Received: 28 October 2020 | Revised: 2 February 2021 | Accepted: 3 March 2021

DOI: 10.1002/sim.8958

RESEARCH ARTICLE

Statistics in Medicine WILEY

Contamination: How much can an individually randomized trial tolerate?

Karla Hemming[1] | **Monica Taljaard**[2] | **Mirjam Moerbeek**[3] | **Andrew Forbes**[4]

[1] Institute of Applied Health Research, University of Birmingham, Birmingham, UK

[2] Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

[3] Department of Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, The Netherlands

[4] School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia

Correspondence
Karla Hemming, Institute of Applied Health Research, University of Birmingham, Birmingham, UK.
Email: k.hemming@bham.ac.uk

Cluster randomization results in an increase in sample size compared to individual randomization, referred to as an efficiency loss. This efficiency loss is typically presented under an assumption of no contamination in the individually randomized trial. An alternative comparator is the sample size needed under individual randomization to detect the attenuated treatment effect due to contamination. A general framework is provided for determining the extent of contamination that can be tolerated in an individually randomized trial before a cluster randomized design yields a larger sample size. Results are presented for a variety of cluster trial designs including parallel arm, stepped-wedge and cluster crossover trials. Results reinforce what is expected: individually randomized trials can tolerate a surprisingly large amount of contamination before they become less efficient than cluster designs. We determine the point at which the contamination means an individual randomized design to detect an attenuated effect requires a larger sample size than cluster randomization under a nonattenuated effect. This critical rate is a simple function of the design effect for clustering and the design effect for multiple periods as well as design effects for stratification or repeated measures under individual randomization. These findings are important for pragmatic comparisons between a novel treatment and usual care as any bias due to contamination will only attenuate the true treatment effect. This is a bias that operates in a predictable direction. Yet, cluster randomized designs with post-randomization recruitment without blinding, are at high risk of bias due to the differential recruitment across treatment arms. This sort of bias operates in an unpredictable direction. Thus, with knowledge that cluster randomized trials are generally at a greater risk of biases that can operate in a nonpredictable direction, results presented here suggest that even in situations where there is a risk of contamination, individual randomization might still be the design of choice.

KEYWORDS

cluster-randomized trials, contamination, individually randomized trials, statistical efficiency

1 INTRODUCTION

Cluster randomization is a commonly used trial design for evaluating interventions that can only be delivered at the cluster-level, although it is often used to evaluate individual-level interventions.[1-3] One of the reasons individual-level interventions are evaluated using cluster randomization is the concern over contamination such that those allocated to the control inadvertently receive the intervention, perhaps because of geographical or social proximity. In a sample of cluster trials published in 2017,[4] 17% evaluated individual-level interventions, including for example the use of oropharynx in intensive care[5] and the treatment of fever with or without antibiotics.[6] In the presence of contamination, evaluation using individual randomization would not lead to an estimate of the estimand of interest (the effect of offering the treatment to an individual) but rather an estimate of the effect in the presence of contamination. The choice of cluster randomization over individual randomization is therefore often justified, in the presence of contamination, so as to estimate the "true" effect of the intervention under the real-world scenario of offering the intervention to everyone. When contamination operates in one direction only (eg, when comparing a novel intervention to usual care), and all other biases being absent, individual randomization will provide a lower bound for the effect that would be realized under the real-world situation of everyone being offered the intervention. That is to say, when contamination can operate in one direction only, individual randomization leads to an attenuated estimate of the estimand of interest. Although of course there may be other biases that might threaten the stability of this result, such as a bias in the measurement of the outcome data.

On the other-hand, while cluster randomization in theory allows estimation of the true estimand of interest, it is less widely appreciated that cluster randomization puts the evaluation at increased risk of other biases.[7] These biases are mostly unique to a particular type of cluster randomized trial, namely one that uses post-randomization identification or recruitment without blinding of the treatments.[8] While these biases do not affect cluster trials in which participants are identified and recruited prior to randomization and whilst blinding can potentially prevent these sources of bias,[7] these biases are nonetheless contributing to a decrease in robustness of evidence generated from cluster randomized trials.[8,9] Reviews suggest that between 20% and 40% of cluster trials are at risk of these biases.[8,10,11] These biases act in an unpredictable direction and do not affect individually randomized trials where individuals are typically recruited prior to randomization. Cluster randomization also of course results in a loss of statistical efficiency such that the sample size that would have been required under individual randomization has to be inflated to allow for clustering.[1] There are also other risks involved in conducting cluster randomized trials, including issues involved with randomizing a small number of clusters.[12,13] Thus opting for a cluster randomized evaluation of an individual-level intervention because of concerns over contamination, will lead not only to an increase in sample size but also put the design at increased risk of bias especially where there is post-randomization identification or recruitment of individuals. On the other hand, when the comparison of interest is compared to usual care, individually randomized trials (while having increased statistical efficiency) risk attenuation of the treatment effect should there be any contamination.

In the situation where contamination can only operate in one direction (eg, in comparisons against usual care), one option is to use individual randomization with an acceptance that the estimate will be an attenuation of the estimand. This option might be particularly of value when concerns around contamination are small; and/or when the trial requires individual identification and recruitment post randomization of clusters and where blinding is not possible. It transpires that because of the increased efficiency of individual randomization, the sample size needed to detect an attenuated treatment effect, can still be smaller than that required under cluster randomization to detect the non-attenuated effect.[14,15] Deciding between these options requires both consideration of how much contamination might be realized (so as to understand by how much any treatment effect will be attenuated) and how much gain in statistical precision can be achieved (so as to understand if the study has power to detect these smaller treatment effects). Others have investigated the gain in statistical precision under a variety of settings, but this gain has not examined under the increasingly common multiple period cluster randomized design. Examples of such designs are the cluster randomized trial with a baseline period (CRT-B), the stepped-wedge cluster randomized trial (SW-CRT), and the cluster randomized cross-over design (CRXO).[16-18] These multiple period designs can have the benefit of decreasing the sample size over the simple parallel arm CRT, but are often at increased risks of other sources of bias[19] so it is of interest to determine when these designs yield a smaller sample size than the individually randomized design with attenuated treatment effect.

2 | OBJECTIVES

In this paper we provide a general framework for determining the amount of contamination that can be *tolerated* in an individually randomized design (to detect an attenuated treatment effect) before a larger sample size is required than a multiple period cluster randomized design (to detect an nonattenuated treatment effect). While we identify that the amount of contamination that can be tolerated is sometimes very high, our aim is not to advocate that this amount of contamination should be tolerated, but rather illustrate that there will often be substantially more power under individual randomization to estimate treatment effects robustly and precisely in the presence of a small amount of contamination, providing they are correctly interpreted as lower bounds on the estimand of interest. Of note we are only considering one form of contamination, that of control arm contamination by the intervention condition. Furthermore, we are considering only continuous outcomes, superiority designs and comparisons of two treatment conditions. We make model-based assumptions underlying sample size calculations, and provide relevant references.

We start by introducing an illustrative case study to motivate our work. We then summarize the existing literature on the methodology for determining sample size needed under individual randomization, considering both simple randomization, stratified randomization, and designs in which repeated measures are taken—all important considerations which impact the required sample size or statistical efficiency of the study. We then go on to extend these concepts for designs which use cluster randomization, specifically considering designs with repeated measures on both the clusters and individuals. This review of design effects for the first time brings these two frameworks together, allowing us then to show how to determine critical values for the amount of contamination that can be "tolerated" under individual randomization before the sample size to detect an attenuated treatment effects becomes greater than that needed under cluster randomization. We show how this critical value can be generalized to be a function of the design effect for clustering, the design effect for the repeated measure nature of the design and the design effect for any design based adjustments such as stratification. Finally we illustrate the meaning of this work for some example study designs and return to our case study to show the implications for practice.

3 | ILLUSTRATIVE CASE STUDY: A NONBLINDED CLUSTER TRIAL FOR THE EVALUATION OF SELF-ADMINISTERED MISOPROSTOL

Our illustrative case study is an evaluation of the self-administration of misoprostol for preventing postpartum hemorrhage in women giving birth at home in Uganda.[20] In Uganda, postpartum mortality is high, in part arising due to postpartum hemorrhage. There are known preventative and effective treatments, one of which is misoporstol particularly useful in some settings as it does not require cold storage. Because many women give birth at home in Uganda, and because misopostol is both known to be effective when administered in health care settings and because it does not need cold storage, it is believed that providing pregnant women with misoprostol for self-administration just before birth is likely to be effective with minimal risks of harm. The objective of the study was therefore to compare two treatment strategies in women giving birth at home: either standard of care or self-administered misoprostol. The treatments were not blinded. The primary outcome was postpartum hemorrhage (for the purpose of illustrative sample size calculations we focus on a continuous version of this binary outcome, haemoglobin value measured in g/dl). The trial was set across six health facilities (the clusters), with all women presenting for antenatal care eligible for inclusion except those who had planned or had in the past a cesarean delivery.

The reason for choosing a cluster randomized design was not clearly reported but was likely in part due to concern over the possibility of contamination of the control arm with the intervention condition (ie, those allocated to the control arm inadvertently receiving the treatment drug). Yet, under cluster randomization the trial was at risk of identification and recruitment bias (primarily because it was an unblinded cluster trial)—a bias that acts in an unpredictable direction; as well as being much less statistically efficient than an individually randomized design. Indeed, a comparison of the characteristics of the included women show multiple suggestions of identification bias (eg, the number recruited into each trial arm was substantially different; as too was the number with HIV and anemia). The trial also adopted a stepped-wedge design, but because of the small number of clusters and because the stepped-wedge design makes strong assumptions about time effects, it is possible that this might also induce an unpredictable bias into the estimated treatment effect.

Under individual randomization the trial was at risk of contamination, but would have had much more precision. Because the two comparisons were a standard of care against an added intervention (here misoprostol) any contamination would lead to an attenuated treatment effect. With prerandomization recruitment the trial would also have been fully concealed at recruitment, would not have been at risk of identification or recruitment bias—and so would have had greater internal validity. Furthermore, given the outcome was objective there would have been minimal concerns around the study not being blinded. We return to this example to consider whether using individual randomization, accepting the possibility of a small degree of contamination, would have have been a robust alternative study design.

