Leyrat, C; Morgan, KE; Leurent, B; Kahan, BC (2017) Cluster randomized trials with a small number of clusters: which analyses should be used? International journal of epidemiology. ISSN 0300-5771 DOI: https://doi.org/10.1093/ije/dyx169

DOI: 10.1093/ije/dyx169

Cluster randomised trials with a small number of clusters: which analyses should be used?

Clémence Leyrat[1,2*], Katy E Morgan[1], Baptiste Leurent[1], Brennan C Kahan[3]

* Correspondence to Clemence Leyrat: clemence.leyrat@lshtm.ac.uk
1 Department of medical statistics, London School of Hygiene and Tropical Medicine, London, UK
2 INSERM CIC 1415, CHRU de Tours, Tours, France
3 Pragmatic Clinical Trials Unit, Queen Mary University of London, London, UK

Abstract

Background: Cluster randomised trials (CRTs) are increasingly used to assess the effectiveness of health interventions. Three main analysis approaches are: cluster-level analyses, mixed-models and generalized estimating equations (GEEs). Mixed models and GEEs can lead to inflated type I error rates with a small number of clusters, and numerous small-sample corrections have been proposed to circumvent this problem. However, the impact of these methods on power is still unclear.

Methods: We performed a simulation study to assess the performance of 12 analysis approaches for CRTs with a continuous outcome and 40 or fewer clusters. These included weighted and unweighted cluster-level analyses, mixed-effects models with different degree-of-freedom corrections, and GEEs with and without a small-sample correction. We assessed these approaches across different values of the intraclass correlation coefficient (ICC), numbers of clusters, and variability in cluster sizes.

Results: Unweighted and variance-weighted cluster-level analysis, mixed models with degree-of-freedom corrections, and GEE with a small-sample correction all maintained the type I error rate at or below 5% across most scenarios, whereas uncorrected approaches lead to inflated type I error rates. However, these analyses had low power (below 50% in some scenarios) when fewer than 20 clusters were randomized, with none reaching the expected 80% power.

Conclusion: Small-sample corrections or variance-weighted cluster-level analyses are recommended for the analysis of continuous outcomes in CRTs with a small number of clusters. The use of these corrections should be incorporated into the sample size calculation to prevent studies from being underpowered.

Keywords: Cluster randomised trial, Small-sample correction, Degrees-of-freedom correction, Type I error rate, Power, Cluster-level analysis, Mixed-model, Generalized Estimating Equations.

Key messages

- Cluster randomised trials (CRTs) with few randomised clusters can lead to inflated type I error rates if they are analysed with standard methods

- A cluster-level analysis weighted by the inverse of the variance, degrees-of-freedom corrections for mixed models or a small-sample correction for GEE can provide an appropriate type I error rate in most situations even with as few as 6-8 clusters

- These type I error corrections lead to a decrease in statistical power; therefore, an inflation of the sample size to account for these corrections is required to ensure trials are appropriately powered

- These analysis approaches, available in common statistical software packages, should be used to analyse CRTs with few clusters, and should be reported in publications to improve transparency and reproducibility

Introduction

Cluster randomised trials (CRTs) are trials in which groups of participants, rather than the participants themselves, are randomised (1). CRTs are commonly used in settings in which individual randomisation is impossible, for example when the intervention is applied at the cluster level, or when there is a risk of contamination between treatment groups (2). The observations within a cluster tend to be correlated, which is usually quantified by the intraclass correlation coefficient (ICC) (3). An appropriate analysis method needs to consider this correlation, as the type I error rate - the probability of finding a statistically significant difference when no true effect exists - will be inflated otherwise (4).

Two families of such methods exist. Cluster-level methods consist of estimating a summary measure of the outcome for each cluster, and then analyzing the summaries using standard methods for independent data. Individual-level methods consist of analyzing individual-level data using a method which takes the clustering into account, usually mixed-effects models (5) or generalized estimating equations (GEEs) (6). Individual-level analyses are used more frequently than cluster-level methods (7), as they typically lead to higher power and allow adjustment for covariates in a more straightforward way than cluster-level methods, which require a two-stage approach to adjust for individual-level covariates (8). However, individual-level analysis approaches can lead to an inflated type I error rate when the number of clusters is small. A recent review showed that 65% of CRTs were at risk of inflated type I error because no small-sample corrections were used despite a small number of clusters randomised (7). The minimum number of clusters required to maintain the type I error rate at 5% has been suggested to be around 30-40 clusters for mixed models and 40-50 for GEEs (1,9), although depending on specific trial characteristics, a larger number of clusters may be required. However, CRTs often involve less than 40 clusters: three reviews of CRTs found median numbers of 21 (10), 25 (7), and 34 (11) of clusters randomised. Therefore, many published CRTs may over report statistically significant intervention effects. Moreover, sample size formulae for CRTs assume a large number of clusters, and are based on the assumption that the statistical method used will have the correct type I error rate. Because this assumption is likely to be violated in trials with a small number of clusters, usual sample size calculation may not be appropriate. It is unclear how much of an impact this violation could have in practice, and whether classic sample size formulae can still be used for trials with a small number of clusters, or whether new formulae need to be developed for these situations.

