimize the number of repeated measures per participant.

In a two-arm parallel CRT and assuming an exchangeable correlation structure as the cluster size increases (for a fixed number of clusters), the incremental increase in power starts to plateau.[35] Furthermore, cluster trials with a small number of clusters are at risk of low internal validity (due to chance imbalances) and low external validity (where the clusters are not representative of the wider population) and might run into complications in analysis when using analytical methods that mostly appeal to large sample theory.[36] Thus, even when restricting the design to a two-arm parallel CRT, researchers need to be aware of the trade-offs between increasing the number of clusters and increasing the cluster sizes: it will usually be preferable to increase the number of clusters where possible.[37]

If the research study allows the possibility of using designs in which clusters are observed under both the control and treatment conditions, such designs might offer greater statistical efficiency. Comparisons of efficiency across these designs are possible. The choice of the most efficient design will depend, ultimately, on the assumed correlation structure. Often this will thus depend on the correlation between cluster period means from the same cluster assessed in different periods. This correlation depends in turn on the cluster period size and other correlation parameters.[32] Comparisons of efficiency need to ensure that the within-period ICC and CAC (and any other

correlations) are transferable across designs. That is to say, a within-period ICC calculated on the basis of a 1-month time period should not be used in the sample size calculation where time is divided into 3-month periods. When making comparisons of efficiency across designs, keeping the cluster periods the same size will help ensure that a like-with-like comparison is being made. Where unequal numbers of clusters are allocated to sequences in a SW-CRT, then allocating more to the first and last sequences will generally maximize statistical power.[38]

Differential clustering across arms and treatment effect heterogeneity across clusters

In addition to clustering arising based on the unit of randomization, clustering can also arise based on how the intervention is delivered. It is possible that an intervention delivered in a group setting induces clustering in one arm only, referred to here as differential clustering across arms. For example, individuals might be randomized to group therapy or to usual care where the usual care arm does not receive any additional treatment and individuals remain independent. This differential clustering also needs to be allowed for in sample size calculations. For a simple parallel design with observations taken at a single cross-section, a variation on the design effect for clustering is available[39,40] (Table 1); see Example 3 below for an illustration.

Related to this, for trials in which treatment is crossed with cluster (e.g. cluster cross-over or stepped-wedge designs), it is possible to test for treatment heterogeneity across clusters, i.e. cluster-to-cluster variation in the treatment effect.[41] Varying treatment effects across clusters creates an implicit differentiation between the correlation of observations within controls and that in treatment clusters (hence the relationship to differential clustering).[42] No closed form design effects are currently available for this formulation, but again power can be determined using numerical methods for inverting matrices. Example 2 illustrates the notion of designing a study to detect treatment effect heterogeneity across clusters.

Small sample corrections

For two-arm parallel CRTs with a small number of clusters, it is conventional to use critical values from the t-distribution rather than the normal distribution in power calculations, with the number of degrees of freedom set to the number of clusters minus two.[43,44] This approach is based on the degrees of freedom that would be obtained from a cluster-level analysis. A similar approach can be taken in multiple-period designs, although exactly what values of degrees of freedom should be used is unclear. One approach is to use the degrees of freedom calculated as the number of cluster periods minus the number of time periods minus one. For a two-arm parallel CRT, this implies the use of the conventional degrees of freedom (total number of clusters minus two). Any use of the t-distribution as opposed to a normal approximation will always be more conservative, irrespective of the degrees of freedom; however, different choices for degrees of freedom might be more or less conservative. Research is still needed on the optimal choice of degrees of freedom.

Varying cluster sizes

Cluster randomized trials that have varying cluster sizes are, all other things being equal, less powerful than trials with no variation between cluster sizes. In a parallel CRT, under the assumption of stratified randomization to ensure balance across arms in the total sample sizes, this variation can be allowed for by a modification of the design effect. Under this assumption the average cluster period size (m) is replaced with m(1+CV^2) where CV is the coefficient of variation of cluster sizes. This is known to be a conservative approach for allowing for varying cluster sizes.[45–47] A modification of this design effect has been shown to be valid for multiple-period cluster trials, under a two-period correlation structure.[45] It is important to note that these design effects are only valid under the assumption of a stratified randomization scheme. This means that in an SW-CRT for example, with one cluster allocated to each sequence, this inflation might not be appropriate.[48] Research is still needed on how these design effects perform under the discrete time exponential decay correlation structure.

Implementing sample size calculations: introducing the Shiny CRT Calculator

The Shiny CRT Calculator is a web-based app in which we have implemented methodology to determine power for cluster trials, including the parallel design, the parallel design with one before and after measure, the two-period cross-over design, the multiple-period cross-over design and the stepped-wedge design. To maximize flexibility of the Shiny CRT Calculator, users can also upload their own designs using a csv file, which can contain missing cluster period cells such as transition periods. This flexibility enables hybrids between parallel and stepped-wedge and other efficient designs such as the dog-leg.[14,32] To verify the upload or to check that the design under calculation corresponds to the intended design, it can be viewed on a 'Diagram of design' tab. For parallel-arm

designs, users can opt to have clustering in one arm only ('differential clustering'). When the design has been uploaded via the 'upload own design', the user is able to select an option to allow for variation in treatment effects across clusters.

Sampling structures allowed are those in which the same participants are measured in each measurement period (closed cohort design) and those in which different participants are measured (cross-sectional design). The app can also accommodate a range of different correlation structures. This includes the conventional exchangeable correlation structure (single within-cluster ICC), structures with a different within- and between-period correlation but where the between-period correlation is constant (two-period correlation structure) and discrete time exponential decay correlation structures. When the user selects anything other than an exchangeable correlation structure, and the number of periods is not implicit in the design (i.e. a two-arm parallel CRT with a discrete time decay correlation), the user must also specify the number of periods (see Example 1 for explanation).

For correlation structures other than the exchangeable correlation structure, the user specifies the within-period ICC and the CAC (under a discrete time decay correlation structure, this CAC is assumed to have an exponential decay over time); for designs with repeated measures on the same participant, the user also specifies the IAC. For an exchangeable correlation structure, the user specifies a single ICC only. The power curve is plotted for a 'base-case' within-period ICC and CAC (where applicable), supplied by the user. To allow examination of sensitivity to the specified CAC, two additional power curves are produced automatically, corresponding to a lower and upper extreme for the CAC. These extremes are set at 80% and 120%, respectively (or 1, whichever is greater). The user can also specify ranges for upper and lower limits of the estimated within-period ICC, which results in a total of six curves for each plot. Users can remove any of these curves by double-clicking the corresponding key on the legend. The bounds for the discrete time decay are the same as for the CAC. For trials with differential clustering, the user is asked to specify the base-case ICC under the treatment condition and the base-case ICC under the control condition (usually expected to be zero). When the option 'variation in treatment effects' is selected, the user is requested to provide a standard deviation of the treatment effect across clusters (see Example 2 for explanation). In this case, the ICC represents the correlation under the control condition (making the somewhat simplified assumption of independence between the correlation in the control and intervention condition).

