Design and analysis of three-arm parallel cluster randomized trials with small numbers of clusters

Watson, Samuel; Girling, Alan; Hemming, Karla

DOI:
10.1002/sim.8828

Document Version
Peer reviewed version

Citation for published version (Harvard):
Watson, S, Girling, A & Hemming, K 2021, 'Design and analysis of three-arm parallel cluster randomized trials with small numbers of clusters', *Statistics in Medicine*, vol. 40, no. 5, pp. 1133-1146. https://doi.org/10.1002/sim.8828

ORIGINAL ARTICLE

Journal Section

Design and analysis of three-arm parallel cluster randomised trials with small numbers of clusters

Samuel I. Watson[1] | **Alan Girling**[1] | **Karla Hemming**[1]

[1]Institute of Applied Health Research, University of Birmingham, B15 2TT, United Kingdom

Correspondence
Samuel I. Watson PhD, Institute of Applied Health Research, University of Birmingham, Birmingham, United Kingdom
Email: s.i.watson@bham.ac.uk

Funding information
SW is funded by the NIHR Global Health Research Unit on Improving Health in Slums and NIHR Applied Research Centre (ARC) West Midlands. KH is funded by a NIHR Senior Research Fellowship SRF-2017-10-002.

In this article we review and evaluate a number of methods used in the design and analysis of small three-arm parallel cluster randomised trials. We conduct a simulation-based study to evaluate restricted randomisation methods including covariate-constrained randomisation and a novel method for matched-group cluster randomisation. We also evaluate the appropriate modelling of the data and small sample inferential methods for a variety of treatment effects relevant to three-arm trials. Our results indicate that small-sample corrections are required for high (0.05) but not low (0.001) values of the intraclass correlation coefficient and their performance can depend on trial design, number of clusters, and the nature of the hypothesis being tested. The Satterthwaite correction generally performed best at an ICC of 0.05 with a nominal type I error rate for single-period trials, and in trials with repeated measures type I error rates were between 0.04 and 0.06. Restricted randomisation methods produce little benefit in trials with repeated measures but in trials with single post-intervention design can provide relatively large gains in power when compared to the most unbalanced possible allocations. Matched-group randomisation improves power but is not as effective as covariate-constrained randomisation. For model-based analysis, adjusting for fewer covariates than were used in

Abbreviations: cRCT, cluster randomised controlled trial; LMM, linear mixed model; ICC, intra-class correlation coefficient

a restricted randomisation process under any design can produce non-nominal type I error rates and reductions in power. Where comparisons to two-arm cluster trials are possible, the performance of the methods are qualitatively very similar.

KEYWORDS
Cluster randomised controlled trial, covariate-constrained randomisation, matching, power, sample size

1 | INTRODUCTION

Cluster randomised controlled trials (cRCT) are a widely-used method to evaluate the effect of interventions applied to groups of individuals such as clinics, hospitals, schools, or villages [1, 2]. Cluster trials are useful when treatments are targeted at 'higher-level' processes rather than the individual, or when it is not possible to avoid interaction between individuals within the same cluster. Many cluster trials include only include a small number of clusters as it can be difficult or impractical to recruit many [3]. Schools and hospitals, for example, are large, complex organisations, and enlisting them into a trial can be an expensive and time-consuming process [4]. There is no standard definition of 'small' for the number of clusters in this setting, but, for an approximate sense of scale around fewer than ten to fifteen clusters per arm is often used in analyses of 'small' cluster trials (e.g. [3, 4, 5, 6]).

A small number of clusters can present issues for the design and analysis of a cRCT [7]. In particular, an imbalance of cluster-level prognostic factors, which we will call *covariates* from here on, between trial arms can lead to low power [8, 6]. And under small sample sizes methods of inference can fail to have the expected statistical properties [5]. Restricted randomisation methods that improve covariate balance can produce greater power but require appropriate model-based analysis and it can be unclear how these interact with any small-sample corrections that are used for statistical inference [6, 9].

There are various types of cRCT designs; in this article we focus on *multi-arm* parallel cluster trials. There are relatively few published examples of multi-arm cRCTs but they are growing in number. They have evaluated either different combinations of interventions against usual care, or different 'doses' of a single type of intervention; they have ranged in size from fewer than ten clusters per arm to over 100; they have included both continuous and binary outcomes; and generally, but not exclusively, include repeated measures. One recent major trial enrolled over 700 clusters into seven trial arms to evaluate the effects of different combinations of water, sanitation, hygiene, and nutrition interventions on the risk of childhood diarrhoeal and respiratory illness in Bangladesh where the presence of disease was recorded post-intervention [10, 11]. Smaller examples include a three-arm trial with 11 clusters per arm of an intervention to reduce time spent sitting at work with baseline and post-treatment measures of sitting time [12]. We have conducted a four-arm trial of different levels of incentives for small businesses to follow a healthy workplace initiative that included both continuous and binary outcomes measured pre- and post-intervention [13]. Other examples include a three-arm cRCT of different clinical management tools for knee osteoarthritis that enrolled 87 clusters, with a continuous outcome score measured at baseline and follow up [14], and an adaptive cluster trial of different levels of incentives for HIV self-testing in Malawi that started with six arms and six clusters per arm that measured a dichotomous outcome only in the post-intervention period [15]. There are evidently many possible variations to the design of multi-arm cRCTs, but little guidance around their design and analysis. This article considers

three-arm cRCTs with a small number of clusters, with continuous outcomes, including single post-intervention and repeated cross-sectional and cohort designs.

Almost all of the evaluation of methods for the design and analysis of small cRCTs has been for two-arm trials. The analysis of multi-arm trials is not as clear-cut as for a two-arm study as there are multiple possible pairwise and joint comparisons that may be of interest [16]. Unlike for simple null hypotheses with one linear restriction, small-sample corrections for test statistics of null hypotheses with multiple linear restriction depend on the whole internal random structure of the covariance matrix of the estimated parameters and the non-linear dependency of the reference distribution on the number of restrictions [17]. Different corrections handle this problem in different ways, which may result in differing performance. Furthermore, their use in trials with repeated measures has not been previously analysed for any number of trial arms. The performance of methods that balance covariates, such as restricted randomisation, is not known. Analyses of these randomisation methods have not previously incorporated trials with repeated measures into their assessments (e.g. [18, 6]) for any number of trial arms. Our objective in this article is to extend and evaluate analysis methods for multi-arm trials.

1.1 | Covariate Imbalance

Trials that use simple randomisation methods with small numbers of randomisation units could end up with a severe imbalance of cluster-level covariates between trial arms [3]. These covariates could include demographics or socioeconomic characteristics of its catchment population, or other factors that may confound intervention effects. While imbalance is not a problem *per se* for causal inference from cRCTs, it might result in lower power than a randomisation method that can ensure a certain level of balance. For cRCTs where all clusters are recruited prior to the start of the trial, a number of restricted randomisation methods have been developed for two-arm trials to balance important prognostic characteristics while maintaining an ignorable treatment allocation [3], which we collectively refer to as 'balancing randomisation'. This includes covariate-constrained randomisation [8, 6, 9] and matched-pairs cluster randomisation [19, 20]. Where we refer to 'imbalance' in the article, we mean an imbalance in cluster-level covariates.