4 | BACKGROUND: SAMPLE SIZE FOR INDIVIDUAL AND CLUSTER RANDOMIZED TRIALS WITH REPEATED MEASURES

In this section we outline previously published formulae for the sample size required for a variety of different designs under individual and cluster randomization. For both individual and cluster randomization, we outline these formulae by considering any inflation or deflation required over a parallel individually randomized design using simple randomization. We consider standard individually randomized designs as well as designs with pre- and post-randomization measurements; and stratified as well as simple randomization. We consider cluster randomized trials with multiple periods of measurement where these measures are taken on either the same (cohort) or different (cross-sectional) participants at each measurement occasion. We use the term *design effect* to denote the inflation (or deflation) in sample size needed over that of simple individual randomization; and outline all these formulae in terms of these design effects. For reasons which become evident later, we also define all formulae in terms of the total number of measurements as opposed to the total number of participants (on whom multiple measurements might be taken).

4.1 | Individually randomized controlled trials

The required sample size per arm for an individually randomized trial, with equal numbers of individuals in each arm, at prespecified power $1-\beta$ to detect a difference of δ (target effect size) for a continuous normally distributed outcome with SD σ^2, is n_I, where:

$$n_I = 2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{\delta^2}\right], \tag{1}$$

and where $z_{\alpha/2}$ is the critical value of the z-distribution with an area $\alpha/2$ in each tail.

4.2 | Individual randomization with stratification

In practice, individually randomized trials often use stratified (sometimes referred to as a randomized block design) as opposed to simple randomization. In a stratified individually randomized trial, individuals are allocated to either of the two treatments at random, but such that within any given stratum (eg, center in a multicenter trial) there is a balance across treatment and control conditions. While stratification is rarely allowed for in sample size calculations, it leads to a smaller sample size to detect the same target effect size compared to an individually randomized trial with simple randomization.[21] The required sample size per arm for a stratified individually randomized trial to detect a difference of δ, for intra-stratum correlation (ISC) ρ_s is n_{ST}, where:

$$n_{\mathrm{ST}} = \underbrace{2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{\delta^2}\right]}_{n_I}[1-\rho_s] = n_I\,\underbrace{[1-\rho_s]}_{\mathrm{DE_{stratification}}}, \tag{2}$$

where ρ_s represents the correlation between outcomes in the same *s*tratum. The reduction in sample size needed due to the stratification can be represented by the design effect for stratification $(1-\rho_s)$. Stratification thus results in a smaller sample size since, as both treatment conditions are balanced within each stratum, each stratum acts as its own control, thereby eliminating between-strata differences.

TABLE 1 Individually randomized controlled trials: Design effects to allow for stratification and repeated measures

Number of pre-measurements	Number of post-measurements	Stratification	Design effect
0	1	No	1
0	1	Yes	$1 - \rho_s$
1	1	No	$2(1 - \rho_i^2)$
0	v	No	$v(\frac{1+(v-1)\rho_i}{v})$
u	v	No	$(v + u) * (\frac{1+(v-1)\rho_i}{v} - \frac{u\rho_i^2}{1+(u-1)\rho_i})$

Notes: The design effect refers to the inflation or deflation in the total sample size (number of measurements not number of participants) over that of simple individually randomization; ρ_s is the correlation between observations from the same stratum; ρ_i is the correlation between two observations from the same individual at different points in time. All formulae for the repeated measures designs are provided in Reference 23.

4.3 | Individual randomization with repeated measures

Individually randomized trials with continuous outcome measures are sometimes supplemented with an adjustment for a prerandomization or baseline measure of the outcome, which we refer to as a repeated measures design. If a baseline measure of the outcome is taken, an ANCOVA analysis (ie, an analysis in which there is an adjustment for the baseline measure) can reduce the required sample size. The sample size per arm under a design supplemented with a baseline measurement of the outcome is:

$$n_B = \underbrace{2\sigma^2 \left[\frac{(z_{\alpha/2} + z_\beta)^2}{\delta^2}\right]}_{n_I} [2(1 - \rho_i^2)] = n_I \underbrace{[2(1 - \rho_i^2)]}_{\text{DE}_{\text{repeatedmeasures}}}, \tag{3}$$

where ρ_i represents the correlation between two measurements on the same individual: one at baseline and one at follow-up.[22,23] Thus, the design effect for an ANCOVA analysis is $2(1 - \rho_i^2)$. For reasons that will become clear in due course, n_B represents the total number of measurements per arm and not the total number of participants per arm each of whom will have two measurements (one pre- and one post-randomization). It is for this reason that the design effect includes the factor 2.

Table 1 extends these design effects for designs with multiple pre and multiple post measures under the assumption of a time averaged treatment effect.[23] Time averaged treatment effects assume interest is in the average effect of the intervention over all of the post-measurements compared to that averaged over all pre-measurements. In addition, all these repeated measures designs assume a correlation structure for which correlations between observations on the same individual are assumed to be constant and do not decay over time (compound symmetry). This assumption might not be tenable in all situations, it is nonetheless a common assumption.[23]

4.4 | Cluster randomized trials

Like the individually randomized trial, the parallel cluster randomized trial can be extended in numerous ways to include repeated measures at the level of the cluster, which we call multiple period cluster randomized designs. These repeated measures might be taken on the same or different individuals overtime (cohort or cross-sectional). We provide design effects here for two common extensions (the cluster randomized trial with a baseline period and the two-period cluster randomized cross-over design) both under the assumption of cross-sectional sampling (repeated measures on different individuals) but provide further design effects for other designs (such as the stepped-wedge and multiple period cross-over design) under both cross-sectional sampling and cohort sampling (Table 2). Our formula can be applied both to multiple period cluster designs where the trial is extended by adding these multiple periods (ie, elongating the design); and where the total study duration is carved up into multiple periods. However, because things simplify for designs in which periods are added, we consider this scenario as a special case. We start by outlining these formulae for the standard parallel cluster randomized trial.

TABLE 2 Design effects for multiple period cluster randomized designs for cohort and cross-sectional sampling

Design	Design effect
CRT	$[1 + (m - 1)\rho]$
CRT*	$[1 - r]^{-1}$
CRT-B	$[1 + (m - 1)\rho_{wp}]2(1 - r^2)$
CRXO	$[1 + (m - 1)\rho_{wp}](1 - r)$
MP-CRXO	$[1 + (m - 1)\rho_{wp}](1 - r)$
SW-CRT	$[1 + (m - 1)\rho_{wp}](t + 1)\frac{3t(1-r)(1+tr)}{(t^2-1)(2+tr)}$

Notes: The design effect refers to the inflation or deflation in the total sample size (number of measurements not number of participants) over that of simple individual randomization; ρ is the intra-cluster correlation; ρ_{wp} is the within-period ICC; and r is the cluster-mean correlation (defined at Equation (9) for cross-sectional sampling and at replaced with r^* at 11 for cohort sampling); t is the number of steps in the SW-CRT. All under the assumption of a block exchangeable and compound symmetry correlation structure. CRT: parallel cluster randomized trial; CRT-B: cluster randomized trial with baseline period; CRXO: two-period cluster randomized cross-over trial; MP-CRXO: multiple period cluster randomized cross-over trial; SW-CRT: stepped-wedge cluster randomized trial. For formula see[16] and;[28] * compared to a stratified individually randomized design.

4.5 | Parallel cluster randomization

The required sample size per arm to detect a difference of δ, in a parallel cluster randomized trial, with an intra-cluster correlation (ICC) ρ and cluster size m is n_{CRT}, where:

$$n_{\text{CRT}} = 2\sigma^2 \left[\frac{(z_{\alpha/2} + z_\beta)^2}{\delta^2}\right] [1 + (m - 1)\rho], \tag{4}$$

$$= n_I \underbrace{[1 + (m - 1)\rho]}_{\text{DE}_{\text{clustering}}}, \tag{5}$$

where $[1 + (m - 1)\rho]$ is the common design effect for clustering and where the ICC measures the extent of the correlation between outcomes measured within the same cluster (assuming an exchangeable correlation structure).[1]

We also determine the inflation needed under cluster randomization over that of individual randomization with stratification so as to illustrate what we might think of as the design effect in the comparison of a cluster design to a stratified individually randomized design. We assume exchangability of the within-stratum and within-cluster correlations ($\rho_s = \rho$). The assumption that $\rho = \rho_S$ would apply when there is an exchangability between the choice of center for stratification under individual randomization and choice of cluster under cluster randomization. This assumption is likely to hold when cluster and strata are the same and when both studies have the same duration. This inflation will be the ratio of n_{CRT} to n_{ST}:

$$\begin{aligned}\frac{n_{\text{CRT}}}{n_{\text{ST}}} &= \frac{n_I[1 + (m - 1)\rho]}{n_I[1 - \rho]} \\ &= \frac{1 + (m - 1)\rho}{(1 - \rho)} \\ &= \frac{1}{1 - r},\end{aligned} \tag{6}$$

where

$$r = \frac{m\rho}{1 + (m - 1)\rho}, \tag{7}$$

is known as the cluster-mean correlation (this parameter arises again in later derivations). So, it turns out that comparing a cluster randomized design against an individually randomized design with stratified randomization, the inflation needed for the clustered aspect of the design is actually $\frac{1}{(1-r)}$ (which is a function of m and ρ) rather than $[1+(m-1)\rho]$.[24]

4.6 | Parallel cluster randomization with a baseline measure

In a parallel cluster randomized trial with baseline measures all clusters are initially in the control condition and then (typically) half receive the intervention.[25] We initially assume cross-sectional sampling such that the participants measured in the first period of the design are different to those measured in the second period of the design. In cluster randomized trials conducted over two periods it is common to assume a correlation structure characterized by two correlation parameters: the within-period ICC (WP-ICC) which allows for measurements within the same cluster-period to be more highly correlated and the between-period ICC (BP-ICC) which allows for measurements in different cluster-periods to be less correlated.[18] The ratio of the within-period to the between-period ICC is called the cluster-auto correlation (CAC). This correlation structure is referred to as a block-exchangeable correlation structure.[26] The sample size per arm to detect a difference of δ, for a within-period ICC of ρ_{wp}, cluster size per period m and where η is the cluster auto-correlation is:

$$n_{\mathrm{CRT-B}} = \underbrace{2\sigma^2\left[\frac{(z_{\alpha/2}+z_{\beta})^2}{\delta^2}\right]}_{n_I}\underbrace{[1+(m-1)\rho_{wp}]}_{\mathrm{DE_{clustering}}}\underbrace{[2(1-r^2)]}_{\mathrm{DE_{multipleperiods}}}, \tag{8}$$

where r is the cluster-mean correlation:

$$r = \frac{m\rho_{wp}\eta}{1+(m-1)\rho_{wp}}, \tag{9}$$

which was first introduced above (Equation (7)) and is here generalized to account for the multiple period aspect of the design. Of note, $n_{\mathrm{CRT-B}}$ denotes the total number of measurements within each arm of the trial under an assumption of an equal number of measurements in the pre and post period. The total number of measurements taken across both arms is $2n_{\mathrm{CRT-B}}$. In this cross-sectional design the number of participants and number of measurements coincide. We note the similarity between this formula and that of the sample size needed under individual randomization with a pre- and post-measurement (Equation (3)). Following others, we have formulated the sample size as the product of the number needed under individual randomization, the design effect for clustering and the design effect for the multiple period aspect of the design.[27] Our parameter ρ_{wp} represents the correlation within a cluster-period. Under the assumption of the duration of the parallel cluster trial being the same as the duration of a single period in the cluster design with baseline measure, then $\rho = \rho_{wp}$.