The relationship between the power, the ICC and the cluster period size is complex and whereas it may be the case that power for a design increases (or decreases) as the ICC increases (or decreases), this depends on the cluster period size. As a result, the displayed curves for lower ICC values will not always be lower than those curves for higher ICC values (and vice versa) and furthermore, displayed curves for lower and higher ICC values may cross each other. The relationship between CAC and power or precision is also complex when CAC decays and the bounds resulting from the lower and upper CAC values may not necessarily fall either side of that using the original CAC value.[22]

In addition to the usual sample size parameters, users must supply specific parameter values for each design. For the stepped-wedge design, users must specify the number of sequences and the number of clusters allocated to each sequence; and for the multiple-period cluster cross-over design, the user specifies the number of periods. The user specifies their desired significance level (for a two-sided test), which will allow calculations under superiority with multiplicity adjustments where desired. For complete designs, the user can explore the influence of varying cluster size by specifying the coefficient of variation of cluster period sizes (CV) (this option is not available when matrix inversion methods have been used, i.e. for the exponential decay correlation structure or when the user has uploaded their own design).

Outcome types incorporated are continuous, binary and count. The user has the option of either assuming the normal approximations or using a t-distribution with degrees of freedom equal to the number of cluster periods minus the number of time periods minus one. This will be of importance when the planned analysis will include a small sample correction.[49] The option of a t-distribution is not available when the user requests the plot to be cluster size versus number of clusters.

The user can plot the resulting power (or precision) as a function of the cluster size (for a fixed total number of clusters), the number of clusters (for a fixed total cluster size) or the number of clusters against cluster size (for a fixed power). Where the design includes differential clustering, the plot of the number of clusters against cluster size (for fixed power) is three-dimensional and shows total sample size, the number of individuals under the control condition and the number of clusters under the treatment condition (for fixed cluster sizes). The function allows the user to vary the ranges of the x-axis on the plot, so as to make the range more applicable to the scenario. Hovering over the curve will provide the user with power values for each point on the curve.

The app is programmed in R and implemented using the R Shiny application.[50] The app can be used via the following link: [https://clusterrcts.shinyapps.io/rshinyapp]. Source code and updates are located on GitHub [https://github.com/karlahemming/Cluster-RCT-Sample-Size-Calculator]. For complete designs or the more straightforward correlation structures, sample size calculations are implemented using design effects. For incomplete designs, designs with varying numbers of clusters per sequence or designs with more complex correlation structures, numerical methods are used to invert the variance covariance matrix and obtain the variance of the treatment effect estimator. Supplementary Tables S1–S4, available as Supplementary data at *IJE* online, validate the implemented code by comparing for a range of designs, estimating power with independent sources (summarized in Supplementary Appendix 4, available as Supplementary data at *IJE* online). Figure 2 illustrates the interface of the application. Users can hover over buttons to read a brief explanation of options. To reduce computational times, when the user has uploaded their own design, the curve is displayed after the user presses the button 'create curve'. Users can also download the corresponding data behind the resulting power or precision curve. The plot has various built-in options that can be assessed using the bar on the top right-hand side of the graph (hidden from view until the users hovers the mouse in that area). These options include toggle bars (to see the corresponding x- and y-values on the plot), an option to download a copy of the plot and options to re-scale the axis.

Example 1: Evaluating the effectiveness of a screening programme for group B streptococcus

Group B streptococcus (GBS) is a bacterium that is found in the birth canal of approximately 20% of pregnant women. In the UK, GBS screening is not routinely available and a UK funding body commissioned a call for proposals for a randomized trial to compare a GBS screening programme with current practice. As the question of interest is to evaluate a screening programme, a cluster randomized trial has been chosen. The primary outcome is the proportion of babies that develop septicaemia. Investigators agreed that a two-period CRXO design was feasible, provided the wash-out period was sufficiently long, although a cross-over design with multiple switches was not feasible. Financial costs for each cluster included were large and a maximum of 50 clusters was available within budgetary constraints. The study duration needed to be less than 2 years, which implies a maximum average cluster size of around 5000 (since the average number of births per hospital over a 2-year period is about 5000).

The trial is to have a superiority design, the outcome is binary, the hypothesis test is two-sided, the level of significance is set at 0.05 and 90% power is desirable. Various sources have been used to inform the specification of the

