UNIVERSITYOF BIRMINGHAM

Key considerations for designing, conducting and analysing a cluster randomized trial

Hemming, Karla; Taljaard, Monica

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

OXFORD

Education Corner

Key considerations for designing, conducting and analysing a cluster randomized trial

Karla Hemming [1],* and Monica Taljaard[2,3]

[1]Institute of Applied Health Research, University of Birmingham, Birmingham, UK, [2]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada and [3]School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, ON, Canada

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

Abstract

Not only do cluster randomized trials require a larger sample size than individually randomized trials, they also face many additional complexities. The potential for contamination is the most commonly used justification for using cluster randomization, but the risk of contamination should be carefully weighed against the more serious problem of questionable scientific validity in settings with post-randomization identification or recruitment of participants unblinded to the treatment allocation. In this paper we provide some simple guidelines to help researchers conduct cluster trials in a way that minimizes potential biases and maximizes statistical efficiency. The overarching theme of this guidance is that methods that apply to individually randomized trials rarely apply to cluster randomized trials. We recommend that cluster randomization be only used when necessary—balancing the benefits of cluster randomization with its increased risks of bias and increased sample size. Researchers should also randomize at the lowest possible level—balancing the risks of contamination with ensuring an adequate number of randomization units—as well as exploring other options for statistically efficient designs. Clustering should always be allowed for in the sample size calculation; and the use of restricted randomization (and adjustment in the analysis for covariates used in the randomization) should be considered. Where possible, participants should be recruited before randomizing clusters and, when recruiting (or identifying) participants post-randomization, recruiters should be masked to the allocation. In the analysis, the target of inference should align with the research question, and adjustment for clustering and small sample corrections should be used when the trial includes less than about 40 clusters.

Key words: Cluster randomized trials, statistical analysis plans, cluster randomization justification, risk of bias

1

Key Messages

- Cluster randomization should be used only when necessary: not only do cluster randomized trials require larger sample sizes than individually randomized trials, but they also have many other complexities.
- The potential for contamination is the most commonly used justification for using cluster randomization, but risks of contamination should be weighed against other risks before adopting cluster randomization.
- A key consideration with cluster randomization is the questionable scientific validity of inferences in settings with post-randomization identification or recruitment unblinded to the treatment allocation.
- When cluster randomization is necessary, adhering to key recommendations for this design is essential to minimize potential biases and maximize statistical efficiency.

Introduction

In individually randomized trials, often referred to as patient randomized trials, individuals are allocated independently to different interventions, referred to here as treatment or control conditions. Instead of individuals, cluster randomized trials randomize entire clusters.[1,2] Examples of clusters in a health care setting include wards, hospitals or primary care clinics and, in a non-health care setting, schools or communities.[3] Cluster randomized trials are used to evaluate diverse types of interventions. Whereas clusters are randomized, the interventions may be delivered at the level of the cluster (e.g. a change to the electronic health care record), professionals within clusters (e.g. education of health care providers) or directly to individuals (e.g. a drug). For example, the CAP trial randomized primary care practices to evaluate an intervention delivered to the individual (cancer screen[4]; whereas the CHIME-GP trial plans to randomize primary care practices to evaluate an educational intervention delivered to general practitioners.[5] Moreover, many cluster trials evaluate complex interventions with components delivered at multiple levels.

Over and above the unit of intervention delivery, cluster randomized trials can have different levels for the unit of analysis. If the unit of randomization is different from the unit of analysis (e.g. medical practices are randomized and outcomes collected on individuals), observations are correlated within clusters, which reduces the effective sample size.[6] Thus, cluster randomized trials require larger sample sizes compared with individually randomized trials.[7,8] Cluster randomized trials can also be at greater risk of bias.[9–11] Despite being at increased risk of bias and requiring larger sample sizes, cluster randomized trials are often selected for logistical reasons: simplifying the logistics of intervention delivery when everyone in the same cluster is treated in the same way.[12] Cluster trials can also be advantageous when access to active interventions might be costly, or when there is a high risk of contamination.[12] This design can also facilitate research embedded within

usual care, particularly when it is ethically appropriate not to take individual-level consent and where routinely collected data are used for outcome assessment. Thus, cluster trials are also aligned with the pragmatic trials' agenda—facilitating the evaluation of interventions in representative populations and under real-world conditions where there will be non-adherence, lack of compliance and co-interventions.[13] Indeed, the use of cluster randomization has been steadily increasing over the past decades.[14]

However, cluster randomized trials are much more complex to design, analyse and report compared with individually randomized trials. Thus, although cluster randomized trials are an essential design in the toolkit of a health researcher, there are several critical requirements for their successful adoption, implementation and interpretation. Numerous systematic reviews have shown that there are major methodological concerns with published cluster randomized trials.[3,15–18] In this manuscript, we provide 10 of the most important requirements for designing, conducting and analysing cluster randomized trials to help researchers overcome their major shortcomings (Box 1).