Constrained randomisation follows the steps: (i) enumerate a large number of possible randomisation schemes, or all of them if the number of clusters is small enough; (ii) remove duplicate schemes; (iii) assess the balance in each enumerated scheme according to some balance score; and (iv) randomly select a scheme from the top q% of schemes, where $q = 10$ is often used. A commonly used balance score for two-arm trials was proposed by Raab and Butcher, which is a sum of standardised mean differences between the two arms [8], but could include any statistic that compares the distance between the distributions of the covariate(s) of interest in the different trial arms. For example, Grischott [21] provides functionality for 15 different measures in an online constrained-randomisation app.

Li et al. [18, 6] evaluated covariate-constrained randomisation methods for both linear and generalised linear models, respectively, for small cluster trials. They identified a number of conclusions through a series of simulation-based studies. Sufficient model-based adjustment is required to ensure that tests have the nominal type I error rate; if not all of the covariates used in the randomisation are used in a model-based analysis then type I errors are conservative and can approach zero. They also show that constrained randomisation provides perhaps only a modest improvement in power compared to simple randomisation *on average*, which is most pronounced when the randomisation space is constricted to the most balanced 10% or less of permutations. Little improvement in power was observed above eight clusters per arm.

Matched-pair cluster randomisation is less widely used than covariate-constrained randomisation as means to ensure covariate balance in cRCTs [19]. Clusters are paired on the basis of their similarity of observed characteristics, and then randomised within pairs to either treatment or control conditions. Some authors have questioned the effi-

ciency of such a randomisation process, and some have recommended against using it altogether (see discussion in Imai et al. [19]). However, in-depth analyses of this design, such as by Imai et al. [19] and Wu et al. [20], have shown that under the right conditions the matched-pair cluster randomisation can improve efficiency for small trials.

Little analysis exists of balancing randomisation methods in the multi-arm cluster trial setting and there does not exist any extension and evaluation of balancing randomisation methods involving matching to a multi-arm setting. Ciolino et al. [22] consider covariate-constrained randomisation in a multi-arm trial setting and demonstrate that it does indeed improve balance between trial arms, but the consequences of this on inferences are not explored.

1.2 | Inference

Balancing randomisation methods reduce the sampling variance of treatment effect estimators by causing a correlation in outcomes between trial arms, since only allocations with a similar covariate distribution in each arm are allowed (see C.1 Supplementary Information as well as Kahan and Morris [23], Wu et al. [20]). However, treatment effect estimators assume independence between trial arms. The correlation between arms is determined by the covariates used in the balancing procedure and any estimators of the variance of treatment effects that do not adjust for these covariates will be biased upwards, i.e. the standard errors of treatment effects will be too large and the type I error rate will not be nominal, which has been demonstrated for two-arm trials [6, 9, 19]. However, the estimator of the treatment effect itself remains unbiased.

Treatment effects are typically estimated from a cRCT using a generalised linear mixed model. However, when the number of clusters is small, test statistics do not have the standard (asymptotic) reference distributions. Use of uncorrected test statistics leads to inflated type I error rates and exaggerated power [4]. A number of small-sample corrections have been proposed in the literature that determine the appropriate (denominator) degrees of freedom for the F distributions of test statistics. The Satterthwaite correction approximates the degrees of freedom based on the first two moments of the parameter estimate [24], and the correction proposed by Kenward and Roger [17] takes Satterthwaite's correction after applying a scale factor based on a small-sample estimate of the covariance matrix to the Wald statistic. In addition, there is also the 'between-within' correction, which sets the number of degrees of freedom for the the test statistic to the number of degrees of freedom at the cluster level. The performance of these corrections may depend on the level of the intraclass correlation coefficient (ICC), which is the proportion of the total variance at the cluster level. Using simulated data, Leyrat et al. [5] found that for very small two-arm cRCTs (less than around six clusters per arm) with an ICC of 0.05, the type I error rate using the the Kenward-Roger correction was too conservative, the Satterthwaite correction anti-conservative, and the 'between-within' correction performed well. However, for low ICC (0.001) uncorrected and the Satterthwaite correction, in some settings, performed nominally, while the others (between-within and Kenward-Roger) were too conservative. Generally Kenward-Roger and Satterthwaite corrections have been found to perform similarly, although evidence from other evaluations suggests the Kenward-Roger correction is more robust to different structures of the covariance matrix (e.g. Luke [25], Arnau et al. [26]).

1.3 | Objectives

The aim of this article is to examine and extend methods for small cRCTs to a multi-arm setting. In particular, using simulated data we examine the power and type I error of different estimators of treatment effects in various small multi-arm cRCT designs, including cross-sectional and repeated measures designs, under both covariate-constrained randomisation and matched-pairs cluster randomisation. We consider the number of covariates used in the randomi-

sation and adjustment, small sample corrections, and the size of the trial among other aspects of the design and analysis. The rest of the article is structured as follows: Section 2 describes our multi-arm trial framework including notation and null hypotheses, randomisation methods, and methods for the simulation study. Section 3 provides the results from the simulation study. Section 4 discusses the results and concludes.

2 | METHODS

2.1 | Multi-arm cluster trial notation and models

We conducted a series of simulation studies to evaluate the design choices and performance of randomisation and analysis methods in the small multi-arm parallel cRCT setting. We examine three-arm trials. We follow the simulation described in Li et al. [6] where possible.

We focus on three parallel designs of a multi-arm cluster randomised trial: one with only a single post-intervention observation in each cluster; one with two cross-sectional observations per cluster, one pre- and the other post-intervention, a repeated cross-sectional design; and a cohort design with the same participants observed in two time periods, pre- and post-intervention. We refer to these designs as the 'post', 'repeated cross-section', and 'cohort' designs, respectively. The choice over the design is often pragmatic; while a multi-period design can afford greater power than a single period design, time and budget may prohibit two rounds of data collection, for example.

We consider a three arm trial with two treatments, so that one arm receives the current 'standard of care' and acts as a control. There are J clusters in total recruited to the trial $j = 1, ..., J$ that are randomised to the three trial arms at a 1:1:1 ratio, $g = 1$, $g = 2$ or $g = 3$, where an 'arm' is a set of clusters all randomly assigned to the same treatment or treatment schedule, and one or two periods of observation: $t = 1$ or $t = 1, 2$. We assume that each cluster has n_{jg} individuals where $\sum_{j=1}^{J} n_{jg} = n$. We explicitly notate the trial arm g to facilitate description of covariate balancing between arms later, where G is the total number of arms and T is the total number of time periods. The two treatments are indicated by dichotomous variables d_{1jgt} and d_{2jgt}, which equal one if cluster j has the treatment at time t and zero otherwise, and which contrast the two treatment conditions to the control condition. We observe S_x time-invariant pre-treatment covariates at the cluster-level $X_{jg} = [x_{1jg}, ..., x_{S_x jg}]'$, $j = 1, ..., J$. We observe R covariates at the individual-level $Z_{ijgt} = [z_{1ijgt}, ..., z_{Rijgt}]'$, $i = 1, ..., n$. We consider only a continuous outcome $y_{ijgt} \in \mathbb{R}$.