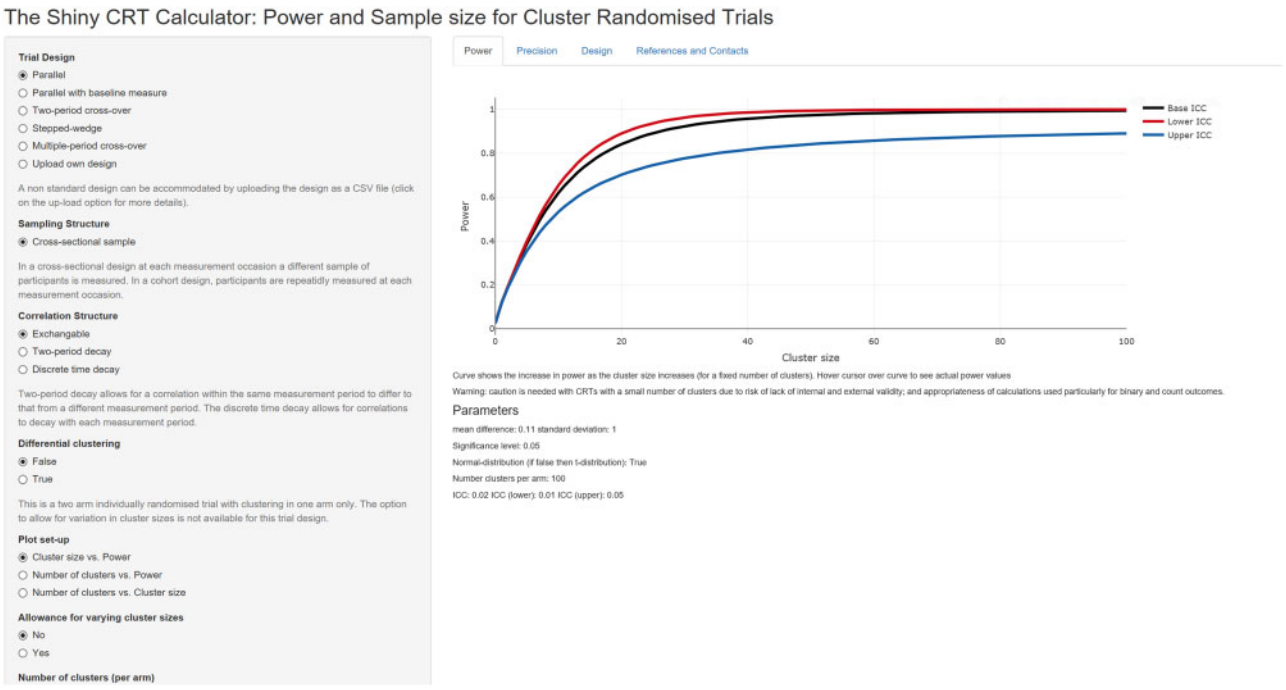

Figure 2. Illustration of the interface of the Shiny CRT Calculator.

prevalence of the outcome, which is anticipated to be 1% in the control condition. We use asymptotic tests here, since expected cluster period counts are greater than five, based on the cluster period sizes of 2500.[51] A relative risk reduction of 30% is considered plausible, informed by smaller individually randomized trials and meta-analyses, although not necessarily the minimal clinically important difference. There are no routinely collected data available from which to estimate correlation parameters, and the specified estimates were therefore informed by patterns observed in the literature: the outcome is a clinical outcome with low prevalence and from large clusters, suggesting that the ICC is likely to be very small. We therefore considered a base-case value of 0.005 with lower and upper ranges of 0.001 and 0.01, respectively. When considering the two-period CRXO design, a natural correlation structure is the two-period correlation structure, for which we additionally need to specify the CAC. There is little empirical evidence to inform likely values of the CAC, and we use a value suggested in the literature of 0.8[6,14] as our base case. The Shiny CRT Calculator automatically varied the CAC between 0.64 (80% of the base-case value) and 0.96 (120% of the base-case value).

Figure 3a illustrates that 90% power is not achievable under a two-arm parallel CRT, assuming an exchangeable correlation structure (with assumed within-cluster correlation ranging from 0.001 to 0.01) even when cluster sizes are 5000 (assuming the trial runs for 2 years); see Supplementary Appendix 2, available as Supplementary data at *IJE* online, where we extend this example under the parallel design for other correlation structures. Figure 3b illustrates how under the two-period CRXO design with these parameter values, 90% power is achievable under a wide range of conditions and shows increasingly negligible impact of increasing the cluster period size (note that here the cluster period sizes represent the expected number of observations over a 1-year period). For example, this example illustrates that 25 clusters per sequence (i.e. 50 clusters in total) with an average of 1000 births per cluster period achieves 90% power when the ICC is 0.005 and the CAC is 0.8. The sensitivity analysis shows that if the CAC were as low as 0.64, then 90% power would achievable with very large cluster sizes; whereas if the CAC were as high as 0.96, a cluster period size of less than 500 would be required.

Example 2: Evaluating the effectiveness of the introduction of a national pre-implantation biopsy histopathology service for kidney transplantation (PITHIA trial)

The PITHIA trial is a SW-CRT designed to determine if the introduction of a pre-implantation biopsy histopathology service increases and improves outcomes of kidney transplants.[52] The intervention will be evaluated using an SW-CRT; each measurement period is 4 months, with four clusters being randomly allocated to each of five sequences. The trial has a superiority design and has two co-primary outcomes: a binary outcome representing acceptance of transplant on first offer (called acceptance on first offer henceforth); and a continuous outcome-recipient estimated glomerular filtration rate (eGFR), measured at 1 year after transplant. Hypothesis tests are two-sided; the level of significance is set at 0.025 to allow for the two primary outcomes. The study has a fixed number of clusters and is needed to run for a fixed duration; thus, the design is fixed by what was deemed logistically possible. We initially consider invoking a two-period correlation structure and then relax this to accommodate a discrete time decay correlation structure.

In this example there were routinely collected data available (UK Transplant Registry, held by NHS Blood and Transplant) to estimate values of the outcomes under the control condition and correlation parameters. For the binary outcome of acceptance on first offer, the estimated prevalence was 28% and the average cluster size per period was 20. A linear mixed effects regression model was fitted to the extracted data (22 clusters, six cluster periods, average cluster period size 20), with a binary outcome (whether the offer resulted in a transplant), a fixed effect for period (4 months), a random effect for cluster and a random interaction between cluster and period (to model a two-period correlation structure (Supplementary Appendix 3, available as Supplementary data at *IJE* online). From this model, the within-period ICC was estimated as 0.025. We consider a lower limit of 0.01 and an upper limit of 0.06, which also are the lower and upper limits of the 95% confidence interval around the within-period ICC. The CAC was estimated as 0.92. For the CAC, the Shiny CRT Calculator automatically considers the lower range of 0.74 and the upper range of 1, i.e. 80% and 120% (rounded down to 1), respectively, of the specified value of 0.92. We specified a target absolute difference of 10% (i.e. prevalence 28% versus 38%). Figure 4a shows that under a two-period correlation structure, power in the region of 80% is achievable under most scenarios. For example, when the within-period ICC is 0.01 and the CAC is 0.92, then a cluster period size of 20 achieves 82% power.

For the continuous outcome (eGFR) we consider a standardized effect size (target mean difference divided by standard deviation) of 0.25. The average cluster size per period was 10. We used the same procedure to obtain estimates for the correlation parameters as described for the co-primary outcome above. STATA and SAS code for doing

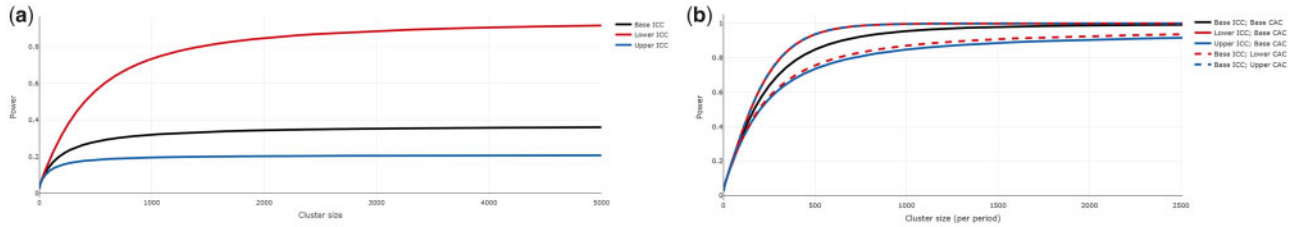

Figure 3. (a) Two-arm parallel CRT. Scenario includes 25 clusters per arm; proportion under control condition is 0.010 and proportion under intervention condition is 0.007; significance level is 0.05; within period ICC is 0.005 (lower value 0.001 and higher value 0.01). Expected cluster size (over 2 years) is 5000. (b) Two-period CRXO design. Scenario includes 25 clusters per sequence; proportion under control condition is 0.010 and proportion under intervention condition is 0.007; significance level is 0.05; within period ICC is 0.005 (lower value 0.001 and higher value 0.01); CAC is 0.8 (lower and higher values 80% and 120% of base case CAC, i.e. 0.64 and 0.96).Expected cluster-size per period (1 year) is 2500.