4.7 | Two-period cluster randomized cross-over design

In a two-period cluster randomized cross-over trial clusters are allocated to one of two sequences.[28] Clusters allocated to the first sequence are initially observed in the control condition and then switch to the intervention condition. Clusters allocated to the second sequence are initially observed under the intervention condition and then under the control condition. We again, initially assume cross-sectional sampling. The sample size under each treatment condition to detect a difference of δ, for within-period ICC ρ_{wp}, cluster size per period m and where η again represents the cluster auto-correlation is:

$$n_{\mathrm{CRXO}} = \underbrace{2\sigma^2\left[\frac{(z_{\alpha/2}+z_{\beta})^2}{\delta^2}\right]}_{n_I}\underbrace{[1+(m-1)\rho_{wp}]}_{\mathrm{DE_{clustering}}}\ \underbrace{[(1-r)]}_{\mathrm{DE_{multipleperiods}}}\ , \tag{10}$$

where again $r = \frac{m\rho_{wp}\eta}{1+(m-1)\rho_{wp}}$ is the cluster-mean correlation. Again, under the assumption of the duration of the parallel cluster trial being the same as the duration of a single period in the cluster design with baseline measure, then $\rho = \rho_{wp}$.

4.8 | Cluster randomization with other extensions

So far these multiple period cluster randomized designs have assumed cross-sectional sampling. However, this is easily extended to cohort designs by capitalizing on the definition of the cluster-mean correlation, which can be rewritten as a function of the individual level correlation (ρ_i):[27]

$$r^* = \frac{m\rho_{wp}\eta + (1-\rho_{wp})\rho_i}{1+(m-1)\rho_{wp}}, \tag{11}$$

and substituting r for r^* in the above. Here the advantage of defining the total sample sizes as the number of measurements (rather than the number of participants) becomes clear: the sample size is the same for both cohort and cross-sectional sampling with the formulae only differing by the definition of the cluster-mean correlation (r or r^*).

These multiple period cluster randomized designs can be extended in other ways too. For example, a cluster randomized cross-over trial might include multiple cross-overs. We provide the design effects for other multiple period cluster randomized designs in Table 2. These design effects have all been derived elsewhere[16] except for the design effect for the multiple cross-over design which is derived in Appendix A. We note that the design effect for the multiple cross-over designs, on first sight, appears to be identical to that of the two period cross-over design. However, in practice there will be differences between the two design effects, as the cluster-period size m and the within-period ICC (ρ_{wp}) will change across two and multiple-period designs. We also note that all these design effects assume a block exchangeable correlation structure. Others have proposed more realistic correlation structures, but these mostly do not simplify to design effects and can often face convergence issues at the analysis stage.[29]

5 | DETERMINING CRITICAL VALUES FOR RATES OF CONTAMINATION

We now compare the sample size needed under individual randomization to detect an attenuated treatment effect with the sample size needed under a multiple period cluster randomized design (to detect the nonattenuated effect). In this way we derive the contamination rate at which an individually randomized trial with an attenuated treatment effect begins to require a larger sample size than a multiple period cluster randomized design. We call this the critical value for the rate of contamination. This is a critical value that, if exceeded, makes the individually randomized trial less statistically *efficient*. We start from the simplest case, by deriving these critical values for a parallel cluster randomized design compared to a range of individually randomized designs; and in so doing are reproducing the work of others. We then extend these derivations to provide critical values for multiple-period cluster designs. We show how this critical value is the ratio of two design effects: that for the multiple period cluster randomized design versus that for the individually randomized design (eg, under stratification).

5.1 | Comparing to parallel CRTs

If it is expected that there may be contamination of the control arm with the intervention, then any attenuation of the treatment effect can be allowed for in the sample size calculation. Under a scenario where the rate of contamination is w, the required sample size per arm for an individually randomized trial (without stratification) to detect an attenuated difference is n_I^* (henceforth we use the notation n^* to denote sample size under individual randomization to detect an attenuated difference), where:

$$n_I^* \approx 2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{((1-w)\delta)^2}\right] = n_I \underbrace{(1-w)^{-2}}_{\text{DE}_{\text{contamination}}}. \tag{12}$$

By attenuated difference from δ we mean the difference that is expected assuming $w\%$ of the control arm receive the full effect of the intervention and where the full effect of the intervention is δ. In the case of partial contamination, w is replaced by pf where p is the proportion of control group that is contaminated and f is the fraction of contamination (constant across subjects) representing the proportion of the intervention condition that the control arm participants

TABLE 3 Critical values for rate of contamination beyond which an individually randomized trial (with or without a single measures and stratification) requires a larger sample size than a parallel cluster randomized design

Design without contamination	Comparator with attenuated effect	Critical value
CRT	iRCT	$1-[1+(m-1)\rho]^{-1/2}$
CRT	iRCT-B	$1-\left[\frac{[1+(m-1)\rho]}{2(1-\rho_i^2)}\right]^{-1/2}$
CRT	iRCT stratified	$1-\left[\frac{1}{(1-r)}\right]^{-1/2}$

Abbreviations: CRT, parallel cluster randomized trial; iRCT, individually randomized trial; iRCT-B, individually randomized trial with baseline measure; ρ, intra-cluster correlation; r, cluster-mean correlation (Equation (9)) to be replaced with r^* at equation (11) for cohort sampling); ρ_i, individual level correlation.

receive.[30] It transpires that the realized effect under contamination is $(1-w)\delta$. The term $[(1-w)^{-2}]$ might be considered as the *design effect* for contamination, where we again use the term design effect to denote the inflation (or deflation) in sample size needed over that of simple individual randomization without any contamination. This result is derived in Appendix B. In Appendix B we also show how these results can be extended to allow for a resulting nonhomogeneous variance across the two arms, as a result of the contamination.

We can therefore determine the contamination rate at which an individually randomized trial, designed to detect an attenuated treatment effect, requires a larger sample size than a cluster randomized trial. For a parallel cluster randomized design this will occur when $n_I^* > n_{\mathrm{CRT}}$:

$$2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{((1-w)\delta)^2}\right] > 2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{\delta^2}\right][1+(m-1)\rho].$$

That is, when:

$$(1-w)^{-2} > [1+(m-1)\rho].$$

$$w > 1-[1+(m-1)\rho]^{-1/2}. \tag{13}$$

Table 3 extends this to comparisons of a cluster randomized design with an individually randomized design with a baseline measure and to an individually randomized design with stratification (see Appendix for full derivation). We see that these critical values are a function of the ratio of the design effect due to clustering to the design effect for repeated measures or stratification used in the individually randomized design.

5.2 | Comparing to multiple-period cluster randomized designs

These critical values can be derived under the cluster randomized design with multiple periods. To this end we might for example compare the individually randomized trial with one pre- and one post-measure to the cluster randomized design with a baseline measure, again determining the contamination rate at which an individually randomized trial with an attenuated treatment effect requires a larger sample size than a cluster randomized trial. This will occur when $n_B^* > n_{\mathrm{CRT-B}}$:

$$2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{((1-w)\delta)^2}\right]2(1-\rho_i^2) > 2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{\delta^2}\right][1+(m-1)\rho_{wp}]2(1-r^2).$$

That is, when:

$$(1-w)^{-2}2(1-\rho_i^2) > [1+(m-1)\rho_{wp}]2(1-r^2).$$

$$w > 1-\left[\frac{[1+(m-1)\rho_{wp}]2(1-r^2)}{2(1-\rho_i^2)}\right]^{-1/2}. \tag{14}$$

TABLE 4 Critical values for rate of contamination beyond which an individually randomized trial requires a larger sample size than a multiple period cluster randomized design (for cohort or cross-sectional sampling)

Design without contamination	Comparator with attenuated effect	Critical value
CRT-B	iRCT-B	$1-\left[\frac{[1+(m-1)\rho_{wp}]2(1-r^2)}{2(1-\rho_i^2)}\right]^{-1/2}$
CRT-B	iRCT	$1-\left[[1+(m-1)\rho_{wp}]2(1-r^2)\right]^{-1/2}$
CRXO	iRCT	$1-\left[[1+(m-1)\rho_{wp}](1-r)\right]^{-1/2}$
MP-CRXO	iRCT	$1-\left[[1+(m-1)\rho_{wp}](1-r)\right]^{-1/2}$
SW-CRT	iRCT	$1-\left[[1+(m-1)\rho_{wp}](t+1)\frac{3t(1-r)(1+tr)}{(t^2-1)(2+tr)}\right]^{-1/2}$

Abbreviations: CRT-B, cluster randomized trial with baseline period; CRXO, two-period cluster randomized cross-over trial; MP-CRXO, multiple period cluster randomized cross-over trial; SW-CRT, stepped-wedge cluster randomized trial; iRCT, individually randomized trial; iRCT-B, individually randomized trial with baseline measure; ρ_{wp}, within-period ICC; r, cluster-mean correlation (Equation (9)) to be replaced with r^* at Equation (11) for cohort sampling); ρ_i, individual level correlation; t is the number of sequences in the SW-CRT.

We see that this critical value is a function of the ratio of the design effects due to clustering and multiple periods to the design effect for repeated measures in the individually randomized design. By replacing r (Equation (9)) with r^* (Equation (11)) in Equation (C5), the above result holds for both cross-sectional and cohort sampling.

Table 4 extends these to include other forms of multiple-period cluster trials introduced earlier. We can also make these comparisons to a stratified individually randomized trial, where if we make the assumption that $\rho_s = \rho_{wp} = \rho$, then again things simplify. Under the assumption of the duration of the individually randomized trial being the same as the duration of a single period in the cluster design with baseline measure, then $\rho_s = \rho_{wp}$ is likely to be a reasonable assumption. Full details are included in Appendix.

Finally this leads to a generic way of determining the critical value

$$1-\left[\frac{\mathrm{DE}_{\mathrm{clustering}}\mathrm{DE}_{\mathrm{multipleperiods}}}{\mathrm{DE}_{\mathrm{stratification}}\mathrm{DE}_{\mathrm{repeatedmeasures}}}\right]^{-1/2}$$

6 | PRACTICAL APPLICATIONS

This paper provides researchers the option of making comparisons between any type of multiple period cluster randomized design with any other type of reference individually randomized design, where the choice of design can be chosen to reflect those that might be feasible in any given scenario.