Some small-sample corrections have been proposed to circumvent the problem of inflated type I error rate. For mixed-models, these involve degrees-of-freedom corrections, such as the one proposed by Satterthwaite (12) or Kenward and Roger (13). For GEEs, the corrections typically involve corrections to standard errors (14,15). Li and Redden assessed type I error rate and power for several small-sample corrections for mixed models (16) and GEEs (17) with binary outcomes, and Johnson *et al*. (18) assessed the type I error rate for cluster-level methods and several degrees-of-freedom corrections for mixed models. However, there is currently no guidance on which small-sample corrections can maintain both the nominal type I error rate and a reasonable power in CRTs with a continuous outcome. We therefore undertook a simulation study to determine which methods perform best in terms of type I error rate and power in CRTs with a small number of clusters and a continuous outcome, and we provide some guidance for the choice of the analysis method.

Methods

We performed a simulation study to compare the performance of 12 analysis methods used in practice for the analysis of continuous outcomes in CRTs with a small number of clusters. These 12 methods include 4 cluster-level methods, 5 corrections for mixed models and 3 corrections for GEE.

Analysis strategies

Cluster-level analyses

Unweighted cluster-level regression: Cluster-level analysis (1,8,19), also known as the two-stage approach (18), consists first of estimating a summary outcome measure by cluster (e.g. the mean) and then fitting a linear regression on these summary measures.

Weighted cluster-level regression: Weighted approaches have been proposed to improve efficiency when clusters are not of the same size (20). A common and straightforward approach is to weight each cluster by its sample size using $w_i = m_i$, where w_i is the weight for cluster *i (i = 1, ..., k)* and m_i the number of participants in cluster i. Another approach has been proposed, in which the weights are set to the inverse of the estimated theoretical variance of the clusters means (18,20):

$$w_i = \frac{1}{\sigma_b^2 + \frac{\sigma_w^2}{m_i}},$$

where σ_b^2 and σ_w^2 are the estimates of the between and within cluster variance, obtained using a one-way analysis of variance. Note that the three cluster-level linear regressions provide different treatment effect estimates and standard errors.

Wilcoxon rank-sum test: A non-parametric approach such as the Wilcoxon test may also be used to compare cluster means between intervention groups, as it does not require the assumption of normality of cluster-level summaries. The Wilcoxon test only provides the p-value and no estimate of the intervention effect.

Individual-level analysis

Individual-level analyses model individual data while taking the hierarchical structure of the data into account. The two main approaches used in practice are mixed models (5) and GEEs (6).

Mixed models: For continuous outcomes from CRTs from two-level CRTs, with patients nested in clusters, mixed models are typically linear regression models which include a random intercept for clusters (1). Inference for the treatment effect estimate is usually based on a Student's t-distribution, where the key issue is the determination of an appropriate number of degrees of freedom (Dof).

There are 5 main approaches described in the literature to determine the Dof for mixed models:

- Method 1 - "uncorrected": inferences are made using a normal distribution and therefore, no Dof computation is needed.

- Method 2 - “uncorrected-t”: the t-distribution with n-2 Dof, n being the total number of observations, is used instead of the normal distribution in Method 1.

Other methods have been proposed to estimate the Dof taking into account the number of clusters:

- Method 3 - “between-within” (following Li and Redden notation (15)): the Dof is defined as k-2 for CRTs where k is the number of clusters. This method assumes all the clusters are the same size.
- Method 4 - “Satterthwaite” (11): the Dof approximation is based on the first two moments of the parameter estimate and does not rely on the assumption of fixed cluster sizes, unlike Method 3.
- Method 5 - “Kenward-Roger” (12): this correction uses the Satterthwaite approximation after applying a scale factor based on a small-sample estimate of the covariance matrix to the Wald statistic (21).