these calculations are presented in Supplementary Appendix 3, available as Supplementary data at *IJE* online. The within-period ICC was estimated as 0.056, and we consider the lower limit of 0.023 and upper limit of 0.13, which correspond to the lower and upper limits of the 95% confidence interval. The CAC was 0.08 with the lower range of 0.064 and upper range of 0.096, corresponding to 80% and 120%, respectively, of the base-case considered in the Shiny CRT Calculator. This CAC value is much lower than expected and so we also considered the implications of a substantially larger CAC of 0.8. Figure 4b shows very large variations in expected power (under an assumed two-period correlation structure)—highly dependent on the estimated within-period ICC. The power plateaus at a value close to 50% when the within-period ICC is 0.13 and not dependent on the CAC (although of note: the considered range for the CAC is narrow). If the CAC is closer to 0.8, then close to 80% power is achievable under all scenarios for a cluster period size in the region of 20 (Figure 4c).

Discrete time exponential decay correlation structure

We continue this example to illustrate the use of discrete time decay correlation structures for the binary outcome acceptance on first offer. SAS is used to fit a model to the binary outcome with an exponential decay structure for the between-period ICCs (code is provided in Supplementary Appendix 3, available as Supplementary data at *IJE* online). The within-period ICC was estimated to be 0.03 (we consider lower and upper bounds of 0.01 and 0.1, respectively); and the CAC for temporally adjacent periods was estimated to be 0.90 (exact value 0.89 but rounded up for illustration). This implies that the CAC for study periods two periods apart (e.g. periods 1 and 3, or periods 2 and 4) is $0.9^2 = 0.81$; for study periods three periods apart (e.g. 1 and 4), the CAC is $0.9^3 = 0.729$; for study periods four periods apart, the CAC is $0.9^4 = 0.6561$, etc. Assuming the same settings as above, Figure 4d indicates that the impact of incorporating discrete time correlation decay structure is to reduce power for the same cluster period size: with a decaying CAC of 0.90 with 20 participants per cluster period, the planned study has 78.6% power, compared with 82% power when the decay in CAC is not accounted for (Figure 4a). More generally, it is known that the relationship between power under a model with a discrete time decay and a model with a two-period correlation structure is complex, and depends on the study design, the CAC and other design parameters.[22]

Incorporating incomplete designs

We also illustrate in this example how the Shiny CRT Calculator can be used to incorporate incomplete designs (for the binary outcome acceptance on first offer). The focus here is allowing for a transition period, but the concept equally applies to other missing cluster periods. We continue with the PITHIA trial (an SW-CRT with five sequences and four clusters allocated to each sequence) but now designate the cluster period immediately after a switch to be a transition period with no data contribution (without increasing the number of periods, see Supplementary Figure S1, available as Supplementary data at *IJE* online). For simplicity we assume an exchangeable correlation structure. To obtain power and sample size for this design, the user must upload a CSV representation of the study and specify all other parameters as above (with blank cells to denote any missing cluster periods). Supplementary Table S2, available as Supplementary data at *IJE* online, shows that the power reduces to 59% as a result of the transition (the effect of the transition period on power could be decreased by adding an extra period at the end of the study).

Unequal number of clusters per sequence

We also illustrate in this example how the Shiny CRT Calculator can be used to explore the power implications