We now make some general observations about the conditions, defined by cluster sizes and ICCs, under which individually randomized designs with attenuated treatment effects likely will require a smaller sample size than a multiple period CRT. Our general observations use both the formula considered explicitly above in the derivations and also use the generic formula. For simplicity, all our practical applications are under the assumption of the duration of the parallel cluster trial is the same as the duration of a single period in the cluster design with baseline measure, so that $\rho = \rho_{wp}$; and the duration of the individually randomized trial is the same as the duration of a single period in the cluster design with baseline measure, so that $\rho_s = \rho_{wp}$. We then provide an illustration of how these results might be useful in practice.

7 | GENERAL OBSERVATIONS

Figure 1 shows critical values for the amount of contamination that can be tolerated in a standard individually randomized trial (single measurement) before the sample size becomes larger than that of a parallel cluster randomized trial. If we take an example with a small cluster size (say $m = 10$) and an ICC of 0.05, we see that up to about 20% contamination can be tolerated before the sample size using individual randomization (powered to detect the attenuated treatment effect) exceeds that of a parallel CRT. For a large cluster size of $m = 500$ and for an ICC of 0.05, up to about 80% contamination can be tolerated before the sample size exceeds that of a cluster randomized design. We also observe that the amount

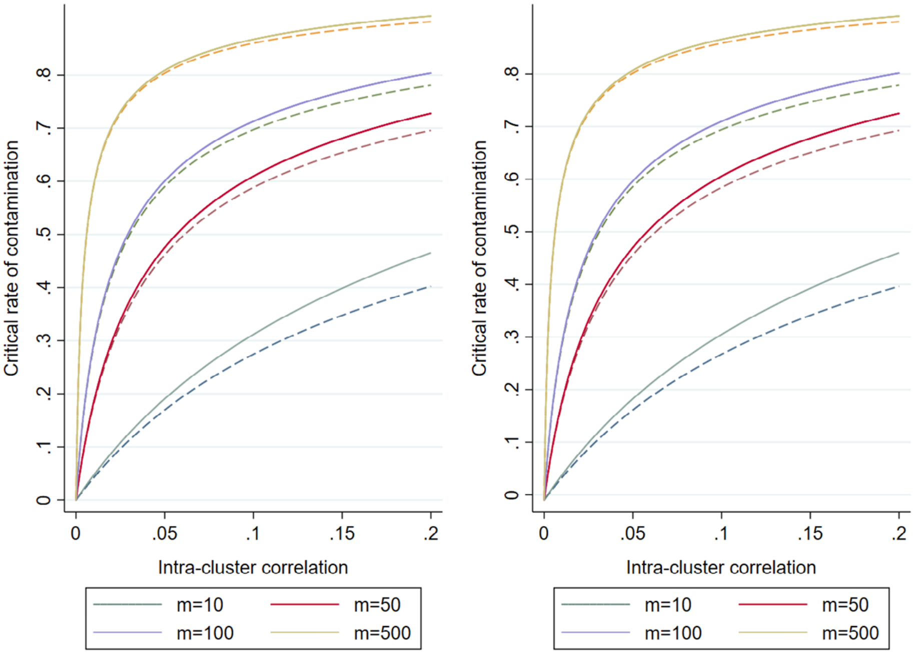


FIGURE 1 Rate of contamination that can be tolerated in an individually randomized trial with simple (solid) and stratified (dash) allocation compared to cluster randomization as a function of cluster size (m) and ICC (for contamination beyond this rate the iRCT is less efficient than the CRT); LHS, for an individually randomized trial with single post-measurement; RHS, for an individually randomized trial with a baseline measure ($\rho_i = 0.7$) [Colour figure can be viewed at wileyonlinelibrary.com]

of contamination that can be tolerated decreases slightly when compared to stratified randomization (dashed lines on figure) and compared to an individually randomized design with a baseline measure (right-hand side figure, $\rho_i = 0.7$). Hence we observe that the amount of contamination that can be tolerated increases with factors known to make the cluster randomized design less statistically efficient (increasing ICC and increasing cluster size); and decreases with factors known to make the individually randomized design more statistically efficient (stratification and adding a baseline measure). Table 5 illustrates the practical implications of these results, showing sample sizes needed under a cluster randomized design (for a range of cluster sizes and correlations) compared to that required under individual randomization with various degrees of contamination.

Figure 2 shows critical values for a parallel CRT with a baseline measure compared to an individually randomized trial with a single post measurement (under both simple and stratified randomization). We again observe that the amount of contamination that can be tolerated increases with the cluster-period size and the within-period ICC. We also note that the amount of contamination that can be tolerated might be larger or smaller than if the alternative trial design was a parallel design without a baseline measure (Figure 1). We take first an example with a small cluster-period size ($m = 10$). Assuming for example a within-period cluster correlation of 0.05 and a cluster auto-correlation of 0.8, up to about 40% contamination can be tolerated under individual randomization before the sample size exceeds that of a CRT with a baseline. While, without the baseline measure only about 20% contamination can be tolerated (Figure 1). Taking next an example with a large cluster-period size ($m = 500$) and again assuming a within-period ICC of 0.05 and a cluster auto-correlation of 0.8. This time we see that the amount of contamination that can be tolerated in a cluster randomized trial with a baseline measure (about 80%, Figure 2) is smaller than can be tolerated in a simple parallel CRT (about 90%, Figure 1). This finding echos results of comparative efficiency research: whether the parallel cluster trial or parallel cluster trial with baseline measures can tolerate more contamination depends on factors known to make the cluster randomized design less statistically efficient than the cluster randomized design with baseline measures (increasing ICC and increasing cluster size).[31] We also observe that compared to an individually randomized trial with stratified (as opposed to simple) randomization, the amount of contamination that can be tolerated again decreases slightly; and also decreases slightly when adding a baseline measure.

Figure 3 shows critical values for a two-period cluster randomized cross-over trial compared to an individually randomized trial with a single post-measurement. Here we see that for some scenarios (eg, small cluster-period size) the amount of contamination that can be tolerated is very small. Two-period cluster cross-over designs are known to be very statistically efficient designs. For example, for small cluster-period sizes and small within period ICCs the design results in very little increase in the sample size over individually randomization.[28] This high efficiency of the design means there for small ICCs and small cluster-period sizes there is little room to tolerate any contamination.

TABLE 5 Sample size (per arm) to detect various standardized effect sizes over a range of intracluster correlations and cluster sizes for the parallel cluster and cluster with baseline measure compared to an individual randomized design with contamination

					Percentage contamination (iRCT)				
WP-ICC	**SES**	**m**	**CRT**	**CRT-BA**	**0**	**5**	**10**	**20**	**50**
0.001	0.1	10	1582	3164	1568	1737	1936	2450	6272
0.001	0.1	50	1645	3284	1568	1737	1936	2450	6272
0.001	0.1	100	1723	3423	1568	1737	1936	2450	6272
0.001	0.15	10	703	1406	697	772	860	1089	2788
0.001	0.15	50	731	1459	697	772	860	1089	2788
0.001	0.15	100	766	1521	697	772	860	1089	2788
0.001	0.3	10	176	352	174	193	215	272	697
0.001	0.3	50	183	365	174	193	215	272	697
0.001	0.3	100	191	380	174	193	215	272	697
0.05	0.1	10	2274	4109	1568	1737	1936	2450	6272
0.05	0.1	50	5410	6217	1568	1737	1936	2450	6272
0.05	0.1	100	9330	7986	1568	1737	1936	2450	6272
0.05	0.15	10	1010	1826	697	772	860	1089	2788
0.05	0.15	50	2404	2763	697	772	860	1089	2788
0.05	0.15	100	4146	3549	697	772	860	1089	2788
0.05	0.3	10	253	457	174	193	215	272	697
0.05	0.3	50	601	691	174	193	215	272	697
0.05	0.3	100	1037	887	174	193	215	272	697
0.1	0.1	10	2979	4621	1568	1737	1936	2450	6272
0.1	0.1	50	9251	7739	1568	1737	1936	2450	6272
0.1	0.1	100	17 091	10 878	1568	1737	1936	2450	6272
0.1	0.15	10	1324	2054	697	772	860	1089	2788
0.1	0.15	50	4112	3440	697	772	860	1089	2788
0.1	0.15	100	7596	4835	697	772	860	1089	2788
0.1	0.3	10	331	513	174	193	215	272	697
0.1	0.3	50	1028	860	174	193	215	272	697
0.1	0.3	100	1899	1209	174	193	215	272	697

Abbreviations: CRT, parallel cluster randomized design; CRT-B, cluster randomized trial with baseline period; iRCT, individually randomized trial; WP-ICC, within-period ICC; cluster-auto correlation assumed to be 0.9 under a cross-sectional design; m, cluster size per period; SES, standardized effect size; all for 80% power.

7.1 | Illustrative case study: a nonblinded *low risk of bias* individually randomized trial to determine if self-administered misoprostol is effective

We return to the case study introduced earlier where the objective was to evaluate the effect of a new treatment strategy on postpartum haemorrhage for women giving birth at home in Uganda. The study was conducted as a stepped-wedge design with six clusters and included 2466 women giving birth at home. The study included three sequences (four measurement periods) and two clusters randomly allocated to each sequence. We assume cross-sectional sampling and expected cluster size of about 400 and so equating to observing a total of 100 women in each cluster-period (total sample size of 2400). The