We will assess the performance of these five methods based on mixed models using a restricted maximum likelihood (REML) estimator. Note that they provide the same intervention effect estimate and standard error but different confidence intervals and p-values. This is because the methods only impact the Dof used for the test of the intervention effect, except for Method 5 which also modifies the test statistic itself.

GEEs: The GEE approach aims to model the mean outcome to give a population-averaged intervention effect while treating the intraclass correlation as a nuisance parameter, instead of estimating cluster-specific effects as in the mixed models approach.

For GEEs, the model-based standard error (SE) estimator assumes the correlation structure to be correctly specified. In CRTs, the correlation structure is usually specified as exchangeable (assuming the same correlation within each cluster). A SE estimator that is robust to misspecification of the working correlation structure may also be used. This “sandwich” estimator relies on the observed between-cluster variability, but this variability is not well estimated if the number of clusters is too small. Therefore, small-sample corrections to the robust SE have been proposed (14,17). The standard errors from these GEE methods differ, but the intervention effect estimate is the same for each, and a normal approximation is used for the test of the intervention effect and the construction of its confidence interval.

Simulation study

A full description of the data generation process and the scenarios studied is given in Appendix. Briefly, data were generated for a two-arm parallel CRT with a continuous outcome and varying cluster sizes using R software. We considered different values for the number of clusters *k*, the ICC ρ, the coefficient of variation of cluster sizes *cv* and the average cluster size *m*. Average cluster sizes were chosen to be close to the lower and upper quartiles of cluster size found in a review of CRTs (7). Parameter values are listed in Table 1. The treatment effect was set to 0 to assess the type I error rate. To assess power, we set the treatment effect to give 80% power based on a standard CRT sample size formula (21). For each scenario, 5000 datasets were generated. Type I error rate and power were then defined as the proportion of these 5000 datasets in which the treatment effect estimates were statistically different from 0 at a 5% significance level, when the true treatment effect was null or positive, respectively.

Results

All the 12 methods compared in our simulation study lead to unbiased estimates of the treatment effect (bias < 0.0005, see supplementary material). The uncorrected mixed models based on the normal distribution ("uncorrected") and based on the t-distribution ("uncorrected-t") gave almost identical results since the DoF for the "uncorrected-t" method was always very high (total number of observations-2, which varies from 198 to 1198). We therefore only present the results for the "uncorrected" approach. Figures 1 and 2 show the power and type I error for two scenarios with ICC values of 0.001 and 0.05 respectively, a coefficient of variation of cluster sizes of 0.8 and the larger of the two mean cluster sizes considered. For GEEs, only the results obtained from convergent models are presented. GEEs converged between 59.8% and 99.9% of the time across the different scenarios – further details are given in the supplementary material. The pattern of results for different values of ICC, coefficient of variation for the cluster size and average cluster size for other scenarios are very like those seen in Figures 1 and 2. These results can be found in the supplementary material.

From 4 to 8 clusters

Cluster-level analyses

With 4 to 8 clusters, only the unweighted and the variance-weighted cluster-level methods maintained the nominal type I error rate of 5% across all scenarios. However, the power was usually low (between 10 and 65%). With fewer than 8 clusters, a Wilcoxon test cannot provide statistically significant results so both the type I error and the power were null.

Mixed models

Both the uncorrected and Satterthwaite approaches had an inflated type I error in some of the scenarios with higher ICC. The between-within and Kenward-Roger methods were too conservative, leading to type I errors lower than 2% in some scenarios, and resulting in a low power. Discrepancies between the Satterthwaite and Kenward-Roger approximations are partially explained by differences in the way the methods are implemented in R (details in supplementary material).

GEEs

GEEs without a small-sample correction always lead to a very high type I error rate (greater than 30% in some scenarios). When a small-sample correction was applied, the type I error rate was below 5% but this approach was generally too conservative when 8 or fewer clusters were randomised.

Although having 8 or fewer clusters may be common in practice (11 out of 78 in a recent published review (7)), our results suggest that the 80% nominal power was not reached among any of the methods which maintained an appropriate type I error rate.

From 10 to 20 clusters

Cluster-level analyses

From 10 to 20 clusters, the size-weighted analysis failed to maintain the type I error rate below 5% across all scenarios. The other approaches maintained an appropriate type I error. Wilcoxon had the lowest power and the variance-weighted method appeared to perform better in terms of power than the unweighted approach when the coefficient of variation of cluster sizes was high.