We focus on a model-based analysis (and data generating process) using a linear mixed effects model. For the cohort model:

$$y_{ijgt} = \mu + X'_{jg}\beta + Z'_{ijgt}\delta + \tau_t + \gamma_1 d_{1jgt} + \gamma_2 d_{2jgt} + \alpha_j + \alpha_{jt} + \alpha_i + u_{ijgt} \tag{1}$$

where $\alpha_j \sim N(0, \sigma^2_{\alpha 1})$ is a between-cluster random effect, $\alpha_{jt} \sim N(0, \sigma^2_{\alpha 2})$ is a within-cluster between-period random effect, $\alpha_i \sim N(0, \sigma^2_{\alpha 3})$ is a between-individual random effect, τ_t is a time period fixed effect, and u_{ijgt} is an independent identically distributed $N(0, \sigma^2)$ error term. In the case of the repeated cross-section design the individual-level random effect α_i is dropped from the model. For the single period post design trials, τ_t and α_{jt} are dropped from the model (set to zero). β, δ, γ, and τ_t are vectors of parameters to be estimated. We also define the (within-period) intraclass correlation coefficient (ICC) as [27]:

$$ICC = \frac{\sigma^2_{\alpha 1} + \sigma^2_{\alpha 2}}{\sigma^2_{\alpha 1} + \sigma^2_{\alpha 2} + \sigma^2_{\alpha 3} + \sigma^2}$$

The cluster autocorrelation coefficient (CAC) is the correlation between cluster-level observations in different time periods and is:

$$CAC = \frac{\sigma^2_{\alpha 1}}{\sigma^2_{\alpha 1} + \sigma^2_{\alpha 2}}$$

and the individual autocorrelation coefficient (IAC), which described the correlation between observations from the same individual in different periods, is:

$$IAC = \frac{\sigma^2_{\alpha 3}}{\sigma^2_{\alpha 3} + \sigma^2}.$$

2.2 | Null hypotheses

There are a number of possible hypothesis tests of interest in the multi-arm trial. CONSORT, a widely accepted set of guidelines for randomised trials, identifies four types of comparison for a three-arm trial,[16] to quote directly:

1. Comparing all 3 groups at once (A vs B vs C); a global test of unordered groups or a test for trend across ordered groups.
2. Comparing 1 group to the other 2 groups combined (A plus B vs C) and then the groups that were combined to each other (A vs B); A and B might be low and high doses of the same drug and the first comparison could be of treated vs untreated, followed by a comparison of the 2 treated groups, or A and B might be 2 antibiotics in the same class vs C as a member of a different class (note: the labeling in this example is arbitrary).
3. All pairwise comparisons: A vs B, A vs C, and B vs C.
4. Comparing A vs C and B vs C, but not A vs B; for example, comparing 2 treatments, separately, to the control but not comparing the 2 treatments to each other.

Based on these different comparisons we specify a number of null hypotheses for the parameters of Equation (1), which are described in Table 1. The hypotheses are not exhaustive but represent the range of possible comparisons we examine in this article. We note also that we assume that for some hypotheses we assume that the value of 'free' treatment effect parameters does not affect the test and leave these unspecified as stated.

2.3 | Randomisation

2.3.1 | Covariate-constrained randomisation

A commonly used balance score is that proposed by Raab and Butcher [8], which is a weighted sum of (standardised) mean differences of covariates. Higher values of this score indicate greater imbalance. The score is motivated by considering that cluster trial analyses based on a linear model are equivalent to a sum of mean differences. In a two-arm trial the score is:

$$\sum_{l=1}^{S} \frac{1}{\sigma^2_{x_l}} (\bar{x}_{l.2} - \bar{x}_{l.1})^2 \tag{2}$$

Comparison	**CONSORT**	Description	H_0	H_1
Pairwise comparison (A vs B, B vs C, etc)	3,4	*a*) {treatment 1} = {controls}, {treatment 2} unspecified *versus* {treatment 1} ≠ {controls}, {treatment 2} unspecified	$\gamma_1 = 0$	$\gamma_1 \neq 0$
Pairwise comparison (A vs B, B vs C, etc)	3,4	*b*) {treatment 2} = {controls}, {treatment 1} unspecified *versus* {treatment 2} ≠ {controls}, {treatment 1} unspecified	$\gamma_2 = 0$	$\gamma_2 \neq 0$
Pairwise comparison (A vs B, B vs C, etc)	2,3	*c*) {treatment 1} = {treatment 2}, {controls} unspecified *versus* {treatment 1} ≠ {treatment 2}, {controls} unspecified	$\gamma_1 = \gamma_2$	$\gamma_1 \neq \gamma_2$
Comparing all three groups (A vs B vs C)	1	*d*) {treatment 1} = {treatment 2} = {controls} *versus* {treatment 1} ≠ {treatment 2} or {treatment 1} ≠ {controls} or {treatment 2} ≠ {controls}	$\gamma_1 = \gamma_2 = 0$	$\gamma_1 \neq 0$ or $\gamma_2 \neq 0$ or $\gamma_1 \neq \gamma_2$
Comparing 1 group to the other two combined (A plus B vs C)	2	*e*) {treatment 1 or treatment 2} = {controls} *versus* {treatment 1 or treatment 2} ≠ {controls}	$\gamma_1 + \gamma_2 = 0$	$\gamma_1 + \gamma_2 \neq 0$

TABLE 1 Null and alternative hypotheses for multi-group trials. $\{a\}$ is shorthand for "the group mean for those with *a*". CONSORT refers to the relevant comparison described by the CONSORT statement.

where S is the number of covariates used in the randomisation, $\sigma^2_{x_l} = Var(x_{ljg})$ and $\bar{x}_{l.g}$ is the mean value of x_{ljg} in arm g for covariate l. In the same way, other measures of the distance between two means, including quadratic distance, absolute or Manhattan distance, or maximum distance could be used. We propose a general multi-arm version of the score in Equation (2) based on the between-arm covariance matrix used to calculate MANOVA test statistics:

$$B = \sum_{g=1}^{G} (\bar{X}_{.g} - \bar{\bar{X}}_{..})(\bar{X}_{.g} - \bar{\bar{X}}_{..})' \tag{3}$$

where $\bar{X}_{.g} = [\bar{x}_{1.g}, ..., \bar{x}_{S.g}]'$ and $\bar{X}_{..} = [\bar{x}_{1..}, ..., \bar{x}_{S..}]$. The score we use is:

$$\text{tr}(B) \tag{4}$$

where the covariates are standardised, as this is the sum of cluster level mean differences.