1. The golden rule: methods that apply to individually randomized trials rarely apply to cluster randomized trials

From planning to reporting, most methods that are used in individually randomized trials do not easily translate and apply to cluster randomized trials. For example, the types of interventions often evaluated in cluster trials, such as complex interventions and implementation strategies, can benefit from careful piloting and refinement before the definitive evaluation.[19] Furthermore, when planning pilot or feasibility studies, there are many nuances to consider in cluster randomized trials—such as how many clusters are required to estimate an intra-cluster correlation coefficient (ICC) with reasonable stability[20] and whether participants be recruited in a way that will not undermine the face validity of the trial.[9] Likewise, these types of evaluations

---

Box 1: The ten commandments for conducting a cluster randomized trial

1. The Golden Rule: methods that apply to individually randomized trials rarely apply to cluster randomized trials
2. Only use cluster randomization when necessary—balancing the benefits of cluster randomization with its increased risks of bias and increased sample size
3. Always randomize at the lowest possible level—balancing the risks of contamination with ensuring an adequate number of randomization units
4. Use the most statistically efficient designs
5. Always allow for clustering in the sample size calculation
6. Consider using restricted randomization and always adjust the analysis for covariates used in the restricted randomization
7. Always recruit participants before randomizing clusters whenever possible and when recruiting (or identifying) participants post-randomization, keep recruiters masked to the cluster allocation
8. Specify the target of inference to align with research questions
9. Always allow for clustering in the analysis
10. Always consider the use of a small sample correction in analysis

---

often benefit from a process evaluation or implementation evaluation alongside trial results, and so often require hybrid trial designs looking at effects on both clinical and implementation outcomes.[21] When it comes to reporting, this should be according to the CONSORT extension to cluster randomized trials[22] or its extension for stepped-wedge trials.[23]

2. Only use cluster randomization when necessary—balancing the benefits of cluster randomization with its increased risks of bias and increased sample size

Cluster randomization leads to an inflation in the required sample size and increases risk of bias compared with individually randomized trials. Exposing participants unnecessarily to the risks of research, either because the same question could have been answered with fewer participants or because the chosen design renders the results uninformative because of bias, is unethical.[22] Because of this, adopting a cluster randomized design should always be carefully justified, and an individually randomized design used in preference wherever possible[12]—a piece of advice that has been around for some time.[24]

Cluster randomization is necessary when the intervention is delivered at the level of the cluster. Apart from this obvious justification, concerns over contamination (i.e. control participants inadvertently receiving the

intervention) is one of the most common justifications for adopting a cluster randomized design.[3,12] Interest in the total effect of an intervention (both its indirect and direct components) is another legitimate reason to use cluster randomization, arising particularly in vaccine trials: here, the contamination can be thought of as desirable.[12] Likewise, opting to use cluster randomization in preference to individual randomization for logistical reasons or to avoid 'disappointment' effects (e.g. participants allocated to the control in an evaluation of a conditional cash transfer[25]) are other common reasons for adopting cluster randomization.[26] Cluster randomization can also help with efforts to increase generalizability of findings—widening inclusion criteria and populations under evaluation.[27] However, cluster randomization should never be adopted under the mistaken perception that it can help avoid seeking individual informed consent.[28]

Where participants have to be identified or recruited post-randomization, identification and recruitment bias operate in an unpredictable direction and can render a cluster trial much like an observational study (below).[11] On the other hand, the impact of contamination is predictable: it will attenuate the treatment effect when those in the control inadvertently receive the intervention.[29] If contamination can be accurately measured and an individually randomized approach adopted, it is possible to estimate the complier average causal effect (that is, estimating the impact among those complying with the treatment allocation). Furthermore, an adjustment can be made to the sample size calculation, and often the inflation to account for contamination is substantially less than the design effect due to clustering. Thus, the use of cluster randomization, particularly to protect against contamination, needs to be balanced against other risks: the use of cluster randomization can render the design at risk of unpredictable sources of bias (when it is used with unblinded post randomization recruitment), thus undermining its scientific validity.[29]

3. Always randomize at the lowest possible level—balancing the risks of contamination with ensuring an adequate number of randomization units

In some cluster randomized trials, there is a choice with respect to the unit of randomization, which can be at a higher or lower level.[8] The choice involves a trade-off between logistics, contamination and statistical efficiency. An example of randomizing at a lower level is choosing to randomize wards rather than hospitals. Logistics are often enhanced when randomizing at a higher level. For example when randomizing hospitals, only one type of intervention is delivered throughout the hospital as opposed to in

different wards. Contamination may also be reduced when randomizing at a higher level. For example, choosing to randomize hospitals rather than wards can prevent contamination due to providers in the same hospital being exposed to both treatment and control conditions. There are additional considerations when choosing clusters based on geographical areas: where clusters are too close in location, contamination might arise.

Nevertheless, statistical efficiency is increased when randomizing at a lower level: randomizing wards rather than hospitals could mean that more units are available for randomization. Statistical efficiency refers to the statistical power achieved from the available number of observations under the given design compared with the power that would have been available with the same number of observations under an alternative design. Power in a cluster trial is increased by increasing the number of clusters or increasing the cluster size. However, increasing the number of clusters leads to the largest increase in power (compared with increasing the cluster size), all other things being equal.[30] When there is a choice to increase either the number of clusters or the cluster sizes, it is always better to increase the number of clusters.[30]

In theory, a parallel-arm cluster randomized trial can be implemented with a minimum of eight clusters (four per arm) which is the minimum required to obtain a *P*-value less than 0.05 under a randomization-based test.[26] However, trials with a greater number of randomization units are much more likely to deliver on their wider objectives (will have increased face validity, will be amenable to an analysis based on standard approaches and are likely more generalizable) irrespective of their total sample size.[31]

4. Use the most statistically efficient design

Whereas increasing the number of clusters is the most statistically efficient way of increasing study power in theory, the number of clusters that can be included in practice is typically limited due to logistical or resource constraints.[32] However, for a fixed number of clusters there are alternative ways of enhancing statistical efficiency, and this can be as simple as including a baseline measure of the response.[33] In closed cohort designs, a baseline measure of the outcome could potentially be taken on all participants (ideally before randomization) so that the same participant is measured before and after randomization.[34] In repeated cross-sectional designs, measurements may be taken on different groups of participants before and after randomization. The baseline period may be prospective or retrospective, or baseline measures may be available in summary form at the cluster level. Using such baseline measures in parallel-

arm designs can have substantial power benefits[35]; in other settings, an unequal allocation of clusters to treatment arms might improve statistical efficiency.[32]