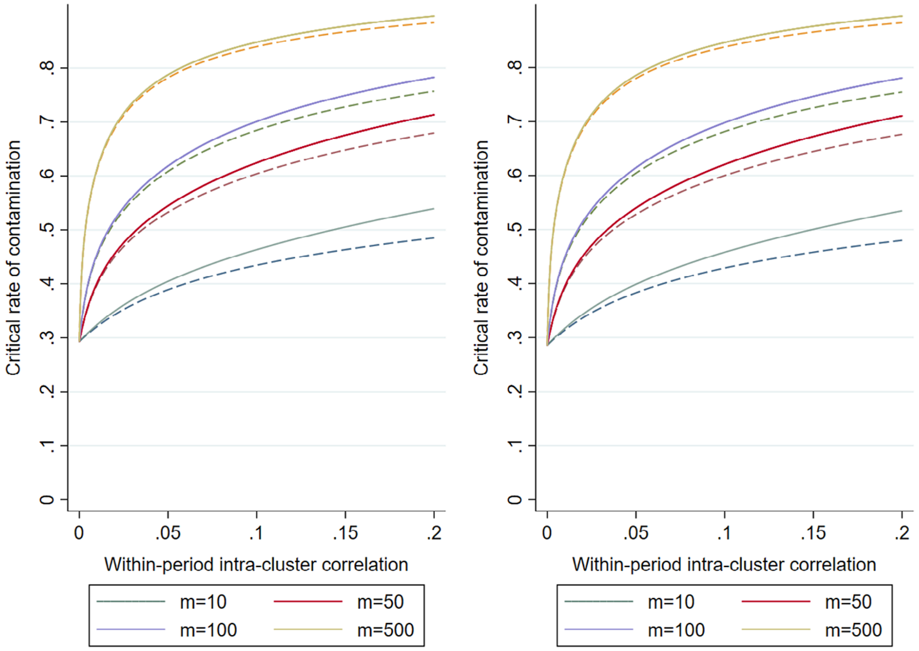


FIGURE 2 Rate of contamination that can be tolerated in an individually randomized trial using simple (solid) and stratified (dash) randomization compared to a (cross-sectional) parallel cluster randomized trial with a baseline measure for different cluster-period sizes (m) assuming $\eta = 0.8$ (for contamination beyond this rate the iRCT is less efficient than the CRT-B); LHS, for an individually randomized trial with single post measurement; RHS, for an individually randomized trial with a baseline measure ($\rho_i = 0.7$) [Colour figure can be viewed at wileyonlinelibrary.com]

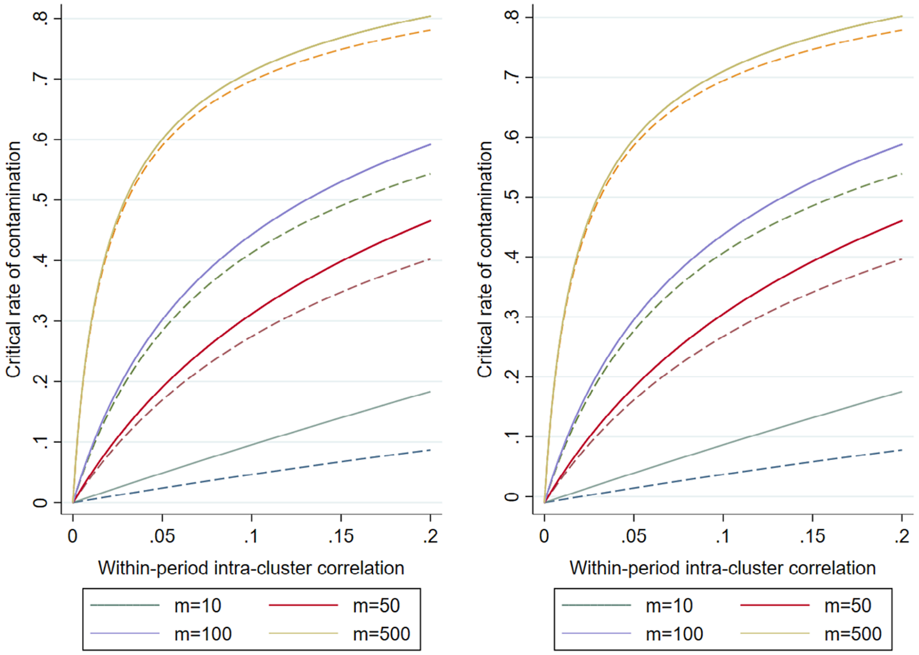


FIGURE 3 Rate of contamination that can be tolerated in an individually randomized trial using simple (solid) and stratified (dash) randomization compared to a (cross-sectional) two-period cluster cross-over trial for different cluster-period sizes (m) assuming $\eta = 0.8$ (for contamination beyond this rate the iRCT is less efficient than the CRXO); LHS, for an individually randomized trial with single post-measurement; RHS, for an individually randomized trial with a baseline measure ($\rho_i = 0.7$) [Colour figure can be viewed at wileyonlinelibrary.com]

trial did not report a within-period ICC, so we assume a typical value of 0.01 and consider sensitivity across a reasonable range.[32] In absence of information on the cluster auto-correlation we consider the value of 0.8.[16]

Figure 4 shows that for likely values of ICCs, about 50% contamination could have been tolerated under individual randomization before the sample size under individual randomization exceeded that needed under the stepped-wedge design. We consider the outcome of blood loss on a standardized effect scale. Under a stepped-wedge design this study would have had about 90% power to detect a 0.25 standardized effect (calculated using the Cluster Shiny App, *https://clusterrcts.shinyapps.io/Cluster-RCT-Sample-Size-Calculator/*[33]). Results presented in this paper suggest that the study would have equivalent power under individual randomization to detect an attenuated target difference of $0.25 * (1 - w) = 0.25 * 0.5 = 0.125$. That is, these results suggest that under individual randomization the study would have been able to detect the smaller standardized effect size of 0.125. Indeed, standard calculations confirm this is approximately correct: under individual randomization, a sample size of 2400 provides approximately 90% power to detect a standardized effect of 0.125 at 5% significance (stata code: power two means 0.125, n(2400)).

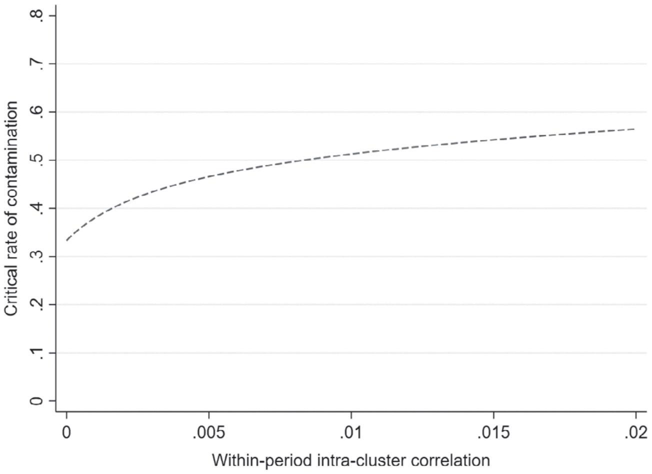


FIGURE 4 Case study: rate of contamination that can be tolerated in an individually randomized trial compared to a stepped-wedge cluster randomized trial (three sequences) for cluster-period sizes ($m = 100$) assuming a cluster-auto correlation of 0.8 (for contamination beyond this rate the iRCT is less efficient than the SW-CRT)

Under individual randomization, this study would thus be powered to detect a much smaller target effect size compared to that under cluster randomization. However, importantly set up as an individually randomized design the study would have been at risk of contamination (in so far as those allocated to the control arm might inadvertently received the intervention). Consequently, any estimated treatment effect under individual randomization would represent an attenuation of that which would have been observed had the control arm not had access to the intervention. While the rate of contamination could not be known in advance, it is unlikely to be hugely problematic in a setting where resources are limited, and likely much smaller than 50% tolerable. This coupled with the fact that the study was at numerous risks of bias run as a stepped-wedge design suggest that individual randomization should have been the design of choice.

8 | DISCUSSION

In summary, our results indicate that individually randomized trials can tolerate a surprisingly large amount of contamination before their sample size requirement exceeds that of a cluster trial to detect the nonattenuated effect. The rate of contamination that can be tolerated depends on within-cluster correlations (including cluster auto-correlations), the cluster-size and the design of the studies being compared (eg, whether the trial includes any repeated measures at the level of the individual or cluster and whether the individually randomized design is stratified). Not surprisingly, but perhaps not commonly appreciated, the rate of contamination that can be tolerated thus increases with factors known to make cluster randomization less statistically efficient; and decreases with factors known to make the individually randomized design more efficient.[31]

When the pragmatic comparison is between a novel treatment and usual care (ie, there are no concerns of contamination of the intervention with the control) any bias that arises due to contamination of the control with the intervention will only attenuate the true treatment effect. This means that using individual randomization leads to an attenuation of the estimand of interest when the evaluation takes a pragmatic stance. In a trial comparing two active interventions it would be important to consider contamination in both directions. For example, in a head-to-head comparison of two treatments for postpartum hemorrhage, contamination might arise across both treatment conditions. Moreover, some studies are designed to show performance under ideal situations (efficacy), and as such contamination of the active treatment with the control (ie, noncompliers) will be important. For example, nonadherence with a novel treatment for postpartum hemorrhage would be important when the objective is to demonstrate potential for effect in those adhering. Furthermore, whatever the primary objective, researchers will naturally have interest in the effect of receiving treatment—that is, in the nonattenuated estimate. Considering contamination of the treatment effect as an issue of noncompliance offers insights through the use of complier average causal effects. Complier average causal effect (CACE) estimates provide estimation (under a number of assumptions) of the effect of receiving the intervention, rather than the effect of offering the intervention. Consequently, in settings where the objective is to estimate the effect of receiving an intervention, with the presence of contamination and under individual randomization, the CACE estimate of the treatment effect could be computed.[34] Indeed a method known as *a contamination adjusted intention to treat analysis* has been proposed[35-37] and implemented.[38]

Conducting an individual randomized trial when there is concern over contamination will always need careful consideration. First and foremost, researchers should always be mindful of ways to prevent contamination. Indeed there are other ways of preventing contamination, other than using cluster randomization, including pseudo cluster randomization[39] and combination of a crossed and nested design.[40] Secondly estimates of likely rates of contamination will help inform decision making: expected small amounts of contamination are likely to provide a convincing case to use individual randomization; whereas if the expected rate of contamination is high then the rate of attenuation will likely be too high to make the treatment estimate meaningful. Resources documenting within-cluster correlations are now reasonably common;[41-43] and there are also established predictors of degree of correlation.[32,44] There have subsequently been calls for overviews of estimates of the rate of contamination.[45] Although much of this research has been focused on educational interventions[46] there are examples of trials that have measured and reported rates of contamination across a range of different settings;[47,48] some work on establishing predictors of contamination;[49] and reviews which suggest that the rate of contamination is very varied and infrequently reported.[50] In those situations where the expected rate of contamination is sufficiently minimal, careful consideration needs to be given to ways of measuring the extent of contamination as this can help triangulate and explain findings. So, in our case study if it was identified that the resulting treatment effect under contamination was very small, knowledge that use of the active treatment had been virtually nonexistent in the control arm, could be useful to support conclusions.