2.3.2 | Matched-group cluster randomisation

There have been no proposed extensions of matched pairs randomisation to a multi-arm setting previously. Indeed, there has been little or no investigation of algorithms to create matched groups with a given number of clusters as members based on their similarity in terms of covariate values. We developed an algorithm to implement a matched-pairs randomisation in a multi-arm trial setting to evaluate whether it could equal or exceed efficiency gains produced by covariate-constrained randomisation. We propose an algorithm that will create $M = \lceil J/G \rceil$ matched groups. The aim of the algorithm is to group clusters into M groups so that $tr(W)$ is minimised, where W is the within-groups covariance matrix as before but where the summation and means are with respect to the matched groups rather than trial arm. The intuition for this is that the closest matching minimises within-matched group variance — a perfect matching would have zero variance within a matched group. The algorithm is shown as Algorithm 1 in Supplementary Information and is a brute-force type algorithm involving continually proposing group swaps, keeping them if they reduce the overall within matched-group variance, and doing this until no more swaps can be made. Each member of a group is then randomly allocated to a different trial arm in a way that ensures there is at most a difference of one in the numbers of cluster per arm.

2.4 | Simulation study

The data generating process for the simulation study follows the model described in Equation (1). Three continuous covariates were simulated at the cluster level

$$x_{sjg} \sim N(0, 1^2), s = 1, 2, 3$$

all of which were used to generate the data. We simulated normal random effects for each cluster and individual:

$$\alpha_j \sim N(0, \sigma^2_{\alpha,1})$$
$$\alpha_{jt} \sim N(0, \sigma^2_{\alpha,2})$$
$$\alpha_i \sim N(0, \sigma^2_{\alpha,3})$$

and then at the individual-level four covariates were simulated:

$$z_{mijgt} \sim N(\mu_{zm}, 2^2), m = 1, ..., 4$$
$$\mu_{zm} \sim Uniform(-2, 2)$$

We do not include any time period effects, so that the true parameter value for each time period is zero. The ICC was specified to be either 0.05 or 0.001. For an ICC of 0.05 and $\sigma = 2$ we have $\sigma^2_{\alpha 1} \approx 0.21$, for example. The cluster autocorrelation coefficient (CAC) and individual autocorrelation coefficient (IAC) were both specified to be 0.8. The coefficients $\beta = [\beta_1, ..., \beta_S]'$ were set at 2 for primary analyses, however this value may be considered high as up to approximately 20% of the variance in the outcome would be accounted for by cluster-level covariates. We also examined a scenario in which β were set at 0.5 so that the cluster-level covariates accounted for approximately 5% of the variance. In all simulations the $\delta = [\delta_1, ..., \delta_4]'$ coefficients were set at 2.

For each simulated data set we use simple randomisation, covariate-constrained randomisation with the score in Equation (2), and matched-group cluster randomisation using the new algorithm. We randomise in a 1:1:1 ratio. Following Li et al. [6], the candidate set size of covariate-constrained randomisation, i.e. the number of enumerated randomisation schemes, was set at 10,000. After enumerating schemes and scoring them we select from schemes in the the upper q quantile of the distribution of the balance score, q was set at 0.1 as Li et al. [6] found little evidence for improvement in power above this figure.

We varied the number of clusters per arm from 3 to 11. For the number of individuals per cluster we considered predominantly unequal cluster sizes. We set an arbitrary distribution of cluster sizes: one third of size 10, one third 25, and one third 40. In scenarios with equal cluster sizes, all clusters had 25 individuals. We do not use cluster size as a covariate to be balanced in the restricted randomisation. All simulated trial data sets had three arms and hence 9 to 33 clusters in total.

2.4.1 | Parameter estimation and small-sample inference

For each simulated data set, we test each of these hypotheses described in Table 1 using the appropriate parameter restrictions for the linear mixed model described in 1. To evaluate the type I error for each of the hypotheses, both γ_1 and γ_2 were set to zero. To evaluate the power we set γ_2 to one and left γ_1 at zero. For the single-period 'post' designs the terms α_{jt} and τ_t were dropped from the analysis model. The number of covariates included in X_{jg} for adjustment in the analysis model was S_x, which is varied for some analyses. All individual-level variables were included in the model.

We use restricted maximum likelihood (REML) (using the R package `lme4`) to estimate the treatment effects as it has shown better small sample properties than standard maximum likelihood estimators, especially with regards to the variance components, which is crucial for accurate inference [28]. Inference for treatment effects, i.e. the

parameter(s) γ in (1), is typically based on a Student's t-distribution. However, naive estimators of the variance of the model parameters are biased downwards with small numbers of clusters [17]. We compare the corrections proposed by Satterthwaite [24], Kenward and Roger [17] (Kenward-Roger), and the 'between-within' correction, along with an *F*-test with no correction. The general version of the 'between-within' correction sets the degrees of freedom correction to the number of degrees of freedom at the cluster level, which in the case of the multi-arm trial setting presented here is $J * T - S_x - (G-1) - (T-1) - 1$ where S_x is the number of cluster-level covariates included in the model-based analysis (specifically, if X is a matrix with all cluster varying covariates, including intercept, cluster-covariates, time dummies, and treatments, then the degrees of freedom is $J * T - \text{rank}(X)$). For a two-arm, post trial with no covariates this would be $J-2$, for example. We conducted 10,000 simulations for each set of simulation parameters. Code for the simulations is available from the authors upon request.

3 | RESULTS

3.1 | Small sample estimation and inference

We present results from the repeated cross-section design and highlight where differences between study designs occur; results from the post and cohort designs are reported in the Supplementary Information. Figure 1 reports the type I errors from the repeated cross-section design. We make a number of observations. In all designs and hypotheses, small sample corrections were not required for an ICC of 0.001, and generally produced non-nominal type I error rates. For an ICC of 0.05, the Satterthwaite correction generally had the best performance. For the repeated cross-section design, it was moderately conservative (type I error of 0.04 to 0.05). For the post design (Figure S2 Supplementary Information), all corrections had a nominal type I error above five clusters per arm, whereas for a cohort design (Figure S6 Supplementary Information) the Satterthwaite correction was moderately anti-conservative (0.05 to 0.06) but generally was the best performing. For the joint null hypothesis $H_0 : \gamma_1 = \gamma_2 = 0$ an uncorrected F-test had nominal type I error rates in post and repeated cross-section designs, but not for the other two types of null hypotheses for which performance was comparable for all corrections.

3.2 | Randomisation methods

The estimated power from the simulations is shown in Figure 2 where we have used the best performing correction from the previous section (Satterthwaite for ICC of 0.05 and uncorrected F-test for an ICC of 0.001), and where all three covariates used in the balanced randomisation procedures were included in model-based adjustment. Covariate-constrained randomisation performed the best in terms of power although there was little advantage in the repeated cross-section or cohort designs (Figure S7 Supplementary Information). The increase in power for the post design was relatively small (Figure S3 Supplementary Information). For example, for a pairwise comparison (e.g. $H_0 : \gamma_1 = 0$) at five clusters per arm and a post design, the power with simple randomisation was 51%, with matched-group cluster randomisation was 55%, and with covariate constrained randomisation 58%. For low ICC (0.001) the conclusions are similar, although the improvement in power was larger in magnitude. For example in the same design as before, the power under simple randomisation was 77%, under matched-group cluster randomisation 82%, and under covariate constrained randomisation 86%. Gains were small for the repeated cross-section design. With the same design parameters as before and an ICC of 0.05, the power for simple, covariate-constrained, and matched-group randomisation was 73%, 75%, and 75% respectively. There was little evidence of any difference in power for the cohort design.