For a fixed number of clusters, the best way to improve statistical efficiency is to adopt a bi-directional, crossover design (i.e. a trial in which clusters switch in both directions, i.e. from an experimental to comparator and visa versa[36,37]). In practice however, interventions evaluated in a typical cluster randomized trial are often not possible to withdraw. Where only unidirectional crossover is possible, adding control periods—either by using a parallel cluster randomized trial with baseline measures or extending this to a stepped-wedge cluster randomized trial—can increase efficiency, with gains depending on the intra-cluster correlations (the higher the correlations and larger the cluster sizes, the greater the rewards).[35,38,39] As with cluster randomization, the risks and rewards of these alternative designs have to be balanced.[40,41] For example, cluster crossover designs should not be used where there are risks of carry-over effects; and stepped-wedge designs might make assumptions about underlying secular trends and about 'light switch' intervention effects (i.e. that the intervention has an immediate and sustained impact).[42,43]

Of note, large cluster sizes can be statistically inefficient, especially when the intra-cluster correlation coefficient is high.[30] This can mean that many observations within a cluster might not make a material contribution to the study power, and this can have implications for both the duration of the study and ethics (it might increase the duration or expose individuals to research risks for no material return).[44] Thus, trialists should consider the implications of decreasing the sizes of clusters on study power as a way of determining if all cluster members are making a material contribution.

5. Always allow for clustering in the sample size calculation

The design and analysis of individually randomized trials assume independence of observations. Observations within clusters tend to be correlated, and this correlation is measured by the intra-cluster correlation coefficient (ICC). When treatment assignment depends on the cluster, as it does in a cluster randomized trial, clustering is said to be 'informative'–in other words, it needs proper consideration in the sample size and analysis.[45] Even when the anticipated ICC is low, the impact on the required sample size can still be important, particularly when the cluster sizes are large.[15]

Sample size estimation should ideally be based on the proposed analysis model (below). In most settings, this can be achieved by inflating the number needed under

individual randomization by a 'design effect'.[46,47] Design effects are available for parallel cluster designs, cluster crossover designs, cluster designs with baseline periods and stepped-wedge designs (known as multiple period designs).[42,48,49] Care is needed when using these design effect approaches for small samples or rare or common binary outcomes—and in these settings determining power by simulation might be warranted[50] or extra clusters be added to each arm to accommodate the use of the t-distribution at the analysis stage (see Commandment 10[26]). Care is also needed in the setting where variation in the cluster sizes is anticipated—again increasing the number of clusters by 25% can protect against this.[51]

These approaches require the specification of measures of correlation, such as the ICC in a parallel design (more extensive correlations are required for multiple period designs), and sensitivity to assumptions about correlations should be investigated.[52] When outcome data are available for a similar setting (perhaps a similar trial or routinely collected data source), correlations can be estimated to inform these calculations; more commonly, researchers have to be guided by likely values considering patterns reported in the literature.[53–55] For binary outcomes, these correlations are needed on the proportions scale and not the logistic scale,[56] and are also known to be dependent on the prevalence of the outcome.[57,58]

6. Consider using restricted randomization and always adjust the analysis for covariates used in the restricted randomization

Cluster trials often randomize a small number of units—substantially fewer than in typical individually randomized trials, and most use some form of restricted randomization.[3,59] Restricted randomization methods can enhance the credibility of the trial results by protecting against imbalances in cluster and participant characteristics (sometimes referred to as enhancing face validity) and can also improve statistical power.[60] These benefits are likely to be greater with a smaller number of randomization units (where the risk of chance imbalance will be larger).[61] Restricted randomization methods use either cluster-level characteristics or cluster-level summaries of individual-level characteristics (e.g. cluster-level mean of primary outcome from a baseline period) to protect against poorly balanced allocations. There is limited guidance on the choice of factors for inclusion in a restricted randomization procedure, but as with covariate adjustment, factors should be chosen on the basis of their prognostic strength, availability and reliability. Examples of common restriction factors in cluster trials include cluster location, cluster size and a baseline measure of the outcome.[16] Unlike

individually randomized trials, where the randomization is usually implemented sequentially, in cluster trials, randomization of clusters is frequently implemented once-off or in batches.

There is a number of different approaches for restricted randomization in cluster trials, including stratified block randomization, minimization, covariate constrained randomization and pair matching.[59] The appropriate method in any given trial often depends on logistical constraints in the setting. For example, minimization allows for sequential randomization, whereas covariate constrained randomization does not. Minimization might be preferred over stratification when there are many covariates to balance.[62,63] Unlike stratification, both minimization and covariate-constrained randomization can balance on continuous covariates and so do not require categorization of prognostic factors.[64] Pair matching has an intuitive appeal, but others have confirmed that its usefulness might be less than expected.[65] Blocking can help prevent large imbalances in numbers of clusters allocated to each arm.[66]

It can be tempting to include many covariates in restricted randomization, but this can be problematic, for example leading to incomplete blocks in stratification[62,67] or overly constrained designs where some pairs of clusters are always allocated to the same arm.[68,69] Furthermore, when the allocation becomes more deterministic it becomes reliant on covariates being truly prognostic and can also lead to subversion of the allocation process where it is easy to predict upcoming allocations. Finally, when restricted randomization has been used, the analysis should adjust for the covariates used in the randomization to ensure nominal type I errors.[61]