There are other issues we have not considered. We have assumed homogeneity of the variance of the outcome across the two treatment arms. In practice it might be the case that in studies with contamination the variance under the control condition would increase reflecting the mixture of observations receiving the treatment and control condition. Thus the variance will depend on the size of the treatment effect. Our consideration of this (Appendix B) suggests that in practice, while increasing the variance of the treatment effect under individual randomization, the practical extent of this is minimal when designing trials to detect standardized effect sizes smaller than about 0.3. For larger effect sizes, particularly in the presence of a large amount of contamination, the approximation to a homogeneous variance will not be valid. A recent review of studies funded in the United Kingdom identified that the average target effect size was 0.3 and the average realized effect size was 0.11.[51] Related to this, treatment effects might vary across centers and this also might have implications.[30] Finally, we have briefly touched on individually randomized designs with repeated measures on the same individual. Here we have assumed that these multiple measurements are equally correlated across all measurement occasions (compound symmetry). While some empirical evidence suggests this assumption may be tenable,[23] logical reasoning would suggests that multiple measurements on the same individual will have decreasing correlation with increasing separation between the time of the measurements, so this assumption might not be appropriate. We have made similar assumptions under the multiple period cluster randomized design in so far as we have assumed a block exchangeable correlation, which also does not allow for decreasing correlation with increasing separation between measurements.[16,29] We have also not considered issues such as varying cluster sizes; the implications of these findings on cluster trials with a small number of clusters; or more complicated correlation structures. However, almost all of these issues are associated with either increased complexity, risk of bias, or decreased statistical efficiency of the cluster randomized design and are likely to reinforce the findings that where individual randomization is theoretically possible and feasible, under small to medium amounts of contamination an individually randomized design should be the design of choice.

9 | CONCLUSION

While individual randomization can be used to estimate an attenuated treatment effect with high statistical efficiency, cluster randomization seems inherently more appealing because it theoretically allows estimation of the nonattenuated treatment effect. CRTs also offer other advantages, including logistical, political and practical advantages; furthermore, when both direct and indirect intervention effects are of interest, the cluster randomized design is the only feasible choice. Broad eligibility criteria enhance generalization of findings and cluster randomization is often perceived, perhaps not always correctly, to be a means to this end. Yet, cluster randomized designs with post randomization recruitment or identification of participants without blinding, are at high risk of bias due to the differential recruitment across treatment arms. This sort of bias operates in an unpredictable direction. Thus, with knowledge that cluster randomized trials are generally at a greater risk of biases that can operate in a nonpredictable direction, results presented here suggest that even in situations where there is a risk of contamination, individual randomization might still be the design of choice even when there is an objective to estimate real-world effectiveness.

ACKNOWLEDGEMENTS

This research was partly funded by the UK NIHR Collaborations for Leadership in Applied Health Research and Care West Midlands initiative. Karla Hemming is funded by a NIHR Senior Research Fellowship SRF-2017-10-002.

ORCID

Karla Hemming https://orcid.org/0000-0002-2226-6550
Mirjam Moerbeek https://orcid.org/0000-0001-5537-1237

REFERENCES

1. Donner A, Klar N. Design and analysis of cluster randomization trials in health research; 2000.
2. Mdege ND, Brabyn S, Hewitt C, Richardson R, Torgerson DJ. The 2 × 2 cluster randomized controlled factorial trial design is mainly used for efficiency and to explore intervention interactions: a systematic review. *J Clin Epidemiol*. 2014;67(10):1083-1092.
3. Rutterford C, Taljaard M, Dixon S, Copas A, Eldridge S. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. *J Clin Epidemiol*. 2015;68(6):716-723.
4. Turner L, Hemming K. Incomplete evidence of impact: challenges and opportunities in reporting findings from cluster randomised trials. Submitted.
5. De Smet AMGA, Kluytmans JAJW, Cooper BS, et al. Decontamination of the digestive tract and oropharynx in ICU patients. *N Engl J Med*. 2009;360(1):20-31.
6. Chinbuah MA, Kager PA, Abbey M, et al. Impact of community management of fever (using antimalarials with or without antibiotics) on childhood mortality: a cluster-randomized controlled trial in Ghana. *Amer J Trop Med Hyg*. 2012;87(5_Suppl):11-20.
7. Eldridge S, Kerry S, Torgerson DJ. Bias in identifying and recruiting participants in cluster randomised trials: what can be done? *BMJ*. 2009;339:b4006.
8. Bolzern J, Mnyama N, Bosanquet K, Torgerson DJ. A review of cluster randomized trials found statistical evidence of selection bias. *J Clin Epidemiol*. 2018;99:106-112.
9. Hahn S, Puffer S, Torgerson DJ, Watson J. Methodological bias in cluster randomised trials. *BMC Med Res Methodol*. 2005;5(1):10.
10. Eldridge S, Ashby D, Bennett C, Wakelin M, Feder G. Internal and external validity of cluster randomised trials: systematic review of recent trials. *BMJ*. 2008;336(7649):876-880.
11. Puffer S, Torgerson D, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *BMJ*. 2003;327(7418):785-789.
12. 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(4):459-463.
13. Kahan BC, Forbes G, Ali Y, et al. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. *Trials*. 2016;17(1):438.
14. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? *BMJ*. 2001;322(7282):355-357.
15. Howe A, Keogh-Brown M, Miles S, Bachmann M. Trials in medical education: fret less about contamination and more about statistical power. Expert consensus on contamination in educational trials elicited by a Delphi exercise. *Med Educ*. 2007;41:196-204.
16. Hooper R, Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *BMJ*. 2015;350:h2925.
17. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials*. 2007;28(2):182-191.
18. Parienti J-J, Kuss O. Cluster-crossover design: a method for limiting clusters level effect in community-intervention studies. *Contemp Clin Trials*. 2007;28(3):316-323.
19. Hemming K, Taljaard M. Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice? *Int J Epidemiol*. 2020.
20. Ononge S, Campbell OMR, Kaharuza F, Lewis JJ, Fielding K, Mirembe F. Effectiveness and safety of misoprostol distributed to antenatal women to prevent postpartum haemorrhage after child-births: a stepped-wedge cluster-randomized trial. *BMC Pregnancy Childbirth*. 2015;15(1):315.
21. Kahan BC, Morris TP. Analysis of multicentre trials with continuous outcomes: when and how should we account for centre effects? *Stat Med*. 2013;32(7):1136-1149.
22. Senn S, Stevens L, Chaturvedi N. Repeated measures in clinical trials: simple strategies for analysis using summary measures. *Stat Med*. 2000;19(6):861-877.
23. Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Stat Med*. 1992;11(13):1685-1704.
24. Moerbeek M, Breukelen GJP, Berger MPF. Design issues for experiments in multilevel populations. *J Educ Behav Stat*. 2000;25(3):271-284.
25. Teerenstra S, Eldridge S, Graff M, Hoop E, Borm GF. A simple sample size formula for analysis of covariance in cluster randomized trials. *Stat Med*. 2012;31(20):2169-2178.
26. Li F. Design and analysis considerations for cohort stepped wedge cluster randomized trials with a decay correlation structure. *Stat Med*. 2020;39(4):438-455.
27. Hooper R, Teerenstra S, Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Stat Med*. 2016;35(26):4718-4728.

28. Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. *Stat Med*. 2008;27(27):5578-5585.
29. Kasza J, Hemming K, Hooper R, Matthews JNS, Forbes AB. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Stat Methods Med Res*. 2019;28(3):703-716.
30. Moerbeek M. Randomization of clusters versus randomization of persons within clusters: which is preferable? *Am Stat*. 2005;59(2):173-179.
31. Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Stat Med*. 2016;35(13):2149-2166.
32. Gulliford MC, Adams G, Ukoumunne OC, Latinovic R, Chinn S, Campbell MJ. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *J Clin Epidemiol*. 2005;58(3):246-251.
33. Hemming K, Kasza J, Hooper R, Forbes A, Taljaard M. A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the Shiny CRT calculator. *Int J Epidemiol*. 2020.
34. Moerbeek M, Van Schie S. What are the statistical implications of treatment non-compliance in cluster randomized trials: a simulation study. *Stat Med*. 2019;38(26):5071-5084.
35. Cuzick J, Edwards R, Segnan N. Adjusting for non-compliance and contamination in randomized clinical trials. *Stat Med*. 1997;16(9):1017-1029.
36. Sussman JB, Hayward RA. An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials. *BMJ*. 2010;340:c2073.
37. Hewitt CE, Torgerson DJ, Miles JNV. Individual allocation had an advantage over cluster randomization in statistical efficiency in some circumstances. *J Clin Epidemiol*. 2008;61(10):1004-1008.
38. Roemeling S, Roobol MJ, Otto SJ, et al. Feasibility study of adjustment for contamination and non-compliance in a prostate cancer screening trial. *Prostate*. 2007;67(10):1053-1060.
39. Borm GF, Melis RJF, Teerenstra S, Peer PG. Pseudo cluster randomization: a treatment allocation method to minimize contamination and selection bias. *Stat Med*. 2005;24(23):3535-3547.
40. Moerbeek M. Cost-efficient designs for three-arm trials with treatment delivered by health professionals: sample sizes for a combination of nested and crossed designs. *Clin Trials*. 2018;15(2):169-177.
41. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clin Trials*. 2005;2(2):99-107.
42. Taljaard M, Donner A, Villar J, et al. Intracluster correlation coefficients from the 2005 WHO global survey on maternal and perinatal health: implications for implementation research. *Paediatr Perinat Epidemiol*. 2008;22(2):117-125.
43. Moerbeek M, Teerenstra S. *Power Analysis of Trials with Multilevel Data*. Boca Raton, FL: Chapman & Hall/CRC Press; 2015.
44. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol*. 2004;57(8):785-794.
45. Rhoads C. The implications of contamination for educational experiments with two levels of nesting. *J Res Educ Effect*. 2016;9(4):531-555.
46. Keogh-Brown MR, Bachmann MO, Shepstone L, et al. Contamination in trials of educational interventions. *Health Technol Assess (Winch Eng)*. 2007;11(43):1–+.
47. Kingston N, Thomas I, Johns L, Moss S, others. Assessing the amount of unscheduled screening ("Contamination") in the control arm of the UK "Age" trial. *Cancer Epidemiol Prev Biomark*. 2010;19(4):1132-1136.
48. Pinsky PF, Black A, Kramer BS, Miller A, Prorok PC, Berg C. Assessing contamination and compliance in the prostate component of the prostate, lung, colorectal, and ovarian (PLCO) cancer screening trial. *Clin Trials*. 2010;7(4):303-311.
49. Howe A, Keogh-Brown M, Miles S, Bachmann M. Expert consensus on contamination in educational trials elicited by a Delphi exercise. *Med Educ*. 2007;41(2):196-204.
50. Magill N, Knight R, McCrone P, Ismail K, Landau S. A scoping review of the problems and solutions associated with contamination in trials of complex interventions in mental health. *BMC Med Res Methodol*. 2019;19(1):4.
51. Rothwell JC, Julious SA, Cooper CL. A study of target effect sizes in randomised controlled trials published in the health technology assessment journal. *Trials*. 2018;19(1):1-13.
52. Day NE. Estimating the components of a mixture of normal distributions. *Biometrika*. 1969;56(3):463-474.