We compared the balance scores for covariate-constrained and matched-group randomisation methods (Figure

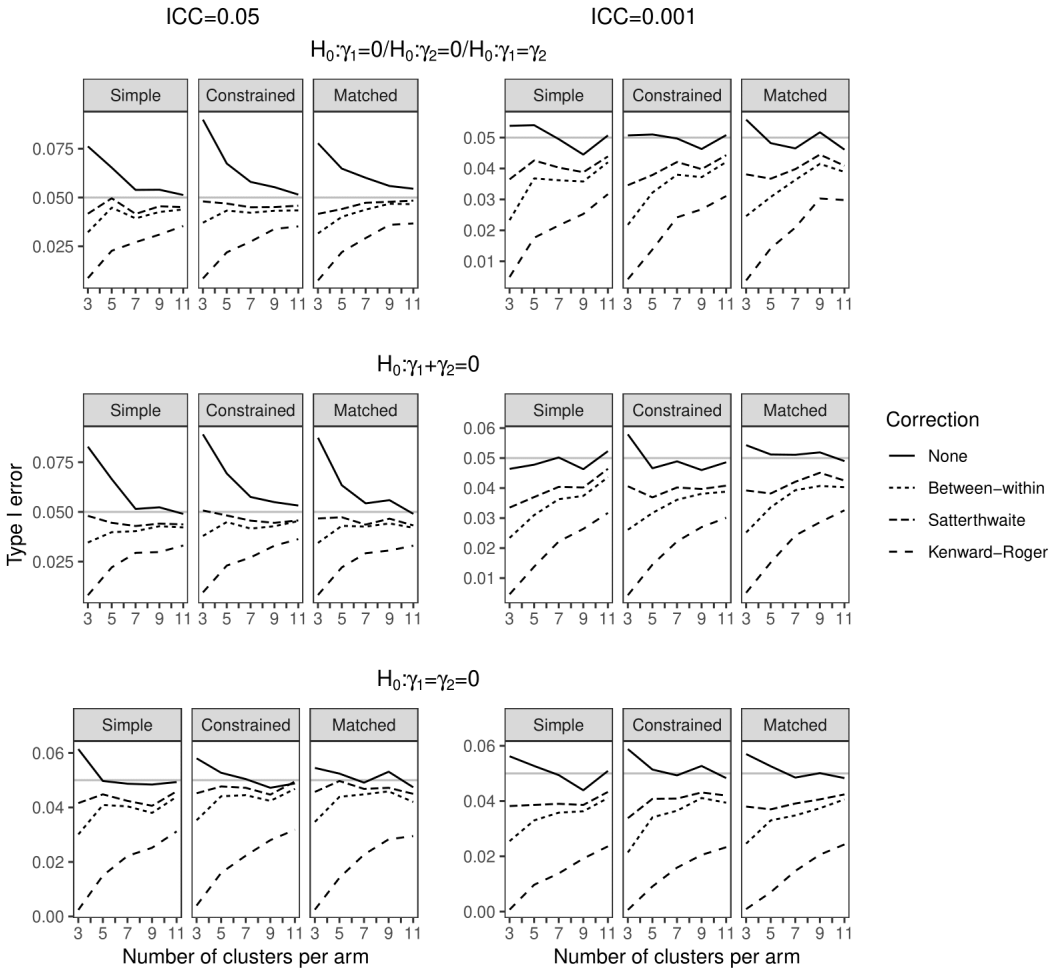

FIGURE 1 Type 1 error for repeated cross-sectional design with different small sample corrections (line types), randomisation methods, and numbers of clusters. Simulations for three-arm trial $G = 3$, unequal cluster sizes, number of covariates in adjustment $S_x = 3$, and number of covariates in randomisation $S = 3$.

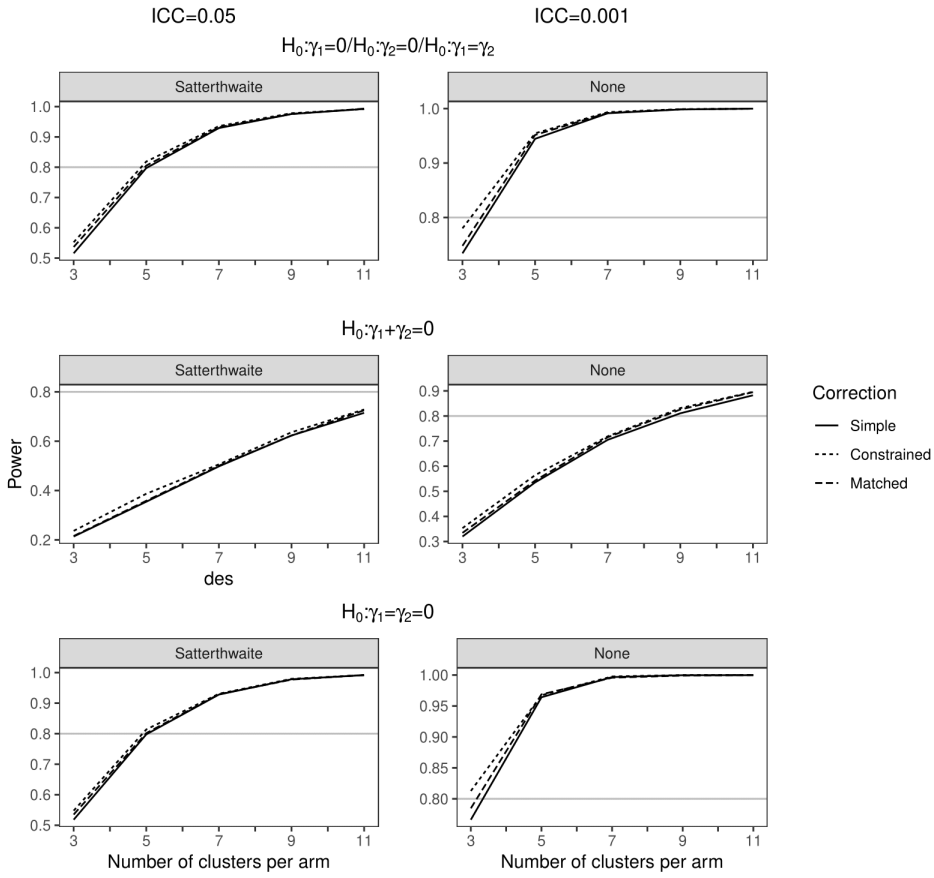

FIGURE 2 Power for different randomisation methods (line types) in the repeated cross-section design, and numbers of clusters. Simulations three-arm trial $G = 3$, unequal cluster sizes, number of covariates in adjustment $S_x = 3$, and number of covariates in randomisation $S = 3$. For ICC=0.05 a Satterthwaite corrected F-test was used and for ICC=0.001 an uncorrected F-test.