7. Always recruit participants before randomizing clusters whenever possible, and when recruiting (or identifying) participants post-randomization, keep recruiters masked to the cluster allocation

To ensure allocation concealment, there is a recommended process which, in individually randomized trials, is almost always followed: individuals are recruited, and once their participation is ratified, they are randomized to the treatment or control condition.[70] In cluster trials, this may be possible in closed cohort designs: for example, in a study where eligible participants are children within a school classroom. Indeed, the majority of cluster trials within a school setting use the closed cohort design.[71] Alternatively, clusters are randomized and participants recruited continuously throughout the trial, with each participant providing a single measurement (known as a 'continuous recruitment' design). Continuous recruitment designs are used when eligible participants are incident populations, for

example, if participants are people with a new diagnosis of hypertension, it would not be possible to recruit all participants before clusters were randomized. This reversal of the ordering can prevent cluster allocation being concealed from recruiters and participants at the time of recruitment, and this can be an important source of bias in cluster trials.[9,72] There are numerous ways this bias can manifest— for example, recruiters and participants in the control arm might differentially recruit and agree to participation in the study compared with those in the intervention arm.[73]

Although it is preferable to recruit participants before randomization of clusters, in practice, this is not always possible. Sometimes cluster trials can be conducted without participant recruitment. For example, cluster randomized trials might involve whole clusters, i.e. a complete enumeration of the entire cluster of eligible participants. This may be appropriate when individuals are not considered research participants and when there are routinely available outcome data (e.g. mortality) and an ethics committee has granted a waiver of participant informed consent.[28] In these settings, the risk of recruitment bias is remove but there remains the possibility that participants could be differentially identified across the study arms. Identification biases are a possible cause of the differential characteristics of participants across study arms in an unblinded evaluation of rapid screening for group B streptococcus in pregnancy using cluster randomization (no individual participant recruitment).[74]

When post-randomization recruitment or identification of participants is unavoidable, there will be a risk of recruitment biases, unless treatment conditions are blinded. However, due to the nature of interventions evaluated in cluster trials, blinding is often not possible. In these settings, mitigation strategies can help prevent recruitment biases.[9,72] First, recruitment by someone independent of the trial who is blind to the cluster status will minimize risks of recruitment bias. Second, recruitment strategies should be consistent across the study arms—for example, consent forms should be similar under treatment and control conditions. Finally, where possible, keeping information on the precise details of the intervention only known to essential people can help.

Reporting whether participants were recruited post-randomization and whether recruiters and participants were aware of their allocation can help others identify these risks when interpreting trial results. Recommended reporting practices include clear reporting of blinding status along the timeline of recruitment and randomization.[75] In addition, making the consent forms available improves transparency on what information was communicated to participants about the trial (e.g. were participants aware that their arm was the active intervention). In some settings

there may be grounds for statistical testing of differences across intervention arms.[76]

8. Specify the target of inference to align with research questions

In individually randomized trials, the target of inference is almost always the individual, i.e. the trial objective is to determine the impact of the intervention on the typical individual. In cluster randomized trials, the trial objective might be to determine the impact of the intervention on either the typical individual or the typical cluster.[26,77] Careful specification of the estimand (target of inference) helps identify the appropriate design and analysis.[78] For example, cohort sampling aligns with an interest in the impact on the typical individual; and repeated cross-sectional sampling aligns with an interest in the impact on the typical cluster.[34,77] Furthermore, depending on whether the target of inference is the typical cluster or typical individual, in conjunction with the possibility that the impact of the intervention might vary with cluster size (known as 'informative cluster sizes'), an analysis approach should be chosen that appropriately targets the effect for the typical individual or the typical cluster (below).[79,80]

Other considerations when specifying the estimand include, but are not limited to, the type of summary measure and covariate adjustment. Whereas there is some debate over the best summary measure to report, it is likely that both an absolute and a relative measure will be appropriate in most trials.[22] Direct covariate adjustment changes the target of inference from unconditional (the expected treatment effect for a typical individual, also known as the marginal effect) to a conditional estimate (the expected treatment effect for a particular subgroup of individuals defined by the covariates in the model).[81] Differences will exist between these two estimands when the summary measure is non-collapsible (odds ratios and hazard ratios). Differences will be more noticeable when the outcome prevalence varies across sub-groups. Furthermore, if effects vary across sub-groups (effect modification), a conditional estimate might not be very useful.[81] Marginal standardization, as an alternative to direct covariate adjustment, can be used to estimate the marginal covariate-adjusted summary measure, as too can inverse probability weighting.[82,83]

Any covariates to be included in a covariate-adjusted estimate of the treatment effect should be pre-specified and should include all covariates used in any restricted randomization.[16,61] Unlike in individually randomized trials, adjustment for covariates might not always increase statistical precision.[84] Adjusting for both cluster-level and individual-level versions (of the same covariate) can help capture residual confounding.[85] Covariate adjustment is

likely to be particularly important when there is post-randomization unblinded recruitment or identification of participants.[82] Whereas care is needed in small samples, small sample corrections can maintain nominal type I errors alongside covariate adjustment.[61,82] When adjusting for covariates in the context of low or high prevalence binary outcomes, model convergence can be problematic, and in these settings, propensity score approaches might help.[82,83] Where covariate data are incomplete, any multiple imputation procedures should appropriately allow for the clustered nature of the design.[86]