Mixed models

The uncorrected mixed model still lead to inflated type I error rates for between 10 and 20 clusters. The other approaches maintained an appropriate type I error in all scenarios. However, the power remained low with these approaches, typically between 60% and 75%. In some scenarios (low ICC and cluster size, high coefficient of variation) Satterthwaite outperformed between-within, which outperformed Kenward-Roger, but these differences grew smaller as the number of clusters increased.

GEEs

GEEs without a small-sample correction failed to maintain the type I error rate below 5%. Corrected GEEs generally had a conservative type I error, and a power close to that observed for Kenward-Roger (usually below the variance-weighted cluster-level and Satterthwaite mixed model).

From 30 to 40 clusters

Cluster-level analyses

With 30 or more clusters, all the cluster-level methods maintain a 5% type I error rate across all scenarios, except the size-weighted method when the ICC or coefficient of variation of cluster sizes were high. Unweighted cluster-level regression and the Wilcoxon test were often under-powered compared to individual-level methods, although a 5% type I error rate was maintained. Only variance-weighted cluster-level regression performed well both in terms of type I error and power.

Mixed models

The Satterthwaite, between-within and Kenward-Roger methods gave similar results for most scenarios (giving both the expected type I error rate and power), but Satterthwaite and between-within seemed to have slightly higher power than Kenward-Roger in some scenarios. Note that the power exceeds 80% for large ICCs because the sample size formula used is known to overestimate the sample size (22).

GEEs

Using a GEE without a small-sample correction lead to inflated type I error rates even with 30 or more clusters. The small-sample correction lead to a 5% type I error rate and a power close to that observed for the mixed models.

200 clusters

For comparison, one scenario with a very large number of clusters was explored. Results are given in the supplementary material. All methods except the size-weighted cluster-level analysis lead to appropriate type I error rates. Only the unweighted cluster-level analysis and the Wilcoxon test were underpowered. All the other approaches performed similarly.

Discussion

Many CRTs have a small number of clusters. Although it is known that small-sample corrections are often required to maintain nominal type I error rates, it is not known what impact these corrections may have on power. Our results confirmed that corrections are needed to maintain a correct type I error rate, but these corrections negatively impact the power of the trial, in some cases reducing power from the nominal 80% to 10%.

Our recommendations for which analysis approaches to use are shown in Table 2. The unweighted, variance-weighted and Wilcoxon cluster-level analyses, mixed-models with Satterthwaite, between-within and Kenward-Roger degree-of-freedom corrections, and GEEs with small-sample correction all had appropriate type I error rates across most scenarios. The method with the highest power varied across the different scenarios, and so should be chosen based on the study characteristics (e.g. number of clusters, anticipated cluster size, etc). Although our results share similarities with Li and Redden's (16,17) studies about small sample corrections for CRTs with a binary outcome, our results are valid for continuous outcomes only. Further research is needed to compare the power of the best corrections for cluster-level analyses, mixed models and GEEs when the outcome is binary.

Among the cluster-level analyses, the variance-weighted analysis performed best as the unweighted approach lead to a loss of efficiency when the number of clusters increased (8). The Wilcoxon test generally had lower power than the other analyses, although may be useful when the outcome cluster-level summaries are not normally distributed, which was not the case in our simulations which used a normal distribution.. A major limitation with cluster-level analyses is the difficulty of adjusting for baseline characteristics, which is a severe drawback given that one third of CRTs are at risk of selection bias (23). This is also an issue when only a small number of clusters are randomised, because chance imbalance is more likely to occur. Some methods have been proposed to adjust for covariates in cluster-level analysis, but individual-level analyses are generally simpler if baseline adjustment is required.

For mixed models, the Satterthwaite, Kenward-Roger, and between-within corrections all generally performed well in terms of type I error rate. GEEs with a small-sample correction also provided correct type I error rates across most settings. However, there were issues with convergence in some cases, especially for a very low number of clusters and a very low intraclass correlation.

Our results for the type I error rate are broadly similar to those from Johnson *et al.* (18). They did not assess the performance of the Satterthwaite approach, which performed well in our simulations, or GEE approaches. Although GEEs are often not recommended for a small number of randomised clusters (9,24), our simulation results suggest that they perform similarly to mixed models when a small-sample correction is used to correct the standard error.