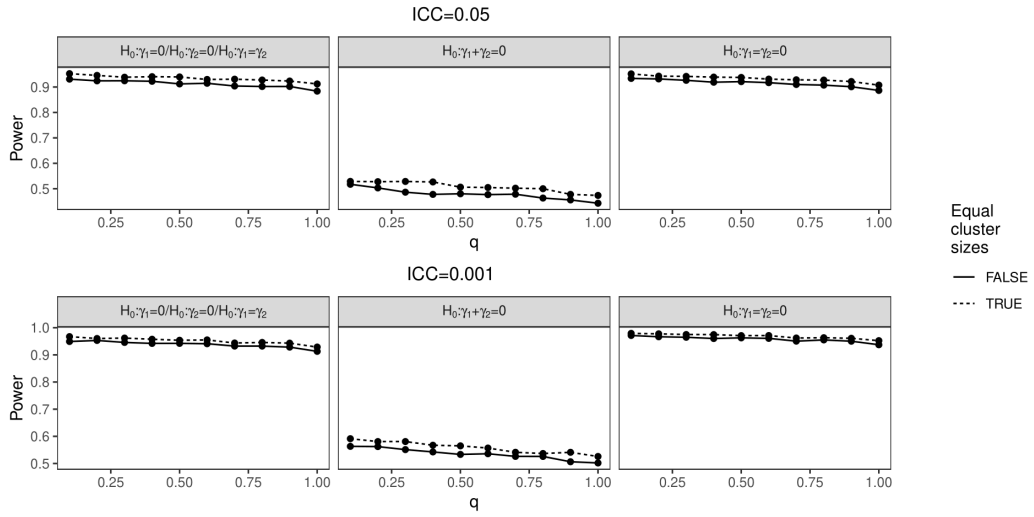

FIGURE 3 Attained power by decile of balance score (q) in an covariate-constrained randomisation. Simulations for $J = 15$ clusters, three-arm trial $G = 3$, number of covariates in adjustment $S_x = 3$, number of covariates in randomisation $S = 3$ and repeated cross-section study design. For ICC=0.05 a Satterthwaite corrected F-test was used and for ICC=0.001 an uncorrected F-test.

S1, Supplementary Information). Matched-group cluster randomisation improves the cluster-level balance of covariates compared to simple randomisation as assessed by the balance score, however there is still a high chance of relatively imbalanced allocations. Covariate-constrained randomisation more reliably produces balanced allocations and hence smaller sampling variation. Figure 3 shows the attained power for different levels of balance according to the balance score. Imbalance makes little difference to power for the repeated cross-section design. For an ICC of 0.001 the difference in attained power displays a similar pattern to the higher ICC. Figure 3 also provides equivalent results for an equal cluster size. As expected, equal cluster sizes result in greater power; the marginal increase in power by decile of balance score is approximately the same between designs equal and unequal cluster size. For a post design allocations in the bottom decile of balance have a power more than 15 percentage points worse than the top decile in this context (Figure S4 Supplementary Information). There was little change in attained power when the parameters β were set at 0.5 instead of 2: the differences in power between the top and bottom deciles for an ICC of 0.05 and covariate constrained randomisation for a post, repeated cross-section, and cohort designs, were, respectively 15, 4, and 0 percentage points.

3.3 | Model adjustment

The effect of different numbers of covariates in the randomisation process and in model-based adjustment for covariate-constrained randomisation and simple randomisation is shown in Figure 4 (for a fixed three covariates in the data generating process). When the covariates used in covariate-constrained randomisation were not all used in model-based adjustment then the type I error rate was below 5% for both levels of ICC. For the repeated cross-section design, type I errors were close to nominal (reflecting the patterns in Figure 1) when the same number or more covariates were used in adjustment as the covariate-constrained randomisation for an ICC of 0.05. When fewer covariates were used

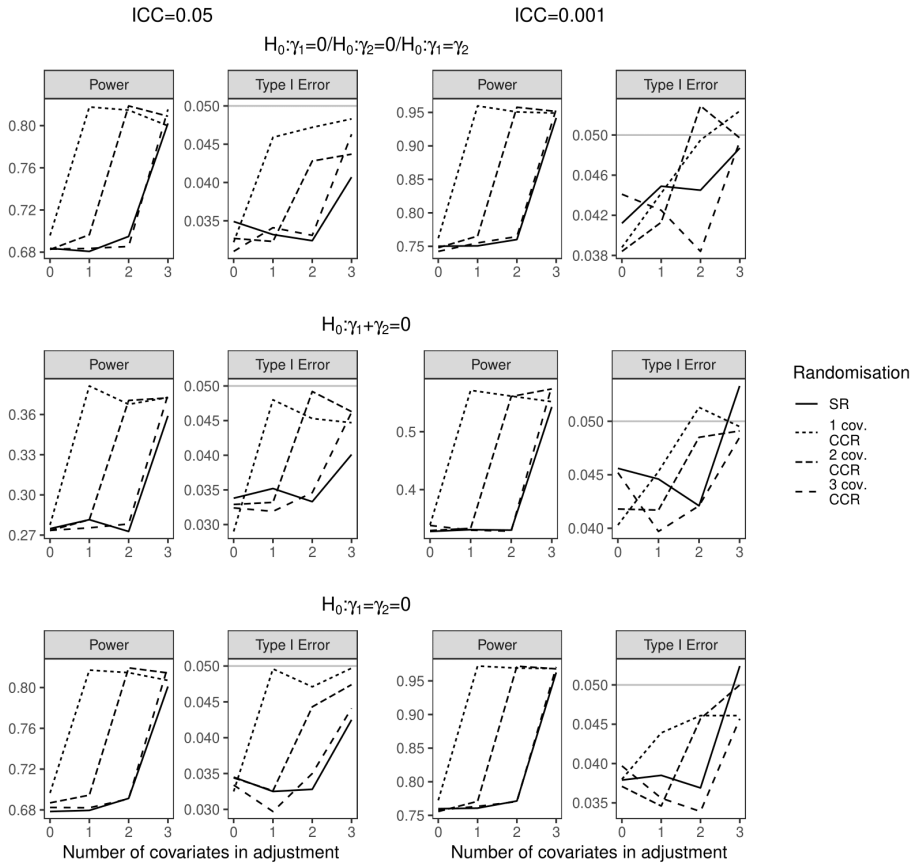

FIGURE 4 Power and type I error with different numbers of covariates used in covariate-constrained randomisation and model adjustment for repeated cross-section design. SR = simple randomisation, *a* cov. CCR = covariate constrained randomisation with *a* covariates. The data generating process used three cluster-level covariates. Simulations for $J = 15$ clusters, three-arm trial $G = 3$, unequal cluster sizes. For ICC=0.05 a Satterthwaite corrected F-test was used and for ICC=0.001 an uncorrected F-test.

for model adjustment than were included in the *data generating process* and the ICC was 0.001 or simple randomisation was used, type I errors were below nominal rates for both repeated cross-section and cohort designs (Figure S9 Supplementary Information), but not the post design. We note that not adjusting for cluster-level covariates has the effect of inflating the ICC, so the effect is likely due to a lack of small sample correction here. Power reflected the pattern of type I error rates: there were reductions in power when fewer covariates were used in the model adjustment than were included in the covariate-constrained randomisation; reductions in power were more modest for the repeated cross-section and cohort designs than for the post design.