9. Always allow for clustering in the analysis

Clustering should always be allowed for in the statistical analysis. Significance testing should not be used to compare models that do and do not allow for clustering. There are a number of ways to allow for the non-independence of observations in the analysis.[87,88] The simplest is to carry out what is known as a cluster-level analysis.[89] Essentially this consists of aggregating cluster-level outcomes by a summary statistic (such as the mean or proportion) and then using conventional analysis methods on these summary statistics (e.g. a t test[90]). For binary outcomes, this approach can be used in conjunction with transformations (e.g. log to report relative risks or logit to report odds ratios). Cluster-level approaches (unweighted by cluster size) allow inferences targeted at estimating the effect of the treatment for the typical cluster.[79] When the objective is to determine the impact on the typical individual, this approach might not work well when there is substantial variation in cluster sizes.[91] Furthermore, the cluster-level approach is more complex when it is desirable to adjust for individual-level covariates,[2] which may explain why cluster-level approaches are infrequently used.[3] However, cluster-level approaches are robust with a small number of clusters.[92]

If an individual-level analysis is preferred, two main approaches to accommodating the non-independence are generalized linear mixed models (GLMM) and generalized estimating equations (GEE).[87] For odds ratios (and other non-collapsible link functions), these two approaches yield different interpretations of the treatment effect; in particular, GLMM yields a cluster-specific estimate, i.e. the effect of the treatment conditional on cluster membership, whereas GEE yields an unconditional (marginal) estimate. These approaches facilitate adjustment for individual-level covariates, but do not work well when there are fewer than about 50 clusters (40 for GLMMs), unless combined with a small sample correction (below).[92] For continuous outcomes, these models typically have reasonable convergence properties, but when using binomial distributions and log or identity links for binary outcomes, convergence problems can arise

(especially with rare or common outcomes). In these settings, modified robust Poisson models can be used to facilitate the estimation of relative risks.[93] Although marginal standardization is a potential alternative option to estimate relative risks and risk differences, its performance has yet to be evaluated in cluster trials. Care is needed when using individual-level methods of analysis in settings where informative cluster size is considered plausible (here, GEE with independent estimating equations might be necessary).[79] For multiple period designs, non-exchangeable correlations should be allowed for to avoid overestimation of statistical precision.[94] For cluster trials with baseline periods, this can be accommodated using either a constrained baseline analysis with random cluster by period effects, or analysis of covariance (ANCOVA)-type approaches (adjusting for baseline values of the outcome—possibly at the level of the cluster).[95]

10. Always consider the use of a small sample correction in analysis

In what was probably the first methodological paper on cluster randomized trials, Cornfield in 1978 referred to two penalties to cluster randomization[6]: variance inflation due to clustering, which is a function of the ICC (also known as the 'design effect') and the degree of freedom penalty, often referred to as a small sample correction. The first of these is widely recognized as an implication of cluster randomization. The second turns out to be consequential but much less widely appreciated.[96] The degree of freedom penalty is essentially a penalty to account for the use of a large-sample approximation (i.e. z test) when a more exact method is required, such as a permutation test or t test.[26,77] It turns out that t tests are required when there are less than about 40 independent observations.[6] When a t test is used, the degrees of freedom must be specified. Under individual randomization, the degrees of freedom for parallel-arm assignment are $2N - 2$, where N is the number of participants per arm.

Appropriate degrees of freedom in parallel cluster trials are not so clear-cut. Options include $2K - 2$ (known as the 'between-within' correction), where K is the number of clusters per arm.[92] This tends to work well for both continuous and binary outcomes when there is limited cluster size variation.[92,97] Other available options are more complicated but might have better statistical properties (i.e. more likely to maintain type I error at 5%). For GLMMs, these degrees of freedom options include Kenward–Roger and Satterthwaite approximations.[97,98] For GEEs, they include but are not limited to Kauermann–Carroll and Mancl–DeRouen.[99] The performance of these is dependent on setting (e.g. number of clusters, cluster sizes, ICC, cluster size variation, outcome prevalence), and none appear to

work well across the board. Typically, these 'corrections' estimate degrees of freedom to be used under a t test, and some of them also consist of making a 'correction' to the standard error of the treatment effect. Indeed, in addition to these degree of freedom corrections, robust standard errors ('sandwich' variance) or restricted maximum likelihood (REML) procedures should be used when fitting GEE or GLMM, respectively.[100] Since cluster randomized trials typically include less than 40 clusters, almost all cluster trials should make allowance for this correction.[3]

Conclusion

Cluster randomized trials face many complexities. Not only do they require larger sample sizes than individually randomized trials, especially for large cluster sizes and large intra-class correlations; other complexities include sometimes being more complicated to implement, being more vulnerable to poor sample size estimation and more vulnerable to invalidating model-based assumptions. Perhaps of most importance is their questionable scientific validity in settings with post-randomization identification or recruitment unblinded to the treatment allocation. The potential for contamination is the most commonly used justification for using cluster randomization, whereas the risks of identification and recruitment biases are less widely appreciated. This needs a shift in balance—and researchers should weigh up the different risks before adopting cluster randomization. Cluster randomization is essential when evaluating cluster-level interventions, but determining the level of the intervention is not always straightforward. When cluster randomization is necessary, these simple commandments should help researchers conduct these evaluations in a way that minimizes potential biases and maximizes statistical efficiency.

Ethics approval

Not applicable.

Data availability

No new data were generated or analysed in support of this research.

Author contributions

K.H. led the development of the idea and led the writing of the paper. M.T. provided critical input and oversight at all stages of development.

Conflict of interest

None declared.