How to cite this article: Hemming K, Taljaard M, Moerbeek M, Forbes A. Contamination: How much can an individually randomized trial tolerate?. *Statistics in Medicine*. 2021;40:3329–3351. https://doi.org/10.1002/sim.8958

APPENDIX A. DESIGN EFFECT FOR MULTIPLE PERIOD CLUSTER RANDOMIZED CROSS-OVER

To derive the design effect for the multiple period cluster randomized cross-over design we extend the work of Hopper 2016.[27] We momentarily use their notation. In that paper, it is shown that the design effect for the multiple period aspect of a cluster design with $T+1$ periods (equation directly below Equation (5)) is:

$$= \frac{L^2(1-r)(1+Tr)}{4\left[LB - D + r(B^2 + LTB - TD - LC)\right]} \tag{A1}$$

where i is sequence; j is period; and A_{ij} denotes exposure to intervention at time period j in sequence i, and L is the number of treatments, where $B = \sum_{ij} A_{ij}$; $C = \sum_i \sum_j (A_{ij})^2$; $D = \sum_j \sum_i (A_{ij})^2$. Then in Equation (6) we are told:

$$n_{\text{bc}} = \text{Def f}_R(r) * \text{Deff}_C(m, \rho) * (T+1) * n_{\text{SI}} \tag{A2}$$

where n_{bc} is the total sample size across the design, n_{SI} is the total sample size across all treatment arms under individual randomization; Def $\text{f}_C(m, \rho)$ is the design effect for clustering and so the remaining elements are what we have defined here as the design effect for the repeated measures aspect of the design. That is, the design effect for the repeated measures aspect of the design (as we have framed it) is:

$$\text{Def f}_R(r)(T+1), \tag{A3}$$

where it is important to note that $T+1$ is the number of time periods in the study (using notation in Hooper 2016). Then for a MP-CRXO design we have $L=2$; $B=T+1$; $C=(T+1)^2/2$; $D=T+1$. So that:

$$\begin{aligned}
\text{Deff}_R(r)(T+1) &= \frac{2^2(1-r)(1+Tr)}{4\left[2(T+1) - (T+1) + r((T+1)^2 + 2T(T+1)^2 - T(T+1)^2 - (T+1)^2)\right]}(T+1) \\
&= \frac{(1-r)(1+Tr)}{(T+1) + rT(r+1)}(T+1) \\
&= \frac{(1-r)}{T+1}(T+1) \\
&= (1-r).
\end{aligned} \tag{A4}$$

So, therefore design effect due to the repeated measures aspect of the multiple period cluster randomized cross-over design is:

$$= (1-r). \tag{A5}$$

APPENDIX B. DERIVATION OF SAMPLE SIZE IN THE PRESENCE OF CONTAMINATION UNDER INDIVIDUAL RANDOMIZATION

Using notation in the main paper, for an individually randomized trial with contamination w, meaning $100w\%$ of control subjects get the intervention, then the expected mean in the control arm is:

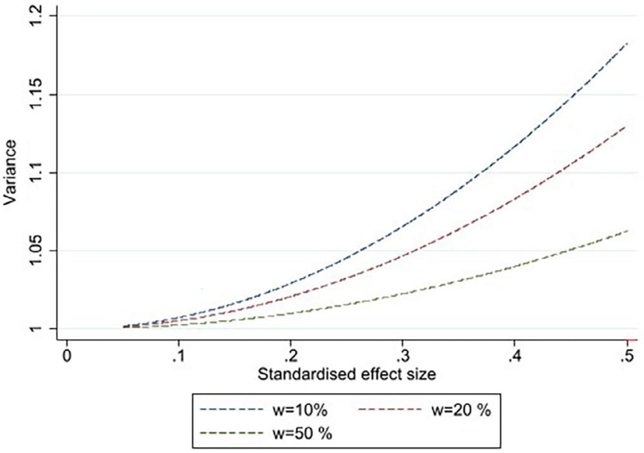
$$(1-w)\mu_c + w\mu_t, \tag{B1}$$

where μ_c and μ_t are the means for the control and treatment conditions, respectively, when received. In the case of partial contamination, w is replaced by pf where p is the proportion of control group that is contaminated and f is the fraction of contamination (constant across subjects) representing the proportion of the intervention condition that the control arm participants receive.[30] It transpires that the realized effect under contamination is $(1-w)\delta$. So the expected (attenuated) treatment difference is:

$$\underbrace{\mu_t}_{\text{mean } Tx} - \underbrace{((1-w)\mu_c + w\mu_t)}_{\text{mean } Tx'} = (1-w)\delta \tag{B2}$$

The distribution of outcomes in the control arm is thus a mixture of two normal distributions: those observations which are uncontaminated are assumed to arise from $N[\mu_c, \sigma^2]$ and those from the contaminated part assumed to arise

FIGURE B1 Variance under the control arm assuming a mixture of two normal distributions ($\sigma^2_{c'}$) to allow comparison with simpler assumption of a standard normal distribution with variance $\sigma^2_c = 1$ [Colour figure can be viewed at wileyonlinelibrary.com]



from $N[\mu_t, \sigma^2]$ with weights $1 - w$ and w respectively. Observations in the control arm therefore arise from a mixture of two normal distributions with mean:[52]

$$\mu_{c'} = (1 - w)\mu_c + w\mu_t,$$

and variance

$$\sigma^2_{c'} = \sigma^2 + w(\mu_c - \mu_{c'})^2 + (1 - w)(\mu_t - \mu_{c'})^2,$$

which will always be larger than σ^2. The resulting sample size per arm becomes:

$$n^*_I = (\sigma^2 + \sigma^2_{c'}) \left[\frac{(z_{\alpha/2} + z_\beta)^2}{((1 - w)\delta)^2} \right], \tag{B3}$$

note the factor 2 (seen in Equation (1)) has disappeared here reflecting the average of the two variance terms $(\sigma^2 + \sigma^2_{c'})/2$.

Figure B1 demonstrates the value of $\sigma^2_{c'}$ for a range of standardized effect sizes. The increase in $\sigma^2_{c'}$ over and above what it would be simplifying by assuming a homogeneous variance ($\sigma^2 = 1$) is negligible for standardized effect sizes less than about 0.2. For higher rates of contamination and large effect sizes its increase over one is not negligible.

Under the assumption of a homogeneous variance, the sample size per arm in an individually randomized trial with contamination is:

$$n^*_I \approx 2\sigma^2 \left[\frac{(z_{\alpha/2} + z_\beta)^2}{((1 - w)\delta)^2} \right]. \tag{B4}$$

The term $[(1 - w)^{-2}]$ might be considered as the *design effect* for contamination, where we again use the term design effect to denote the inflation (or deflation) in sample size needed over that of simple individual randomization without any contamination.

$$n^*_I \approx \underbrace{2\sigma^2 \left[\frac{(z_{\alpha/2} + z_\beta)^2}{\delta^2} \right]}_{n_I} \underbrace{(1 - w)^{-2}}_{\text{DE}_{\text{contamination}}}. \tag{B5}$$

APPENDIX C. DETERMINING CRITICAL VALUES FOR RATES OF CONTAMINATION

We now compare the sample size needed under individual randomization to detect an attenuated treatment effect with the sample size needed under a multiple period cluster randomized design (to detect the nonattenuated effect). In this way we derive the contamination rate at which an individually randomized trial with an attenuated treatment effect begins to require a larger sample size than a multiple period cluster randomized design. We call this the critical value for the rate of contamination. This is a critical value that, if exceeded, makes the individually randomized trial less statistically *efficient*.

TABLE C1 Critical values for rate of contamination beyond which an individually randomized trial (with or without repeated measures) requires a larger sample size than a multiple period cluster randomized design (for cohort or cross-sectional sampling)

Design without contamination	Comparator with attenuated effect	Critical value
CRT	iRCT	$1-[1+(m-1)\rho]^{-1/2}$
CRT-B	iRCT	$1-\left[[1+(m-1)\rho_{wp}]2(1-r^2)\right]^{-1/2}$
CRXO	iRCT	$1-\left[[1+(m-1)\rho_{wp}](1-r)\right]^{-1/2}$
MP-CRXO	iRCT	$1-\left[[1+(m-1)\rho_{wp}]\frac{v(1-r)}{v}\right]^{-1/2}$
SW-CRT	iRCT	$1-\left[[1+(m-1)\rho_{wp}](t+1)\frac{3t(1-r)(1+tr)}{(t^2-1)(2+tr)}\right]^{-1/2}$
CRT	iRCT-B	$1-\left[\frac{[1+(m-1)\rho]}{2(1-\rho_i^2)}\right]^{-1/2}$
CRT-B	iRCT-B	$1-\left[\frac{[1+(m-1)\rho_{wp}]2(1-r^2)}{2(1-\rho_i^2)}\right]^{-1/2}$
CRXO	iRCT-B	$1-\left[\frac{[1+(m-1)\rho_{wp}](1-r)}{2(1-\rho_i^2)}\right]^{-1/2}$
MP-CRXO	iRCT-B	$1-\left[\frac{[1+(m-1)\rho_{wp}]2\frac{v(1-r)}{v}}{2(1-\rho_i^2)}\right]^{-1/2}$
SW-CRT	iRCT-B	$1-\left[\frac{[1+(m-1)\rho_{wp}](t+1)\frac{3t(1-r)(1+tr)}{(t^2-1)(2+tr)}}{2(1-\rho_i^2)}\right]^{-1/2}$
CRT	iRCT stratified	$1-\left[\frac{1}{(1-r)}\right]^{-1/2}$
CRT-B	iRCT stratified	$1-\left[\frac{[1+(m-1)\rho_{wp}]2(1-r^2)}{(1-\rho_s)}\right]^{-1/2}$
CRXO	iRCT stratified	$1-\left[\frac{[1+(m-1)\rho_{wp}](1-r)}{(1-\rho_s)}\right]^{-1/2}$
MP-CRXO	iRCT stratified	$1-\left[\frac{[1+(m-1)\rho_{wp}]2\frac{v(1-r)}{v}}{(1-\rho_s)}\right]^{-1/2}$
SW-CRT	iRCT stratified	$1-\left[\frac{[1+(m-1)\rho_{wp}](t+1)\frac{3t(1-r)(1+tr)}{(t^2-1)(2+tr)}}{(1-\rho_s)}\right]^{-1/2}$
CRT- with multiple periods	iRCT with repeated measures or stratification	$1-\left[\frac{\mathrm{DE}_{\text{clustering}}\mathrm{DE}_{\text{multiple periods}}}{\mathrm{DE}_{\text{stratification}}\mathrm{DE}_{\text{repeated measures}}}\right]^{-1/2}$

Note: CRT, parallel cluster randomized trial; CRT-B, cluster randomized trial with baseline period; CRXO, two-period cluster randomized crossover trial; MP-CRXO, multiple period cluster randomized crossover trial; SW-CRT, stepped-wedge cluster randomized trial; iRCT, individually randomized trial; iRCT-B, individually randomized trial with baseline measure; ρ, intracluster correlation; ρ_{wp}, within-period intracluster correlation; r, cluster-mean correlation (Equation (9)) to be replaced with r^* at Equation (11) for cohort sampling); ρ_i, individual level correlation; ρ_s within-stratum correlation; v is the number of periods in the MP-CRXO; t is the number of sequences in the SW-CRT.