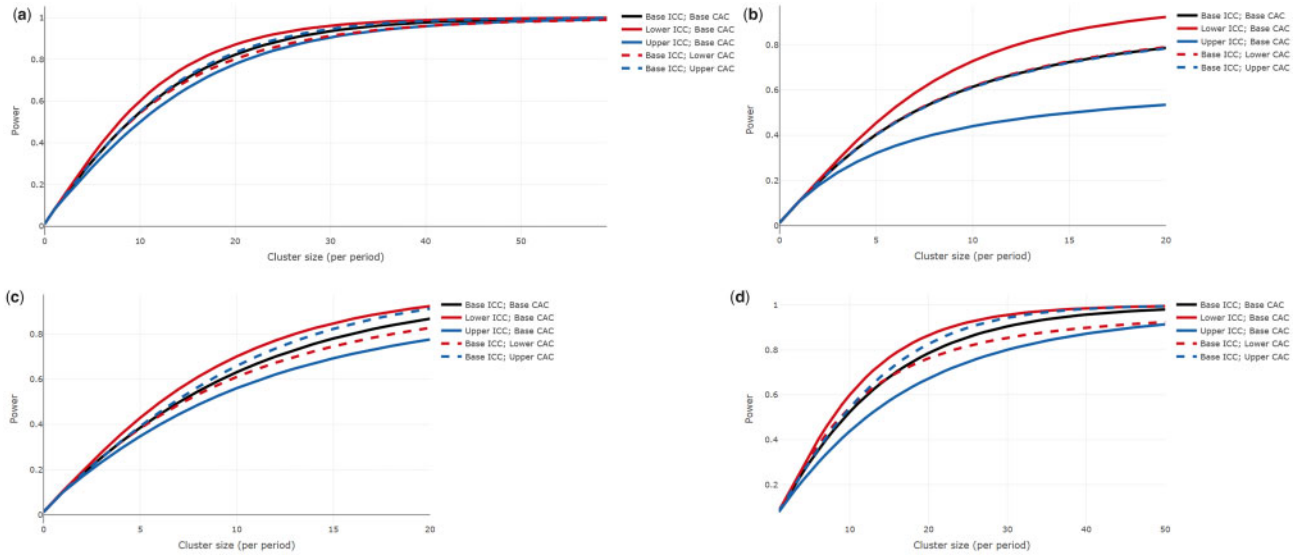

Figure 4. (a) Binary outcome. Scenario includes four clusters per sequence and five sequences in a stepped-wedge design; proportion under control condition is 0.28 and proportion under intervention condition is 0.38; significance level is 0.025; a two-period correlation structure; within period ICC is 0.025 (lower value 0.01 and higher value 0.06); CAC is 0.92 (lower and higher values are 0.74 (80% of base-case) and 1). The expected average cluster-period size is 20. (b) Continuous outcome. Scenario includes four clusters per sequence and five sequences in a stepped-wedge design; to detect a standardized mean difference of 0.25; significance level is 0.025; a two-period correlation structure; within period ICC is 0.056 (lower value 0.023 and higher value 0.13); CAC is 0.08 (lower and higher values 80% and 120% of base case CAC, i.e. 0.064 and 0.096). The expected average cluster-period size is 10. (c) Continuous outcome (high CAC). Scenario includes four clusters per sequence and five sequences in a stepped-wedge design; to detect a standardized mean difference of 0.25; significance level is 0.025; a two-period correlation structure; within period ICC is 0.056 (lower value 0.023 and higher value 0.13); CAC is 0.8 (lower and higher values 80% and 120% of base case CAC, i.e. 0.64 and 0.96). The expected average cluster-period size is 10. (d) Binary outcome, discrete time decay. Scenario includes four clusters per sequence and five sequences in a stepped-wedge design; proportion under control condition is 0.28 and proportion under intervention condition is 0.38; significance level is 0.025; within period ICC is 0.03 (lower value 0.01 and higher value 0.1); CAC is 0.9 (lower and higher values are 0.74 (80% of base-case) and 1). The expected average cluster-period size is 20.

of having an unequal number of clusters per sequence. We do this for the second outcome (continuous outcome eGFR), assuming a complete design and a two-period correlation structure. Assuming the base-case values for the CAC (0.08) and ICC (0.056) and cluster period size (m = 10), the study will have 61% power. We now consider the implications of an extra cluster that we can allocate to any of the five sequences. We modify the design (and upload a CSV file of the design) to allocate this cluster to each of the five sequences in turn (see design diagrams and resulting power in Supplementary Figure S1, available as Supplementary data at *IJE* online). Allocating this cluster to sequence 1 or 5, for example, will increase the power to 69%.To maximize statistical efficiency, researchers might thus consider a trial with five sequences, with four clusters allocated to sequences 2 to 5 and five clusters allocated to sequence 1 (all randomly allocated).

Treatment heterogeneity across clusters

We now use the Shiny CRT Calculator to illustrate the impact of treatment effect heterogeneity across clusters for the binary outcome (acceptance of transplant on first offer). This option only occurs when users upload the design via the upload design option. One way of considering the extent of treatment effect heterogeneity is by considering the expected range of treatment effects across clusters (dividing the range by four, under an assumption of normality, to approximate the standard deviation for the cluster treatment effect).[41] We assume the target is 80% power and assume an average intervention effect across clusters of a 10% absolute difference. Furthermore, we assume that the standard deviation of this treatment effect across clusters is 1%. That is, we assume that the effect of the treatment varies across the clusters such that 95% of the cluster-specific treatment effects are within the range of 10% ± 2%. Note that in order to activate this option, the user must upload the study design via the 'Upload own design' option. For a cluster period size of 20 under the scenario of no treatment heterogeneity, the study would have approximately 89% power (base-case within-period ICC and base-case CAC). When specifying treatment heterogeneity by entering a standard deviation of 0.01 (calculated as 0.04 for the range divided by four), power declines to 88%. In other examples the impact of treatment effect heterogeneity might be much greater.

Example 3: Differential clustering

The ABA trial is an individually randomized trial to answer the question of whether the introduction of a peer support network for new mothers increases numbers of women breastfeeding at 6 weeks. The trial is to have a superiority design and has a single binary outcome representing the proportion of women breastfeeding at 6 months. Hypothesis tests are two-sided; the level of significance is set at 0.05. In this example there are no routinely collected data available to inform estimates of correlations. The study has a limited number of peer support workers (the clusters). We therefore use the Shiny CRT Calculator to illustrate the determination of sample size in a trial with clustering in one arm only.

The estimated prevalence of the primary outcome is around 50%. We considered a target effect size of 10% absolute difference (i.e. prevalence of 50% versus 60%). Under individual randomization, this design would need about 400 per arm for 80% power. We use a moderate value of 0.1 for the ICC, as this is a process outcome (lower range of 0.05 and the upper range of 0.15). To represent individual randomization and no clustering in the control arm, we select the differential clustering option. In this example, clustering occurs only in the intervention arm, but the calculator nevertheless requires specification of the 'cluster size' and ICC in the control arm. In the absence of clustering in the control arm, the 'cluster size' under the control condition is set to 1 (i.e. each individual is their own cluster) and the ICC to 0. We then fix the number of clusters under the treatment condition ('number of clusters (per arm)') to be 30, as an initial start to exploring design options.

Figure 5a shows that under the assumed values for the correlation parameters and when the number of individuals under the control condition is set to be 400 (based on that needed under individual randomization) 80% power is never achieved with 30 clusters in the treatment arm (the maximum achievable power is about 70% under base-case ICC). Changing the number of individuals under the control condition to 700 (but retaining the number of clusters in the intervention arm at 30), in spite of allowing a small increase in power is of limited benefit (Figure 5b). Consequently having 30 clusters in the intervention condition makes the design either infeasible or very inefficient. Increasing the number of clusters under the intervention condition to 40, and having 400 observations under the control condition, enables 80% power to be achievable but does require large cluster sizes under the intervention (Figure 5c).

Figure 6 illustrates how the Shiny CRT Calculator can be used to identify the combination of number of individuals in the control arm and number of clusters in the intervention arm (for a fixed cluster size in the intervention arm) which minimizes the total sample size. This plot reveals that the number in the control arm decreases rapidly as the number of clusters in the intervention condition increases (Figure 6). With a cluster size in the intervention arm of 20, the figure shows that the total sample size is minimized (for 80% power) when the number of clusters in the treatment arm is around 45 and the number of observations in the control arm around 550 (for the base-case ICC).

Conclusions

This paper presents a tutorial on the calculation of sample sizes for different types of cluster randomized trials. The accompanying validated and web-based Shiny CRT Calculator will allow researchers to implement recent methodological advances. In multiple-period cluster randomized designs, correlations within clusters depend not only on cluster membership but also on the time separation between measurements. This is a shift in thinking which is now widely appreciated in the methodological literature but much less known in the applied literature. For researchers who have routinely collected data to hand, we have outlined ways of estimating these correlations. For researchers who do not have routinely collected data, we have provided recommendations for how to determine base-case estimates and then encouraged researchers to investigate sensitivity.