References

1. Murray DM. *Design and Analysis of Group Randomized Trials.* New York, NY: Oxford University Press, 1998.
2. Hayes RJ, Moulton LH. *Cluster Randomized Trials,* 2nd edn. London: CRC Press, 2017. https://doi.org/10.4324/9781315370286.
3. Turner EL, Platt AC, Gallis JA *et al.*; CRT Binary Outcome Reporting Group. Completeness of reporting and risks of overstating impact in cluster randomized trials: a systematic review. *Lancet Glob Health* 2021;**9**:e1163–68.
4. Martin RM, Donovan JL, Turner EL *et al.*; CAP Trial Group. Effect of a low-intensity PSA-based screening intervention on prostate cancer mortality: the CAP randomized clinical trial. *JAMA* 2018;**319**:883–95.
5. Bonney A, Metusela C, Mullan J *et al.* Clinical and healthcare improvement through My Health Record usage and education in general practice (CHIME-GP): a study protocol for a cluster-randomized controlled trial. *Trials* 2021;**22**:569.
6. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol* 1978;**108**:100–02.
7. Turner EL, Li F, Gallis JA, Prague M, Murray DM. Review of recent methodological developments in group-randomized trials. Part 1: design. *Am J Public Health* 2017;**107**:907–15.
8. Murray DM, Taljaard M, Turner EL, George SM. Essential ingredients and innovations in the design and analysis of group-randomized trials. *Annu Rev Public Health* 2020;**41**:1–19.
9. Eldridge S, Kerry S, Torgerson DJ. Bias in identifying and recruiting participants in cluster randomized trials: what can be done? *BMJ* 2009;**339**:b4006.
10. Bolzern J, Mnyama N, Bosanquet K, Torgerson DJ. A review of cluster randomized trials found statistical evidence of selection bias. *J Clin Epidemiol* 2018;**99**:106–12.
11. Easter C, Thompson JA, Eldridge S, Taljaard M, Hemming K. Cluster randomized trials of individual-level interventions were at high risk of bias. *J Clin Epidemiol* 2021;**138**:49–59.
12. Taljaard M, Goldstein CE, Giraudeau B *et al.* Cluster over individual randomization: are study design choices appropriately justified? Review of a random sample of trials. *Clin Trials* 2020;**17**:253–63.
13. Luce BR, Kramer JM, Goodman SN *et al.* Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. *Ann Intern Med* 2009;**151**:206–09.
14. Murray DM. Influential methods reports for group-randomized trials and related designs. *Clin Trials* 2022;**19**:353–62.

15. Rutterford C, Taljaard M, Dixon S, Copas A, Eldridge S. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. *J Clin Epidemiol* 2015;**68**:716–23.

16. Wright N, Ivers N, Eldridge S, Taljaard M, Bremner S. A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice. *J Clin Epidemiol* 2015;**68**:603–09.

17. Murray DM, Pals SL, George SM *et al.* Design and analysis of group-randomized trials in cancer: A review of current practices. *Prev Med* 2018;**111**:241–47.

18. Offorha BC, Walters SJ, Jacques RM. Statistical analysis of publicly funded cluster randomized controlled trials: a review of the National Institute for Health Research Journals Library. *Trials* 2022;**23**:115.

19. Pearson N, Naylor PJ, Ashe MC, Fernandez M, Yoong SL, Wolfenden L. Guidance for conducting feasibility and pilot studies for implementation trials. *Pilot Feasibility Stud* 2020;**6**:167.

20. Eldridge SM, Costelloe CE, Kahan BC, Lancaster GA, Kerry SM. How big should the pilot study for my cluster randomized trial be? *Stat Methods Med Res* 2016;**25**:1039–56.

21. Wolfenden L, Foy R, Presseau J *et al.* Designing and undertaking randomized implementation trials: guide for researchers. *BMJ* 2021;**372**:m3721.

22. Campbell MK, Piaggio G, Elbourne DR, Altman DG; CONSORT Group. Consort 2010 statement: extension to cluster randomized trials. *BMJ* 2012;**345**:e5661.

23. Hemming K, Taljaard M, McKenzie JE *et al.* Reporting of stepped wedge cluster randomized trials: extension of the CONSORT 2010 statement with explanation and elaboration. *BMJ* 2018;**363**:k1614.

24. Mainland D. *Elementary Medical Statistics. The Principles of Quantitative Medicine.* 6th edn. Philadelphia, PA: W B Saunders 1952, p.114.

25. Robertson L, Mushati P, Eaton JW *et al.* Effects of unconditional and conditional cash transfers on child health and development in Zimbabwe: a cluster-randomized trial. *Lancet* 2013;**381**:1283–92.

26. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research.* Chichester, UK: Wiley, 2000.

27. Eldridge S, Ashby D, Bennett C, Wakelin M, Feder G. Internal and external validity of cluster randomized trials: systematic review of recent trials. *BMJ* 2008;**336**:876–80.

28. Nix HP, Weijer C, Brehaut JC, Forster D, Goldstein CE, Taljaard M. Informed consent in cluster randomized trials: a guide for the perplexed. *BMJ Open* 2021;**11**:e054213.

29. Hemming K, Taljaard M, Moerbeek M, Forbes A. Contamination: how much can an individually randomized trial tolerate? *Stat Med* 2021;**40**:3329–51.

30. Hemming K, Eldridge S, Forbes G, Weijer C, Taljaard M. How to design efficient cluster randomized trials. *BMJ* 2017;**358**:j3064.

31. Taljaard M, Teerenstra S, Ivers NM, Fergusson DA. Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. *Clin Trials* 2016;**13**:459–63.

32. Copas AJ, Hooper R. Optimal design of cluster randomized trials allowing unequal allocation of clusters and unequal cluster size between arms. *Stat Med* 2021;**40**:5474–86.

33. Teerenstra S, Eldridge S, Graff M, de Hoop E, Borm GF. A simple sample size formula for analysis of covariance in cluster randomized trials. *Stat Med* 2012;**31**:2169–78.