Further investigation is required to assess the robustness of the small-sample corrections studied in this paper to model misspecification. All the methods assume that the cluster-level means are normally distributed. Further work is needed to assess the robustness of the small-sample corrections if the data do not match this assumption., which can be likely when only few clusters are randomised. For GEEs, an exchangeable working correlation matrix is often chosen, assuming the same correlation for each pair of individuals within a cluster, that is the same across all clusters. If the level of correlation varies from one cluster to another (25), this working matrix is no longer appropriate and the impact on the small-sample standard error is unknown. Similarly, if the correlation is not constant across clusters, the estimates of the within and between-cluster variance components for the weights used in the variance-weighted cluster-level analysis may be misleading.

Method performance varied with the ICC. In practice the ICC will be unknown when deciding the analysis method, and ICC estimates based on previous data can be imprecise. Therefore, even though an uncorrected mixed model can maintain a type I error rate close to 5% for a very small ICC, we do not recommend this approach in practice. We therefore encourage researchers to use a method that maintains an appropriate type I error rate across a range of possible ICC values when a small number of clusters are randomised. These corrections are now available in most standard statistical software packages and their implementation is straightforward. Table 3 gives the commands to implement the methods proposed in this paper in R, SAS and Stata. A brief overview is given also by Cook (26).

Among the methods with appropriate type I error, none lead to the expected power of 80%. Therefore, adjustments to the sample size are required in order to maintain the expected level of power. Because there is no sample size formula for such adjustments, we advise researchers to use simulations. Some crude methods have been proposed to correct the sample size when only few clusters are available, such as adding one cluster per arm (22,27). For a given number of clusters, increasing the cluster size may not always be sufficient to reach an 80% power, and thus the randomization of a larger number of clusters is preferable whenever feasible. CRTs with few clusters may only be appropriate for relatively large treatment effects.

In practice, when a small-sample correction is used, clear reporting is needed in the method sections of the trial's publications. The name of the method along with the software command is needed to ensure transparency and reproducibility of the results. Two recent reviews highlighted that small-sample corrections are not clearly reported in published CRTs (7,24).

In summary, we strongly recommend the use of appropriate methods for the analysis of continuous outcomes in CRTs when few clusters are randomised. The choice of the appropriate methods must be driven by the context of the study along with consideration of statistical arguments. These methods are now available in most statistical software packages. However, these corrections lead to a decrease in power so adjustment must be made in the sample size calculation.

Supplementary material

The full results for all studied scenarios are displayed in the supplementary material.

References

1. Donner A, Klar N. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.

2. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. Health Technol Assess Winch Engl. 1999;3(5):iii-92.

3. Eldridge SM, Ukoumunne OC, Carlin JB. The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. Int Stat Rev. 2009;77(3):378–94.

4. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health. 2004 Mar;94(3):423–32.

5. Hedeker D, Gibbons RD, Flay BR. Random-effects regression models for clustered data with an example from smoking prevention research. J Consult Clin Psychol. 1994;62(4):757–65.

6. Zeger SL, Liang K-Y, Albert PS. Models for Longitudinal Data: A Generalized Estimating Equation Approach. Biometrics. 1988 Dec;44(4):1049–60.

7. 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 Sep 6;17(1):438.

8. Eldridge S, Kerry S. A Practical Guide to Cluster Randomised Trials in Health Services Research. John Wiley & Sons; 2012. 299 p.

9. Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. BMJ. 2011 Sep 26;343(sep26 1):d5886–d5886.

10. Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. Clin Trials Lond Engl. 2004 Feb;1(1):80–90.

11. Satterthwaite FE. An Approximate Distribution of Estimates of Variance Components. Biom Bull. 1946;2(6):110–4.

12. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics. 1997 Sep;53(3):983–97.

13. Fay MP, Graubard BI. Small-Sample Adjustments for Wald-Type Tests Using Sandwich Estimators. Biometrics. 2001;57(4):1198–206.

14. Mancl LA, DeRouen TA. A Covariance Estimator for GEE with Improved Small-Sample Properties. Biometrics. 2001;57(1):126–134.

15. Li P, Redden DT. Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. BMC Med Res Methodol. 2015;15:38.

16. Li P, Redden DT. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. Stat Med. 2015 Jan 30;34(2):281–96.

17. Johnson JL, Kreidler SM, Catellier DJ, Murray DM, Muller KE, Glueck DH. Recommendations for choosing an analysis method that controls Type I error for unbalanced cluster sample designs with Gaussian outcomes. Stat Med. 2015 Nov 30;34(27):3531–45.

18. Hayes RJ, Moulton LH. Cluster Randomised Trials. Taylor & Francis; 2009. 338 p.

19. Campbell MJ, Walters SJ. How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research. John Wiley & Sons; 2014. 374 p.