We have shown how researchers can use the Shiny CRT Calculator to determine power across different designs, different sampling structures and different correlation structures, importantly allowing users to investigate the implications of key correlation parameters (the ICC and CAC), to allow for the correlation to decay over time, to incorporate clustering in one arm only and to allow for treatment effect heterogeneity across clusters in power calculations. Finally, the Shiny CRT Calculator allows for incomplete designs and flexibility with respect to the range of designs that can be considered, including individually randomized trials with clustering in one arm only.

Strengths and limitations

The methodology considered here makes some assumptions. First, the methodology used uses large-sample normal approximations and, for binary and count outcomes, approximates the variances of a difference between two proportions/counts. These approximations might break down when there are low proportions/counts or small cluster sizes. We have also assumed generalization of the concept of estimation of ICCs for binary outcomes on the

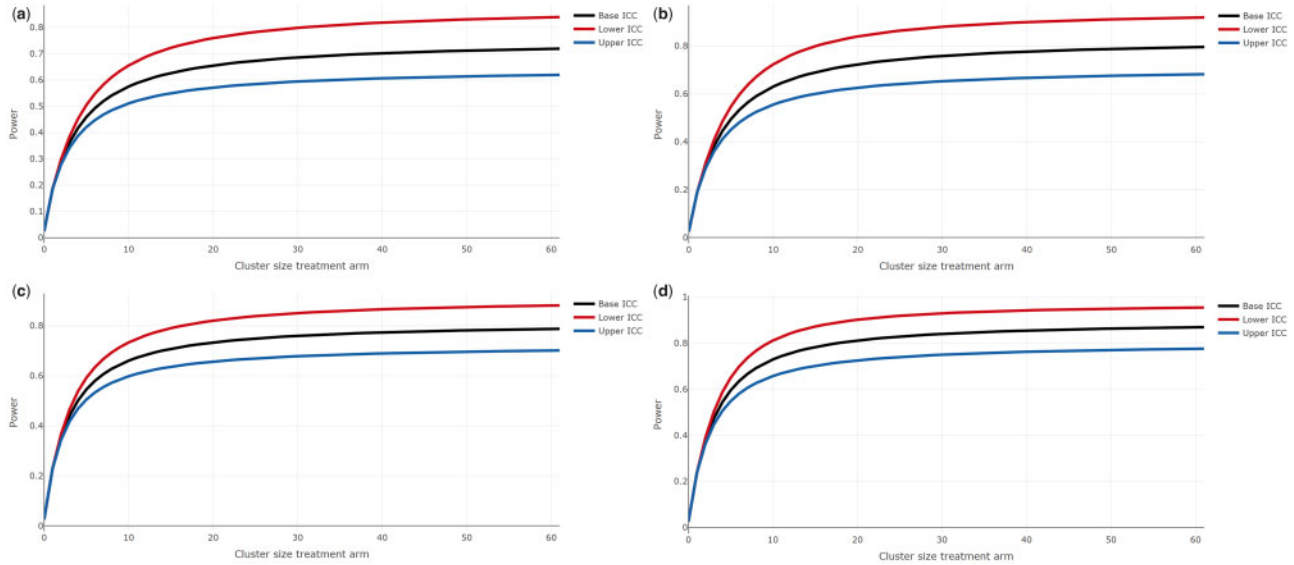

Figure 5. (a) Example 3: power as a function of cluster size in treatment arm for a trial with clustering in one arm only (30 clusters in treatment arm; 400 in control arm). Scenario includes individuals randomized to one of two arms. Assumes 400 individuals are randomized to the control arm [intra-cluster correlation (ICC) 0]; and that there are 30 clusters in the intervention arm (ICC 0.1; lower value 0.05 and higher value 0.15); proportion under control condition is 0.5 and proportion under intervention condition is 0.6; significance level is 0.05. X-axis is cluster size under treatment condition. (b) Example 3: power as a function of cluster size in treatment arm for a trial with clustering in one arm only (30 clusters in treatment arm; 700 in control arm). Scenario includes individuals randomized to one of two arms. Assumes 700 individuals are randomized to the control arm [intra-cluster correlation (ICC) 0]; and that there are 30 clusters in the intervention arm (ICC 0.1; lower value 0.05 and higher value 0.15); proportion under control condition is 0.5 and proportion under intervention condition is 0.6; significance level is 0.05. X-axis is cluster size under treatment condition. (c) Example 3: power as a function of cluster size in treatment arm for a trial with clustering in one arm only (40 clusters in treatment arm; 400 in control arm). Scenario includes individuals randomized to one of two arms. Assumes 400 individuals are randomized to the control arm [intra-cluster correlation (ICC) 0]; and that there are 40 clusters in the intervention arm (ICC 0.1; lower value 0.05 and higher value 0.15); proportion under control condition is 0.5 and proportion under intervention condition is 0.6; significance level is 0.05. X-axis is cluster size under treatment condition. (d) Example 3: power as a function of cluster size in treatment arm for a trial with clustering in one arm only (40 clusters in treatment arm; 700 in control arm). Scenario includes individuals randomized to one of two arms. Assumes 700 individuals are randomized to the control arm [intra-cluster correlation (ICC) 0]; and that there are 40 clusters in the intervention arm (ICC 0.1; lower value 0.05 and higher value 0.15); proportion under control condition is 0.5 and proportion under intervention condition is 0.6; significance level is 0.05. X-axis is cluster size under treatment condition.

proportion scale using linear mixed models, to estimation of CACs for binary outcomes on the proportion scale using linear mixed models. Further work is required to determine the conditions under which the use of these methods is appropriate. We limit our models to the assumption that the variance in the outcome is the same across both arms, whereas others have relaxed this assumption.[40,53]

Furthermore, we have relaxed large-sample assumptions by using critical values from the t-distribution, but the corresponding degrees of freedom used have not been verified for designs other than the simple parallel CRT with continuous outcomes.[49,54] For parallel CRTs, these degrees of freedom corrections result in up to four extra clusters being added to each arm, and our calculations suggest that in stepped-wedge trials the extra number of clusters per sequence might also be high (Supplementary Table S5, available as Supplementary data at *IJE* online). However, we do urge caution in the use of the t-distribution for stepped-wedge studies and other multiple-period designs with small numbers of clusters, as the appropriateness of this degree of freedom correction is unclear.