We start from the simplest case, by deriving these critical values for a parallel cluster randomized design compared to a range of individually randomized designs; and in so doing are reproducing the work of others. We then extend these derivations to provide critical values for multiple-period cluster designs. We show how this critical value is the ratio of two design effects: that for the multiple period cluster randomized design versus that for the individually randomized design (eg, under stratification).

C.1 Individually randomized designs with single post-measurements

We now determine the contamination rate at which an individually randomized trial, designed to detect an attenuated treatment effect, requires a larger sample size than a cluster randomized trial. In that follows we assume homogeneity of variances (B4). For a parallel cluster randomized design this will occur when $n_I^* > n_{\text{CRT}}$:

$$2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{((1-w)\delta)^2}\right] > 2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{\delta^2}\right][1+(m-1)\rho].$$

That is, when:

$$(1-w)^{-2} > [1+(m-1)\rho].$$

$$w > 1-[1+(m-1)\rho]^{-1/2}. \tag{C1}$$

We can also determine the contamination rate at which an individually randomized trial with an attenuated treatment effect requires a larger sample size than a cluster randomized trial with baseline measures. This will occur when $n_I^* > n_{\text{CRT-B}}$. That is when:

$$2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{((1-w)\delta)^2}\right] > 2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{\delta^2}\right][1+(m-1)\rho_{wp}]2(1-r^2).$$

That is when:

$$(1-w)^{-2} > [1+(m-1)\rho_{wp}]2(1-r^2)$$

$$w > 1-[[1+(m-1)\rho_{wp}]2(1-r^2)]^{-1/2}. \tag{C2}$$

Because the sample size is the number of measurements, this formula is applicable to both cross-sectional and cohort sampling designs by replacing r by r^* for cohort sampling. Likewise we determine the contamination rate at which an individually randomized trial with an attenuated treatment effect requires a larger sample size than a two-period cluster cross-over design. This will occur when $n_I^* > n_{\text{CRXO}}$. That is when:

$$2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{((1-w)\delta)^2}\right] > 2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{\delta^2}\right][1+(m-1)\rho_{wp}](1-r).$$

That is, when:

$$(1-w)^{-2} > [1+(m-1)\rho_{wp}](1-r).$$

$$w > 1-[[1+(m-1)\rho_{wp}](1-r)]^{-1/2}. \tag{C3}$$

These critical values are all a function of the design effect for clustering and the design effect for the multiple period aspect of the cluster design (Table 2).

C.2 Individually randomized designs with repeated measures

Above the cross-sectional multiple period cluster designs were compared against individually randomized designs with a single measurement only. We now extend this to individually randomized designs with repeated measures on the same individual. We thus introduce the notation n_B^* to denote the sample size needed under an individually randomized design with a baseline measure (repeated measure on same individual), to detect an attenuated treatment effect.

For a parallel cluster randomized design (with a single period rather than multiple periods) this will occur when $n_B^* > n_{\text{CRT}}$:

$$2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{((1-w)\delta)^2}\right]2(1-\rho_i^2) > 2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{\delta^2}\right][1+(m-1)\rho].$$

That is, when:

$$(1-w)^{-2}2(1-\rho_i^2) > [1+(m-1)\rho].$$

$$w > 1-\left[\frac{[1+(m-1)\rho]}{2(1-\rho_i^2)}\right]^{-1/2}. \tag{C4}$$

We see that this critical value is a function of the ratio of the design effect due to clustering to the design effect for repeated measures in the individually randomized design. Note that this is comparing a cluster randomized design with a single measurement to an individually randomized design with two measurements per individual.

We expand this to the cluster randomized design with a baseline measure. To this end we compare the individually randomized trial with one pre- and one post-measure to the cluster randomized design with a baseline measure, again determining the contamination rate at which an individually randomized trial with an attenuated treatment effect requires a larger sample size than a cluster randomized trial. This will occur when $n_B^* > n_{\text{CRT-B}}$:

$$2\sigma^2 \left[\frac{(z_{\alpha/2}+z_\beta)^2}{((1-w)\delta)^2}\right] 2(1-\rho_i^2) > 2\sigma^2 \left[\frac{(z_{\alpha/2}+z_\beta)^2}{\delta^2}\right] [1+(m-1)\rho_{wp}]2(1-r^2).$$

That is, when:

$$(1-w)^{-2}2(1-\rho_i^2) > [1+(m-1)\rho_{wp}]2(1-r^2).$$

$$w > 1 - \left[\frac{[1+(m-1)\rho_{wp}]2(1-r^2)}{2(1-\rho_i^2)}\right]^{-1/2}. \tag{C5}$$

We see that this critical value is a function of the ratio of the design effects due to clustering and multiple periods to the design effect for repeated measures in the individually randomized design. By replacing r (Equation (9)) with r^* (Equation (11)) in equation (C5), the above result holds for both cross-sectional and cohort sampling.

C.3 Stratified individually randomized designs

We can also make these comparisons to a stratified individually randomized trial. That is, we make a comparison between a trial with a stratified design using individual randomization and powered to detect an attenuated treatment effect, compared to a multiple period parallel cluster randomized design. The individually randomized design will require a larger sample size when $n_{\text{ST}}^* > n_{\text{CRT}}$, where we introduce the notation n_{ST}^* to represent the sample size under individual and stratified randomization in the presence of an attenuated treatment effect. That is when:

$$2\sigma^2 \left[\frac{(z_{\alpha/2}+z_\beta)^2}{((1-w)\delta)^2}\right] (1-\rho_s) > 2\sigma^2 \left[\frac{(z_{\alpha/2}+z_\beta)^2}{\delta^2}\right] [1+(m-1)\rho].$$

So, when:

$$\begin{aligned}(1-w)^{-2}(1-\rho_s) &> [1+(m-1)\rho]\\ w &> 1 - \frac{[1+(m-1)\rho_s]^{-1/2}}{(1-\rho)^{-1/2}}\end{aligned}$$

If we make the assumption that $\rho = \rho_S$, then things simplify nicely so that:

$$w > 1 - \frac{[1+(m-1)\rho]^{-1/2}}{(1-\rho)^{-1/2}} = 1 - \left[\frac{1}{(1-r)}\right]^{-1/2},$$

since for single period designs (eg, single measurement occasion) the cluster-mean correlation, r (Equation (9)) reduces to $r = \frac{m\rho}{[1+(m-1)\rho]}$ and $1-r = \frac{(1-\rho)}{[1+(m-1)\rho]}$. The assumption that $\rho = \rho_S$ would apply when there is an exchangability between the choice of center for stratification under individual randomization and choice of cluster under cluster randomization. This assumption is likely to hold when cluster and strata are the same and when both studies have the same duration.

To aid understanding we also determine the inflation needed under cluster randomization over that of individual randomization with stratification (to detect the same nonattenuated target effect size) so as to illustrate what we might think of as the design effect in the comparison of a cluster design to a stratified individually randomized design. We again assume exchangability of the within-stratum and within-cluster correlations ($\rho_s = \rho$). This inflation will be the ratio of n_{CRT} to n_{ST}:

$$\begin{aligned}\frac{n_{\text{CRT}}}{n_{\text{ST}}} &= \frac{n_I[1+(m-1)\rho]}{n_I[1-\rho]}\\ &= \frac{1+(m-1)\rho}{(1-\rho)}\\ &= \frac{1}{1-r}.\end{aligned} \tag{C6}$$

So, it turns out that comparing a cluster randomized design against an individually randomized design with stratified randomization, the inflation needed for the clustered aspect of the design is actually $\frac{1}{(1-r)}$ (which is a function of m and ρ) rather than $[1+(m-1)\rho]$.[24]

We can extend the comparison against a stratified individually randomized design to the other multiple period cluster designs. For example, we make a comparison between a trial with a stratified individually randomized design, powered to detect an attenuated treatment effect, compared to a parallel cluster randomized design with a baseline measure. The individually randomized design will require a larger sample size when $n^*_{\text{ST}} > n_{\text{CRT-B}}$. That is when:

$$2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{((1-w)\delta)^2}\right](1-\rho_s) > 2\sigma^2\left[\frac{(z_{\alpha/2}+z_\beta)^2}{\delta^2}\right][1+(m-1)\rho_{wp}]2(1-r^2),$$

where $r=\frac{m\rho_{wp}\eta}{1+(m-1)\rho_{wp}}$ for the multiple period design under cross-sectional sampling. So, when:

$$\begin{aligned}(1-w)^{-2}(1-\rho_s) &> [1+(m-1)\rho_{wp}]2(1-r^2)\\ w &> 1-\frac{([1+(m-1)\rho_{wp}]2(1-r^2))^{-1/2}}{(1-\rho_s)^{-1/2}}.\end{aligned} \tag{C7}$$

If we make the assumption that $\rho_s=\rho_{wp}=\rho$, then: $\frac{r}{\eta}=\frac{m\rho}{(1+(m-1)\rho}$ and $1-\frac{r}{\eta}=\frac{1-\rho}{(1+(m-1)\rho}$, so that:

$$w > 1-\left[\frac{1}{1-\frac{r}{\eta}}\right]^{-1/2}(2(1-r^2))^{-1/2}. \tag{C8}$$

Under the assumption of the duration of the individually randomized trial being the same as the duration of a single period in the cluster design with baseline measure, then $\rho_s=\rho_{wp}$ is likely to be a reasonable assumption.

Again we see that the rate of contamination that can be tolerated depends on the inflation in sample size for clustering, the multiple period aspects of the design, and on the increase in precision that the individually randomized design might afford due to stratification here.