20. Kerry SM, Bland JM. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. Stat Med. 2001 Feb 15;20(3):377–90.

21. PROC MIXED: Mixed Models Theory :: SAS/STAT(R) 9.2 User's Guide, Second Edition [Internet]. [cited 2016 Nov 10]. Available from: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mixed_sect022.htm

22. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. Int J Epidemiol. 2006 Jan 10;35(5):1292–300.

23. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. Int J Epidemiol. 2015 Jun;44(3):1051–67.

24. Brierley G, Brabyn S, Torgerson D, Watson J. Bias in recruitment to cluster randomized trials: a review of recent publications. J Eval Clin Pract. 2012 Aug;18(4):878–86.

25. Huang S, Fiero MH, Bell ML. Generalized estimating equations in cluster randomized trials with a small number of clusters: Review of practice and simulation study. Clin Trials. 2016 Apr 19;1740774516643498.

26. Fang X, Li J, Wong WK, Fu B. Detecting the violation of variance homogeneity in mixed models. Stat Methods Med Res. 2016 Dec 1;25(6):2506–20.

27. Cook A. Small-Sample Robust Variance Correction for Generalized Estimating Equations for Use in Cluster Randomized Clinical Trials [Internet]. 2015 [cited 2016 Nov 15]. Available from: https://www.nihcollaboratory.org/Products/Variance-correction-for-GEE_V1.R0.pdf

28. Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. Int J Epidemiol. 1999 Apr;28(2):319–26.

29. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. Clin Trials Lond Engl. 2005;2(2):99–107.

Tables

Table 1. List of parameters varied across scenarios (total number of scenarios = 90)

Parameter	Notation	Values	Source
Intraclass correlation coefficient	ρ	0.001,0.01,0.05	Campbell *et al.* (28)
Number of clusters randomised	k	4,6,8,10,20,30,40,200	Donner and Klar (1)

Average cluster size	m	m=50,300 when k = 4,6,8
		m=25,280 when k=10
		m=15,200 when k=20
		m=10,150 when k=30
		m=7,40 when k=40
		m=50 when k=200
			Kahan *et al.* (7)

Coefficient of variation of cluster sizes	cv	0.4,0.8	Eldridge *et al.* (21)

Table 2. Recommended* methods

Analysis strategy	**Number of clusters**		
	k=4-8	**k=10-20**	**k=30-40**
Cluster-level analyses	←	Variance-weighted	→
Mixed models	Between-within	Satterthwaite	Satterthwaite Between-within Kenward-Roger
GEEs	←	Small-sample correction	→

*The "recommended" method is that which tended to have the highest power across our simulations, among the methods maintaining a type I error rate below 5% in all scenarios evaluated. Other factors should also be considered for the choice of methods, such as the ICC value and the variability in cluster sizes. Note that different methods may perform equally well in several scenarios, and that the differences between methods tend to become negligible as the number of clusters increases.

Table 3. R, SAS and Stata commands to correct the type I error rate when the number of randomised clusters is small

		R	SAS	Stata
Weighted cluster-level analysis		*lm* function with the *weights* option. Weights can be computed using the between and within variance components from the package *ICC*	*proc glm* with the *weights* option. Weights can be computed using the between and within variance components estimated from a mixed model with *proc mixed*	*regress* command with *aweights* option. Weights can be computed using the between and within variance components estimated with the command *xtreg*
Mixed model	*Satterthwaite*	Mixed model with the *lmer* function from *lme4* package then *anova* command with the option *DDF="Satterthwaite"*. Package *lmerTest* required	*proc mixed* with the option *DDFM=SAT*	*mixed* command with the *dfmethod(satterthwaite)* option
	Kenward-Roger	Mixed model with the *lmer* function from *lme4* package then *anova* command with the option *DDF="Kenward-Roger"*. Package *pbkrtest* required	*proc mixed* with the option *DDFM=KR*	*mixed* command with the *dfmethod(kroger)* option
	Between-Within	Mixed model with the *lmer* function from *lme4* package then *anova* command to get the F value. Finally, the *pf* function to get the corresponding p-value with 1 and k-2 degrees of freedom	*proc mixed* with the option *DDFM=BW*	*mixed* command with the *dfmethod(repeated)* option
Small sample correction for GEE		*gee* function from the *gee* package then *saws* function applied on the gee output. Package *saws* required	*proc GLIMMIX* with the option *EMPIRICAL=FIROEEQ*	*xtgee* command with the *vce(bootstrap)* option to obtain correct standard errors with bootstrapping*

*Because small-sample corrections are not available yet in Stata

Figure 1: Power and type I error rate of the compared methods for the analysis of CRTs with a small number of clusters. The value of the intraclass correlation coefficient ρ for the outcome is 0.001. The average cluster size corresponds to the upper quartile of the distribution observed in a review of CRTs [7] with coefficient of variation for cluster size of 0.8. 5000 simulations were carried out per scenario.