34. Atienza AA, King AC. Community-based health intervention trials: an overview of methodological issues. *Epidemiol Rev* 2002;**24**:72–79.

35. Hooper R, Copas AJ. Optimal design of cluster randomized trials with continuous recruitment and prospective baseline period. *Clin Trials* 2021;**18**:147–57.

36. Bellomo R, Forbes A, Akram M, Bailey M, Pilcher DV, Cooper DJ. Why we must cluster and cross over. *Crit Care Resusc* 2013;**15**:155–57.

37. Grantham KL, Kasza J, Heritier S, Hemming K, Litton E, Forbes AB. How many times should a cluster randomized crossover trial cross over? *Stat Med* 2019;**38**:5021–33.

38. Lawrie J, Carlin JB, Forbes AB. Optimal stepped wedge designs. *Stat Probab Lett* 2015;**99**:210–14.

39. Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Stat Med* 2016;**35**:2149–66.

40. Hemming K, Taljaard M. Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice? *Int J Epidemiol* 2020;**49**:1043–52.

41. Hooper R, Eldridge SM. Cutting edge or blunt instrument: how to decide if a stepped wedge design is right for you. *BMJ Qual Saf* 2021;**30**:245–50.

42. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007;**28**:182–91.

43. Kenny A, Voldal EC, Xia F, Heagerty PJ, Hughes JP. Analysis of stepped wedge cluster randomized trials in the presence of a time-varying treatment effect. *Stat Med* 2022;**41**:4311–39.

44. Bacchetti P, Wolf LE, Segal MR, McCulloch CE. Ethics and sample size. *Am J Epidemiol* 2005;**161**:105–10.

45. Kahan BC, Morris TP. Assessing potential sources of clustering in individually randomized trials. *BMC Med Res Methodol* 2013;**13**:58.

46. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol* 2006;**35**:1292–300.

47. Hemming K, Kasza J, Hooper R, Forbes A, Taljaard M. A tutorial on sample size calculation for multiple-period cluster randomized parallel, crossover and stepped-wedge trials using the Shiny CRT Calculator. *Int J Epidemiol* 2020;**49**:979–95.

48. Hooper R, Bourke L. Cluster randomized trials with repeated cross sections: alternatives to parallel group designs. *BMJ* 2015;**350**:h2925.

49. Kasza J, Hooper R, Copas A, Forbes AB. Sample size and power calculations for open cohort longitudinal cluster randomized trials. *Stat Med* 2020;**39**:1871–83.

50. Reich NG, Myers JA, Obeng D, Milstone AM, Perl TM. Empirical power and sample size calculations for cluster-randomized and cluster-randomized crossover studies. *PLoS One* 2012;**7**:e35564.

51. van Breukelen GJ, Candel MJ. Comments on 'Efficiency loss because of varying cluster size in cluster randomized trials is smaller than literature suggests'. *Stat Med* 2012;**31**:397–400.

52. Kasza J, Forbes AB. Inference for the treatment effect in multiple-period cluster randomized trials when random effect

correlation structure is misspecified. *Stat Methods Med Res* 2019;**28**:3112–22.

53. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol* 2004;**57**:785–94.

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

55. Korevaar E, Kasza J, Taljaard M *et al.* Intra-cluster correlations from the CLustered OUtcome Dataset bank to inform the design of longitudinal cluster trials. *Clin Trials* 2021;**18**:529–40.

56. Yelland LN, Salter AB, Ryan P, Laurence CO. Adjusted intra-class correlation coefficients for binary data: methods and estimates from a cluster-randomized trial in primary care. *Clin Trials* 2011;**8**:48–58.

57. Gulliford MC, Adams G, Ukoumunne OC, Latinovic R, Chinn S, Campbell MJ. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *J Clin Epidemiol* 2005;**58**:246–51.

58. Mbekwe Yepnang AM, Caille A, Eldridge SM, Giraudeau B. Association of intracluster correlation measures with outcome prevalence for binary outcomes in cluster randomized trials. *Stat Methods Med Res* 2021;**30**:1988–2003.

59. Ivers NM, Halperin IJ, Barnsley J *et al.* Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials* 2012;**13**:120.

60. Moerbeek M, van Schie S. How large are the consequences of covariate imbalance in cluster randomized trials: a simulation study with a continuous outcome and a binary covariate at the cluster level. *BMC Med Res Methodol* 2016;**16**:79.

61. Li F, Turner EL, Heagerty PJ, Murray DM, Vollmer WM, DeLong ER. An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes. *Stat Med* 2017;**36**:3791–806.

62. Therneau TM. How many stratification factors are "too many" to use in a randomization plan? *Control Clin Trials* 1993;**14**: 98–108.

63. Martin J, Middleton L, Hemming K. Minimisation for the design of parallel cluster-randomized trials: An evaluation of balance in cluster-level covariates and numbers of clusters allocated to each arm. *Clin Trials* 2023;**20**:111–20.

64. Xiao L, Yank V, Ma J. Algorithm for balancing both continuous and categorical covariates in randomized controlled trials. *Comput Methods Programs Biomed* 2012;**108**:1185–90.

65. Chondros P, Ukoumunne OC, Gunn JM, Carlin JB. When should matching be used in the design of cluster randomized trials? *Stat Med* 2021;**40**:5765–78.

66. Suresh K. An overview of randomization techniques: an unbiased assessment of outcome in clinical research. *J Hum Reprod Sci* 2011;**4**:8–11.

67. de Hoop E, Teerenstra S, van Gaal BG, Moerbeek M, Borm GF. The "best balance" allocation led to optimal balance in cluster-controlled trials. *J Clin Epidemiol* 2012;**65**:132–37.