4 | DISCUSSION

Both covariate-constrained randomisation and matched-pair cluster randomisation improve the balance of covariates between trial arms in two arm parallel cluster randomised trials with small numbers of clusters [6, 18, 22]. We examined the performance of constrained-randomisation and matched group randomisation (a proposed multi-arm extension to matched-pair randomisation) for three-arm trials in terms of the type I error and power of treatment effect estimators. Covariate-constrained randomisation performed the best in terms of power. However, there was little benefit in terms of power in trials with repeated measures. Indeed, the sampling distribution of the treatment effect estimator was affected only marginally by cluster-level covariate imbalance in the repeated cross-section design even compared to an equivalently powered post design. Our explanation for this is that having baseline measurements ensures an equivalent distribution of covariates in *treatment* and *control* conditions thus effectively providing adjustment for cluster-level differences, even if there is imbalance between *arms*.

For small trials and a reasonable ICC (ICC=0.05) we demonstrated the necessity of a small sample correction to ensure the appropriate type I error rates of the estimators. However, for cluster trials with repeated measures, either of repeated cross-section or cohort designs, none of the small-sample corrections provided exactly nominal type I error rates in small samples, although they were approximately nominal (+/- <0.01). The between-within and Satterthwaite corrections both performed reasonably, although for very small trials (fewer than five clusters per arm) our evidence suggests only the Satterthwaite correction has close to nominal type I error rates. However, for very low ICC values (ICC=0.001), the uncorrected test statistic was preferred and small-sample corrections generally produced non-nominal type I error rates. This reflects the findings of Leyrat et al. [5], who found for post design trials at very low values of the ICC, uncorrected analyses performed close to nominally in a mixed model setting. They also found evidence the Satterthwaite correction performed close to nominally, while the results in this article suggest it is moderately conservative for very low ICCs and in repeated measures designs. The performance of the corrections also depended to some extent on the hypothesis being tested. For joint null hypotheses comparing all treatment groups at once to the control, we found the uncorrected F-test to perform well and the between-within correction to be more conservative, with a high ICC.

We also showed that if the covariates used in the constrained randomisation procedure were not used in the analysis the standard errors of the treatment effect estimators would likely be biased, the type I error rate would likely be significantly conservative, which would also result in significant drop-offs in power. This is consistent with the findings of Li et al. [6] for two-arm trials and is due to the uncorrected correlation induced by the constrained randomisation procedure (see Supplementary Information). We also noted that if (strongly prognostic) cluster-level covariates that were in the data generating process were not adjusted for in analyses of repeated measures cluster trials then error rates were not nominal, which was explained as this would inflate the ICC. It is unknown whether the small sample correction would be as effective at even higher levels of the ICC.

Given the only moderate benefit in terms of power and potentially severe consequences of model mis-specification, one might ask whether using a constrained randomisation procedure is warranted even for the post design. For smaller trials constrained randomisation may be worth it, although much of the gain in power may come from adjustment using these covariates rather than the constrained-randomisation itself. As Lin [29] discusses at length, an estimator of a treatment effect from an adjusted model will generally be at least as efficient as one from an unadjusted model. However, comparisons of balancing randomisation to simple randomisation are typically *on average* [6, 18]; we showed that a simple randomisation procedure can produce severely imbalanced allocations with large reductions in 'attained power' [30] (i.e. larger standard errors) for a trial with only post-intervention measures. Constrained randomisation (and matched-group randomisation) removes the risk of generating a highly imbalanced allocation in this context,

which may significantly undermine the power of a trial. An additional benefit to constrained randomisation may also come from an improvement in the face validity of the randomisation procedure. More balanced trial arms at baseline may be improve confidence in reported results from cluster trials.

Multi-arm cluster trials raise the question of corrections for multiple testing as there is more than one treatment effect. However, there is no clear consensus on whether this required or not. Wason et al. [31] estimate around half of individual-level clinical trials do correct for multiple testing and half do not. Multiplicity adjustments are likely to have a greater effect when the treatments are independent than when they are correlated (for example, different doses of the same treatment) [31]. However, multiplicity corrections still require the underlying statistical tests to have the appropriate error rates. Our results are therefore important to identifying these tests in the context of multi-arm cluster trials, but we leave the question of multiplicity adjustments to future research.

4.1 | Limitations

We acknowledge a number of limitations to the work and the scope of the scenarios we investigated. We did not examine directly the effect of varying q, setting it at 10% for all simulations. This was based on Li et al. [6]'s conclusions. We did show variations in power for different deciles of the distribution of the balance statistic. Further improvements could be improved by focusing on the top 5% or 1% of allocations, but we suggest they would likely be small and for small numbers of clusters this may result in candidate set of allocations with only a small number of members. There are also a variety of other scores that could be used for assessing balance for covariated-constrained randomisation, although there is no obvious way to generalise many of these to a multi-arm multivariate setting. MANOVA test statistics are an exception and are based on the eigenvalues of $W^{-1}B$, where W is the within-group covariance matrix. For example Wilks' Λ is $\prod_{s=1}^{S} \frac{1}{1+\lambda_s} = \frac{det(W)}{det(W+B)}$ where λ_s are the eigenvalues. The reason for this is that these functions of the eigenvalues of $W^{-1}B$ are ratios of between sums of squares to within (or total) sums of squares for particular discriminant function variates of the covariates, and thus replicate ANOVA analyses in a multivariate setting [32]. However, since our concern is only to assess the scale of between group variance rather than conduct any inference, information contained in W is somewhat redundant and MANOVA scores are unlikely to perform any better than the score we proposed here.

We did not compare model-based analyses with other means of estimating treatment effects with small numbers of clusters, in particular generalised estimating equations (GEE). Previous simulation studies have shown GEEs to perform no better than, and often worse than, model-based methods for small numbers of clusters [28, 18]. As such we opted to examine only model-based analyses. We also did not examine more complex multi-arm designs, such as the stepped-wedge design, in which different randomisation sequences can be considered separate 'arms'. However we expect when there are repeated measures that the conclusion relating to the repeated cross-section design here, namely that imbalance in cluster-level covariates between arms is unlikely to have much effect, will apply. Finally, one can examine the validity of a given randomisation procedure using a test of randomness [33]. This would be a useful means to check the validity of a given covariate-constrained randomisation procedure should one be used. Intuitively the Raab-Butcher score ought to perform best since the model based comparisons are in essence estimates of adjusted mean differences [8]. Finally, we note that the simulations did not examine trials with dichotomous outcomes, which is an important area for future investigation given some of the potential inferential issues we have identified in analyses with continuous outcomes.

4.2 | Conclusions and recommendations

Our findings echo the results of a number of other articles in other contexts and several conclusions and recommendations arise from the research presented in this article. First, balancing randomisation procedures can successfully be extended to the three-arm cluster trial. Second, restricted randomisation procedures may not produce much benefit in trials with repeated measures but can reduce potentially substantial losses in power in three-arm cluster trials with a single cross-sectional observation. Third, the choice of small-sample correction is dependent on the ICC, which is typically unknown *ex ante*. Moreover, estimates of the ICC are likely to be noisy or biased with small numbers of clusters, so future research is needed to identify best-practice in these scenarios. Caution is needed with inference from trials with small numbers of clusters where the ICC is unknown. Fourth, appropriate model-based adjustment is needed to ensure reliable inferences when using balancing randomisation methods.

conflict of interest

None declared.

references

[1] Murray DM. Design and Analysis of Group Randomised Trials. New York, NY: Oxford University Press Inc.; 1998.