Figure 2: Power and type I error rate of the compared methods for the analysis of CRTs with a small number of clusters. The value of the intraclass correlation coefficient ρ for the outcome is 0.05. The average cluster size corresponds to the upper quartile of the distribution observed in a review of CRTs [7] with coefficient of variation for cluster size of 0. 5000 simulations were carried out per scenario.

Appendix - Simulation plan

Data generation

We generated datasets corresponding to a two-arm parallel CRT with a continuous outcome and varying cluster sizes.

Cluster size: The number of individuals within each cluster was allowed to vary, with a coefficient of variation cv. This coefficient of variation is defined as:

$$cv = \frac{s_m}{m},$$

where s_m is the standard deviation of the cluster size. Fixing the average cluster size m and the coefficient of variation *cv*, cluster sizes were generated in a negative binomial distribution, as in Ref. [A1] with NegBin$\left(\frac{m^2}{s_m^2 - m}, \frac{m}{s_m^2}\right)$ with $s_m^2 = (cv \times m)^2$. This distribution was truncated to have a minimum cluster size of 2.

Outcome model: the individual-level continuous outcome Y_{ij} was generated under the following mixed-effects model:

$$Y_{ij} = \delta T_i + \gamma_i + \varepsilon_{ij},$$

for the individual *j (j = 1,…,m_i)*, in the *ith* cluster *(i = 1,…, k)*. δ is the difference in means between the two groups. T_i is an indicator variable for the intervention (T_i=0 in the control group and 1 in the intervention group). γ_i is the random effect for cluster i and ε_{ij} the residual error for the *jth* subject in cluster i where $\gamma_i \sim N\left(0, \sigma_b^2\right)$ and $\varepsilon_{ij} \sim N(0, \sigma_w^2)$ with $\gamma_i \perp \varepsilon_{ij}$. The intraclass correlation coefficient (ICC) for the outcome is defined as [A2]:

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}.$$

Without loss of generalizability, we constrained $Var(Y|T_i) = \sigma_b^2 + \sigma_w^2 = 1$. Thus, $\sigma_b^2 = \rho$ and $\sigma_w^2 = 1 - \rho$. This implies that $Y_{ij} \sim N(\delta T_i, 1)$.

Scenarios studied

We designed our simulation study in order to reflect situations observed in practice. Because our focus is on CRTs with a small number of clusters *k*, we chose the following values for the total number of randomised clusters *k: k = (4; 6; 8; 10; 20; 30; 40)*. The values for the average cluster size *m* were chosen according to those observed in a review of 100 CRTs looking at the use of small sample corrections in practice [A3]. The distribution of cluster sizes for CRTs in this review with less than 45 clusters are displayed in Table A1.

Table A1. Distribution of the average cluster size observed in 78 CRTs with a small number of clusters [A3]

Number of clusters	Number of trials	Median	[Q1-Q3]
4-8	11	93	[49-289]
9-15	20	53	[25-284]
16-25	20	34	[16-73]
26-35	11	55	[12-158]
36-45	15	16	[7-41]

For each value of *k*, we considered two values of *m*, based on the 1st and the 3rd quartile of the average cluster size distribution observed in the review for the corresponding number of clusters. Values were rounded for simplicity. Thus, we have: *m = (50, 300)* for *k = (4, 6, 8)*; *m = (25, 280)* for *k = 10*; *m = (15, 200)* for *k = 20*; *m = (10, 150)* for *k = 30*; *m = (7, 40)* for *k = 40*. As a "control" scenario, we also looked at *k = 200* with *m = 50* to assess the performance of the different approaches when the number of randomised clusters is no longer small.

Because cluster size is variable in most CRTs, we assessed scenarios with both a moderate and a high variability in the average cluster size, using values of 0.4 and 0.8 for the coefficient of variation of the cluster size. These two values are the smallest and the largest values observed in the examples presented by Eldridge *et al.* [A4]. Finally, we used 3 different ICC values for the outcome *ρ= (0.001, 0.01, 0.05)*. An ICC of 0.05 corresponds to the median ICC value observed in Campbell's study [A5], but we also considered smaller values since the ICC tends to be smaller in large clusters [A5].