Finally, whereas the correlation structures considered allow for correlations to decay with increasing separation between periods of measurements, in settings where data are continuously accrued a more intuitive correlation structure is one that allows correlations to decay with increasing separation between 'actual times' of measurement.[31] Furthermore, although we have allowed for within-cluster correlations to depend on time period of measurement at least in some form, in designs that repeatedly measure the same individual we have not allowed for individual-level correlations to depend on time of measurement. Importantly, where the correlation structure is misspecified, sample size can be under- or over-estimated.[19]

The Shiny CRT Calculator provides a user-friendly and flexible means of estimating power across a full range of different cluster trial designs. Yet there are some limitations in its functionality that users should consider. First, we do not recommend that researchers use the Shiny CRT Calculator to identify the minimum number of clusters necessary, as studies with a small number of clusters risk lack of internal and external validity and questionable

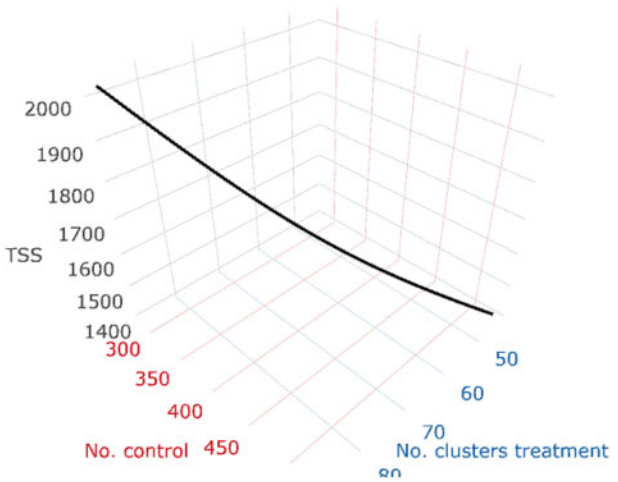

Figure 6. Example 3: power as a function of cluster size in treatment arm, number in control arm and total sample size for a trial with clustering in one arm only (cluster size 20 in treatment arm). Scenario includes individuals randomized to one of two arms. [intra-cluster correlation (ICC) 0.1; lower value and higher values not shown for ease of presentation]; proportion under control condition is 0.5 and proportion under intervention condition is 0.6; significance level is 0.05. Axes show number of clusters under treatment condition; number of individuals randomized to the control condition and the resulting total sample size (TSS). Power is 80%.

suitability of statistical design and analysis methods.[36] Second, for multiple-period designs, the methodology assumes that intervention effects of interest are expressed as time-averaged treatment effects; and alternative approaches are needed where this is not the primary focus.[55] Perhaps more importantly, the Shiny CRT Calculator does rely on asymptotic approximations and users should be mindful of this fact. Whenever cluster sizes, numbers of clusters or proportions of events are small, the calculations will be at increased risk of not meeting the required assumptions. In such circumstances, it might be necessary to consider power by simulation as an alternative, or perhaps complementary, approach.[56]

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

This research was partly funded by the UK NIHR Collaborations for Leadership in Applied Health Research and Care West Midlands initiative. K.H. is funded by an NIHR Senior Research Fellowship (SRF-2017–10-002).

Acknowledgements

Laura Pankhurst (laura.pankhurst@nhsbt.nhs.uk) NHS Blood and Transplant Clinical Trials Unit performed the sample size calculations used for the PITHIA trial with Karla Hemming, which is shown in the first part of example 2. The authors are grateful to all the transplant centres in the UK who contributed data to the UK Transplant Registry, held by NHS Blood and Transplant, which was used in this example. Jim Hughes (jphughes@u.washington.edu) coded the treatment heterogeneity functionality into the app.

Author Contributions

K.H. is guarantor, conceived of the idea, led the writing of the manuscript, developed the application and produced the figures. M.T. co-led the writing and development of ideas with K.H. All authors contributed to writing, drafting and editing the paper. J.K. carried out the calculations for the examples where the correlations followed exponential decay, developed the code for the app for where matrix algebra is used and wrote sections of the paper relating to these elements.

Conflict of interest: None declared.

References

1. Donner A, Klar N. Pitfalls of and controversies in cluster randomization trials. *Am J Public Health* 2004;**94**:416–22.
2. Eldridge S, Kerry S. *A Practical Guide to Cluster Randomized Trials in Health Services Research*. Chichester, UK: Wiley, 2012.
3. Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. *Int J Epidemiol* 1999;**28**:319–26.
4. Teerenstra S, Eldridge S, Graff M, de Hoop E, Borm GF. A simple sample size formula for analysis of covariance in cluster randomized trials. *Stat Med* 2012;**31**:2169–78.
5. Hooper R, Forbes A, Hemming K, Takeda A, Beresford L. Analysis of cluster randomized trials with an assessment of outcome at baseline. *BMJ* 2018;**360**:k1121.
6. Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. *Stat Med* 2008;**27**:5578–85.
7. Arnup SJ, McKenzie JE, Hemming K, Pilcher D, Forbes AB. Understanding the cluster randomized crossover design: a graphical illustraton of the components of variation and a sample size tutorial. *Trials* 2017;**18**:381.
8. Forbes AB, Akram M, Pilcher D, Cooper J, Bellomo R. Cluster randomized crossover trials with binary data and unbalanced cluster sizes: application to studies of near-universal interventions in intensive care. *Clin Trials* 2015;**12**:34–44.
9. Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McKenzie JE. Appropriate statistical methods were infrequently used in cluster-randomized crossover trials. *J Clin Epidemiol* 2016;**74**:40–50.
10. Matthews JN. Multi-period crossover trials. *Stat Methods Med Res* 1994;**3**:383–405.
11. Hemming K, Lilford R, Girling AJ. Stepped-wedge cluster randomized controlled trials: a generic framework including parallel and multiple-level designs. *Stat Med* 2015;**34**:181–96.
12. Hemming K, Taljaard M, McKenzie JE *et al*. Reporting of stepped wedge cluster randomized trials: extension of the CONSORT 2010 statement with explanation and elaboration. *BMJ* 2018;**363**:k1614.
13. Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G, Hargreaves JR. Designing a stepped wedge trial: three main designs, carry-over effects and randomization approaches. *Trials* 2015;**16**:352.