[2] Eldridge S, Kerry S. A practical guide to cluster randomised trials in health services research. Chichester, UK: Wiley; 2012.

[3] Ivers NM, Halperin IJ, Barnsley J, Grimshaw JM, Shah BR, Tu K, et al., Allocation techniques for balance at baseline in cluster randomized trials: a methodological review; 2012.

[4] Kahan BC, Forbes G, Ali Y, Jairath V, Bremner S, Harhay MO, 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;.

[5] Leyrat C, Morgan KE, Leurent B, Kahan BC. Cluster randomized trials with a small number of clusters: Which analyses should be used? International Journal of Epidemiology 2018;47(1):321–331.

[6] Li F, Lokhnygina Y, Murray DM, Heagerty PJ, DeLong ER. An evaluation of constrained randomization for the design and analysis of group-randomized trials. Statistics in Medicine 2016 may;35(10):1565–1579. http://doi.wiley.com/10.1002/sim.6813.

[7] Taljaard M, Teerenstra S, Ivers NM, Fergusson DA, Substantial risks associated with few clusters in cluster randomized and stepped wedge designs; 2016.

[8] Raab GM, Butcher I. Balance in cluster randomized trials. Statistics in Medicine 2001;20(3):351–365.

[9] Li F, Morgan KL, Zaslavsky AM. Balancing Covariates via Propensity Score Weighting. Journal of the American Statistical Association 2018 jan;113(521):390–400. https://www.tandfonline.com/doi/full/10.1080/01621459.2016.1260466.

[10] Luby SP, Rahman M, Arnold BF, Unicomb L, Ashraf S, Winch PJ, et al. Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural Bangladesh: a cluster randomised controlled trial. The Lancet Global Health 2018 mar;6(3):e302–e315. https://linkinghub.elsevier.com/retrieve/pii/S2214109X17304904.

[11] Ashraf S, Islam M, Unicomb L, Rahman M, Winch PJ, Arnold BF, et al. Effect of Improved Water Quality, Sanitation, Hygiene and Nutrition Interventions on Respiratory Illness in Young Children in Rural Bangladesh: A Multi-Arm Cluster-Randomized Controlled Trial. The American Journal of Tropical Medicine and Hygiene 2020 may;102(5):1124–1130. http://www.ajtmh.org/content/journals/10.4269/ajtmh.19-0769.

[12] Edwardson CL, Biddle SJH, Clarke-Cornwell A, Clemes S, Davies MJ, Dunstan DW, et al. A three arm cluster randomised controlled trial to test the effectiveness and cost-effectiveness of the SMART Work & Life intervention for reducing daily sitting time in office workers: study protocol. BMC Public Health 2018 dec;18(1):1120. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-018-6017-1.

[13] Thrive at Work Wellbeing Programme Collaboration. Evaluation of a policy intervention to promote the health and wellbeing of workers in small and medium sized enterprises – a cluster randomised controlled trial. BMC Public Health 2019 dec;19(1):493. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-6582-y.

[14] Cagnin A, Choinière M, Bureau NJ, Durand M, Mezghani N, Gaudreault N, et al. A multi-arm cluster randomized clinical trial of the use of knee kinesiography in the management of osteoarthritis patients in a primary care setting. Postgraduate Medicine 2020 jan;132(1):91–101. https://www.tandfonline.com/doi/full/10.1080/00325481.2019.1665457.

[15] Choko AT, Corbett EL, Stallard N, Maheswaran H, Lepine A, Johnson CC, et al. HIV self-testing alone or with additional interventions, including financial incentives, and linkage to care or prevention among male partners of antenatal care clinic attendees in Malawi: An adaptive multi-arm, multi-stage cluster randomised trial. PLOS Medicine 2019 jan;16(1):e1002719. https://dx.plos.org/10.1371/journal.pmed.1002719.

[16] Juszczak E, Altman DG, Hopewell S, Schulz K. Reporting of Multi-Arm Parallel-Group Randomized Trials. JAMA 2019 apr;321(16):1610. http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2019.3087.

[17] Kenward MG, Roger JH. Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. Biometrics 1997 sep;53(3):983. https://www.jstor.org/stable/2533558?origin=crossref.

[18] Li F, Turner EL, Heagerty PJ, Vollmer WM, Delong ER, Murray DM. An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes. Statistics in Medicine 2017;(October 2016):3791–3806.

[19] Imai K, King G, Nall C. The essential role of pair matching in cluster-randomized experiments, with application to the Mexican Universal Health Insurance Evaluation. Statistical Science 2009;24(1):29–53.

[20] Wu Z, Frangakis CE, Louis TA, Scharfstein DO. Estimation of treatment effects in matched-pair cluster randomized trials by calibrating covariate imbalance between clusters. Biometrics 2014;70(4):1014–1022.

[21] Grischott T. The Shiny Balancer - software and imbalance criteria for optimally balanced treatment allocation in small RCTs and cRCTs. BMC Medical Research Methodology 2018 dec;18(1):108. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0551-5.

[22] Ciolino JD, Diebold A, Jensen JK, Rouleau GW, Koloms KK, Tandon D. Choosing an imbalance metric for covariate-constrained randomization in multiple-arm cluster-randomized trials. Trials 2019;20(1):1–10.

[23] Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. Statistics in Medicine 2012;31(4):328–340.

[24] Satterthwaite FE. An Approximate Distribution of Estimates of Variance Components. Biometrics Bulletin 1946;.

[25] Luke SG. Evaluating significance in linear mixed-effects models in R. Behavior Research Methods 2017;.

[26] Arnau J, Bono R, Vallejo G. Analyzing small samples of repeated measures data with the mixed-model adjusted F test. Communications in Statistics: Simulation and Computation 2009;.

[27] Hemming K, Lilford R, Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. Statistics in Medicine 2015 jan;34(2):181–196. http://doi.wiley.com/10.1002/sim.6325.

[28] McNeish D, Stapleton LM. Modeling Clustered Data with Very Few Clusters. Multivariate Behavioral Research 2016;51(4):495–518. http://dx.doi.org/10.1080/00273171.2016.1167008.

[29] Lin W. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. Annals of Applied Statistics 2013;7(1):295–318.

[30] Wong H, Ouyang Y, Karim ME. The randomization-induced risk of a trial failing to attain its target power: Assessment and mitigation. Trials 2019;20(1):1–8.

[31] Wason JMS, Stecher L, Mander AP. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? Trials 2014 dec;15(1):364. https://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-15-364.

[32] Bray JH, Maxwell SE. Omnibus MANOVA Tests. In: Multivariate Analysis of Variance 2455 Teller Road, Newbury Park California 91320 United States of America: SAGE Publications, Inc.;.p. 14–39. http://methods.sagepub.com/book/multivariate-analysis-of-variance/n2.xml.

[33] Moulton LH, Golub JE, Durovni B, Cavalcante SC, Pacheco AG, Saraceni V, et al. Statistical design of THRio: A phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. Clinical Trials 2007;.