In a first set of simulations, we set $\delta = 0$ to evaluate the type I error rate. Then, we went on to examine scenarios with non-zero δ. We used values of the treatment effect δ obtained from the sample size formula for a mean difference accounting for the design effect (DE) [A4]:

$$\delta = \sqrt{2DE \times \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2}{\frac{mk}{2}}},$$

with a power of 80% ($z_{\beta} = z_{0.2}$) and a type I error of 5% ($z_{\alpha/2} = z_{0.025}$) and

$$DE = 1 + \left[\left(\frac{cv^2\left(\frac{k}{2} - 1\right)}{\frac{k}{2} + 1}\right)m - 1\right]\rho,$$

where k is the total number of clusters, m the average cluster size, cv the coefficient of variation of cluster size, ρ the ICC of the outcome, and z_{γ} the γ^{th} percentile of the standard normal distribution. By doing this, we fixed the nominal power at 80% in order to assess if the compared analysis strategies reach the expected power. Across the scenarios evaluated (see below), the median of the δ values used was 0.33, with values within the range [0.08-0.82].

The combination of possible parameters values for k, m, cv and ρ lead to 90 different scenarios. For each of the 90 scenarios, 5000 simulated datasets were generated with δ set to 0 to assess the type I error rate and another 5000 datasets to assess the power ($\delta \neq 0$). In summary, for each scenario, the simulations followed these steps:

1. Create k clusters
2. Draw cluster size m_i for each cluster from a negative binomial distribution
3. Generate individual continuous outcomes according to the mixed model described above
4. Analyse the dataset using the 12 methods described in the main paper
5. Store treatment effect estimates, standard errors and two-sided p-values.

Assessment of results

Results were assessed in terms of:

- Type I error rate: defined as the proportion of simulations in which the p-value for the intervention effect was < 0.05 when the true treatment effect $\delta = 0$
- Power: defined as the proportion of simulations in which the p-value for the intervention effect was < 0.05 when the true treatment effect $\delta \neq 0$

- Bias of the treatment effect: $B(\hat{\delta}) = E(\hat{\delta}) - \delta$, calculated as the difference between the average treatment effect across the simulations and the true treatment effect δ
- Variability ratio: defined as the ratio of the mean model-based standard error to the empirical standard deviation of the treatment effect estimate
- Rate of convergence and estimation issues: defined as the percentage of models for which convergence problems arise.

Simulations were performed using R software version 3.1. For mixed models, the Satterthwaite correction was obtained from the *lmerTest* package [A6] and the *pbkrtest* [A7] package was used for Kenward-Roger estimator. The small sample correction for GEE was obtained using the package *saws* [A8] after estimating GEE parameters using the package *gee* [A9].

[A1] Zou, G., Donner, A.: Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. Biometrics 60(3), 807-811 (2004)

[A2] Donner, A., Klar, N.: Design and Analysis of Cluster Randomization Trials in Health Research. Arnold, London (2000)

[A3] Kahan, B.C., Forbes, G., Ali, Y., Jairath, V., Bremner, S., Harhay, M.O., Hooper, R., Wright, N., Eldridge, S.M., Leyrat, C.: Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. Trials 17(1), 438 (2016)

[A4] Eldridge, S.M., Ashby, D., Kerry, S.: Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. International Journal of Epidemiology. 35(5), 1292-1300 (2006)

[A5] Campbell, M.K., Fayers, P.M., Grimshaw, J.M.: Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. Clinical Trials (London, England) 2(2), 99-107 (2005)

[A6] Kuznetsova, A., Bruun Brockho, P., Haubo Bojesen Christensen, R.: lmerTest: Tests in Linear Mixed Effects Models. (2015). http://CRAN.R-project.org/package=lmerTest

[A7] Halekoh, U., Hojsgaard, S.: Parametric Bootstrap and Kenward Roger Based Methods for Mixed Model Comparison. (2016). https://cran.r-project.org/web/packages/pbkrtest/pbkrtest.pdf

[A8] Fay, M.P., Graubard, B.I.: Small-sample adjustments for wald-type tests using sandwich estimators. Biometrics 57, 1198-1206 (2001)

[A9] Carey VJ (2002). gee: Generalized Estimation Equation Solver. R package version 4.13-10; Ported from S-PLUS to R by Thomas Lumley (versions 3.13 and 4.4) and Brian Ripley (version 4.13).