68. Bailey RA, Rowley CA. Valid randomization. *Proc R Soc Lond A Math Phys Sci* 1987;**410**:105–24.

69. Moulton LH. Covariate-based constrained randomization of group-randomized trials. *Clin Trials* 2004;**1**:297–305.

70. Chalmers I. Why the 1948 MRC trial of streptomycin used treatment allocation based on random numbers. *J R Soc Med* 2011; **104**:383–86.

71. Parker K, Nunns M, Xiao Z, Ford T, Ukoumunne OC. Characteristics and practices of school-based cluster randomized controlled trials for improving health outcomes in pupils in the United Kingdom: a methodological systematic review. *BMC Med Res Methodol* 2021;**21**:152.

72. Giraudeau B, Ravaud P. Preventing bias in cluster randomized trials. *PLoS Med* 2009;**6**:e1000065.

73. Heim N, van Stel HF, Ettema RG, van der Mast RC, Inouye SK, Schuurmans MJ. HELP! Problems in executing a pragmatic, randomized, stepped wedge trial on the Hospital Elder Life Program to prevent delirium in older patients. *Trials* 2017;**18**: 220.

74. Daniels JP, Dixon E, Gill A *et al.*; GBS2 Collaborative Group. Rapid intrapartum test for maternal group B streptococcal colonisation and its effect on antibiotic use in labouring women with risk factors for early-onset neonatal infection (GBS2): cluster randomized trial with nested test accuracy study. *BMC Med* 2022;**20**:9.

75. Caille A, Kerry S, Tavernier E, Leyrat C, Eldridge S, Giraudeau B. Timeline cluster: a graphical tool to identify risk of bias in cluster randomized trials. *BMJ* 2016;**354**:i4291.

76. Bolzern JE, Mitchell A, Torgerson DJ. Baseline testing in cluster randomized controlled trials: should this be done? *BMC Med Res Methodol* 2019;**19**:106.

77. Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomization-based inference for community intervention trials. *Stat Med* 1996;**15**:1069–92.

78. Jin M, Liu G. Estimand framework: delineating what to be estimated with clinical questions of interest in clinical trials. *Contemp Clin Trials* 2020;**96**:106093.

79. Kahan BC, Li F, Copas AJ, Harhay MO. Estimands in cluster-randomized trials: choosing analyses that answer the right question. *Int J Epidemiol* 2023;**52**:107–18.

80. Hemming K, Taljaard M. Estimands in cluster trials: thinking carefully about the target of inference and the consequences for analysis choice. *Int J Epidemiol* 2022. doi: 10.1093/ije/dyac174.

81. Morris TP, Walker AS, Williamson EJ, White IR. Planning a method for covariate adjustment in individually randomized trials: a practical guide. *Trials* 2022;**23**:328.

82. Leyrat C, Caille A, Donner A, Giraudeau B. Propensity score methods for estimating relative risks in cluster randomized trials with low-incidence binary outcomes and selection bias. *Stat Med* 2014;**33**:3556–75.

83. Zhu AY, Mitra N, Hemming KO, Harhay H, Li F. Leveraging baseline covariates to analyze small cluster-randomized trials with a rare binary outcome. *Biom J* 2023:e2200135.

84. Wang B, Harhay MO, Small, DS, Morris TP, Li F. On the mixed-model analysis of covariance in cluster-randomized trials. Preprint arXiv:2112.00832

85. Begg MD, Parides MK. Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Stat Med* 2003;**22**:2591–602.

86. Díaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomized

trials? A systematic review and guidelines. *Clin Trials* 2014;**11**: 590–600.

87. Turner EL, Prague M, Gallis JA, Li F, Murray DM. Review of recent methodological developments in group-randomized trials: part 2-analysis. *Am J Public Health* 2017;**107**:1078–86.

88. Su F, Ding P. Model-assisted analyses of cluster-randomized experiments. *J R Stat Soc Ser B (Stat Methodol)* 2021;**83**: 994–1015.

89. Rao JN, Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics* 1992;**48**:577–85.

90. Lakshman RR, Sharp SJ, Ong KK, Forouhi NG. A novel school-based intervention to improve nutrition knowledge in children: cluster randomized controlled trial. *BMC Public Health* 2010;**10**:123.

91. Ukoumunne OC, Forbes AB, Carlin JB, Gulliford MC. Comparison of the risk difference, risk ratio and odds ratio scales for quantifying the unadjusted intervention effect in cluster randomized trials. *Stat Med* 2008;**27**:5143–55.

92. Leyrat C, Morgan KE, Leurent B, Kahan BC. Cluster randomized trials with a small number of clusters: which analyses should be used? *Int J Epidemiol* 2018;**47**:321–31.

93. Zou GY, Donner A. Extension of the modified Poisson regression model to prospective studies with correlated binary data. *Stat Methods Med Res* 2013;**22**:661–70.

94. Li F, Hughes JP, Hemming K, Taljaard M, Melnick ER, Heagerty PJ. Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: an overview. *Stat Methods Med Res* 2021;**30**:612–39.

95. Hooper R, Forbes A, Hemming K, Takeda A, Beresford L. Analysis of cluster randomized trials with an assessment of outcome at baseline. *BMJ* 2018;**360**:k1121.

96. Kahan BC, Forbes G, Ali Y *et al*. Increased risk of type I errors in cluster randomized trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. *Trials* 2016;**17**: 438.

97. 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.

98. 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;**34**:3531–45.

99. Li P, Redden DT. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Stat Med* 2015;**34**:281–96.

100. McNeish D, Stapleton LM. Modeling clustered data with very few clusters. *Multivariate Behav Res* 2016;**51**:495–518.