14. Hooper R, Bourke L. Cluster randomized trials with repeated cross sections: alternatives to parallel group designs. *BMJ* 2015; **350**:h2925.
15. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007;**28**:182–91.
16. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. Hoboken, NJ: Wiley-Interscience, 2004.
17. Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. *Int Stat Rev* 2009;**77**:378–94.
18. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. *Int J Epidemiol* 2015; **44**:1051–67.
19. Kasza J, Forbes AB. Inference for the treatment effect in multiple-period cluster randomized trials when random effect correlation structure is misspecified. *Stat Methods Med Res* 2019;**88**:3112–22.
20. Feldman HA, McKinlay SM. Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. *Stat Med* 1994;**13**:61–78.
21. Hemming K. Sample size calculations for stepped wedge trials using design effects are only approximate in some circumstances. *Trials* 2016;**17**:234.
22. 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 randomized trials. *Stat Methods Med Res* 2019;**28**:703–16.
23. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomized trials. *Stat Med* 2016;**35**:4718–28.
24. Yelland LN, Salter AB, Ryan P, Laurence CO. Adjusted intraclass correlation coefficients for binary data: methods and estimates from a cluster-randomized trial in primary care. *Clin Trials* 2011;**8**:48–58.
25. Martin J, Girling A, Nirantharakumar K, Ryan R, Marshall T, Hemming K. Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomized controlled trials for type-2 diabetes in UK primary care. *Trials* 2016;**17**:402.
26. Eldridge SM, Costelloe CE, Kahan BC, Lancaster GA, Kerry SM. How big should the pilot study for my cluster randomized trial be? *Stat Methods Med Res* 2016;**25**:1039–56.
27. Campbell MK, Grimshaw JM, Elbourne DR. Intracluster correlation coefficients in cluster randomized trials: empirical insights into how should they be reported. *BMC Med Res Methodol* 2004;**4**:9.
28. Cook JA, Bruckner T, MacLennan GS, Seiler CM. Clustering in surgical trials - database of intracluster correlations. *Trials* 2012; **13**:2.
29. Gulliford MC, Ukoumunne OC, Chinn S. Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: data from the Health Survey for England 1994. *Am J Epidemiol* 1999;**149**: 876–83.
30. Martin J. Advancing knowledge in stepped-wedge cluster randomized trials.PhD thesis. Institute of Applied Health Research, University of Birmingham, UK. 2017.
31. Grantham KL, Kasza J, Heritier S *et al*. Accounting for a decaying correlation structure in sample size determination for cluster randomized trials with continuous recruitment. *Stat Med* 2019; **38**:1918–34.
32. Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Stat Med* 2016;**35**:2149–66.
33. Kasza J, Forbes A. Information content of cluster period cells in stepped wedge trials. *Biometrics* 2019;**75**:144–52.
34. Lundström E, Isaksson E, Wester P, Laska AC, Näsman P. Enhancing Recruitment Using Teleconference and Commitment Contract (ERUTECC): study protocol for a randomized, stepped-wedge cluster trial within the EFFECTS trial. *Trials* 2018;**19**:14.
35. Hemming K, Eldridge S, Forbes G, Weijer C, Taljaard M. How to design efficient cluster randomized trials. *BMJ* 2017;**358**: j3064.
36. Taljaard M, Teerenstra S, Ivers NM, Fergusson DA. Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. *Clin Trials* 2016;**13**:459–63.
37. Campbell MK, Thomson S, Ramsay CR, MacLennan GS, Grimshaw JM. Sample size calculator for cluster randomized trials. *Comput Biol Med* 2004;**34**:113–25.
38. Lawrie J, Carlin JB, Forbes AB. Optimal stepped wedge designs. *Stat Probab Lett* 2015;**99**:210–14.
39. Roberts C, Roberts SA. Design and analysis of clinical trials with clustering effects due to treatment. *Clin Trials* 2005;**2**: 152–62.
40. Moerbeek M, Wong WK. Sample size formulae for trials comparing group and individual treatments in a multilevel model. *Stat Med* 2008;**27**:2850–64.
41. Hughes JP, Granston TS, Heagerty PJ. Current issues in the design and analysis of stepped wedge trials. *Contemp Clin Trials* 2015;**45**:55–60.
42. Hemming K, Taljaard M, Forbes A. Modeling clustering and treatment effect heterogeneity in parallel and stepped-wedge cluster randomized trials. *Stat Med* 2018;**37**:883–98.
43. Murray DM. *Design and Analysis of Group Randomized Trials*. New York, NY: Oxford University Press, 1998.
44. van Breukelen GJP, Candel M. How to design and analyse cluster randomized trials with a small number of clusters? Comment on Leyrat *et al*. *Int J Epidemiol* 2018., Apr 18. doi: 10.1093/ije/-dyy061. [Epub ahead of print.]
45. Girling AJ. Relative efficiency of unequal cluster sizes in stepped wedge and other trial designs under longitudinal or cross-sectional sampling. *Stat Med* 2018, Sept 12. doi: 10.1002/sim.7943.
46. Candel MJ, Van Breukelen GJ. Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. *Stat Med* 2010;**29**:1488–501.
47. van Breukelen GJ, Candel MJ. Calculating sample sizes for cluster randomized trials: we can keep it simple and efficient! *J Clin Epidemiol* 2012;**65**:1212–28.
48. Martin JT, Hemming K, Girling A. The impact of varying cluster size in cross-sectional stepped-wedge cluster randomized trials. *BMC Med Res Methodol* 2019;**19**:123.
49. Leyrat C, Morgan KE, Leurent B, Kahan BC. Cluster randomized trials with a small number of clusters: which analyses should be used? *Int J Epidemiol* 2018, Mar 27. doi: 10.1093/ije/dyy057. [Epub ahead of print.]

50. Chang W, Cheng J, Allaire J. *Shiny: Web Application Framework for R*, Version 1.4. 2019. https://CRAN.R-project.org/package=shiny (6 December 2017, date last accessed).
51. Lydersen S, Fagerland MW, Laake P. Recommended tests for association in 2 x 2 tables. *Stat Med* 2009;**28**:1159–75.
52. Ayorinde JO, Summers DM, Pankhurst L *et al*. PreImplantation trial of histopathology in renal allografts (PITHIA): a stepped-wedge cluster randomized controlled trial protocol. *BMJ Open* 2019;**9**:e026166.
53. Candel M, Van Breukelen G. Sample size calculation for treatment effects in randomized trials with fixed cluster sizes and heterogeneous intraclass correlations and variances. *Stat Methods Med Res* 2015;**24**:557–73.
54. Lemme F, Van Breukelen GJP, Candel M, Berger M. The effect of heterogeneous variance on efficiency and power of cluster randomized trials with a balanced 2x2 factorial design. *Stat Methods Med Res* 2015;**24**:574–93.
55. Hemming K, Taljaard M, Forbes A. Analysis of cluster randomized stepped wedge trials with repeated cross-sectional samples. *Trials* 2017;**18**:110.
56. Hooper R. Versatile sample size calculation using simulation. *STATA J* 2013;**13**:21–38.