

Best (but oft-forgotten) practices: designing, analyzing, and reporting cluster randomized controlled trials^{1,2}

Andrew W Brown,^{3,4*} Peng Li,³ Michelle M Bohan Brown,⁶ Kathryn A Kaiser,^{3,4} Scott W Keith,⁷ J Michael Oakes,⁸ and David B Allison^{3-5*}

³Office of Energetics, ⁴Nutrition Obesity Research Center, and ⁵Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL; ⁶Department of Food, Nutrition, and Packaging Sciences, Clemson University, Clemson, SC; ⁷Division of Biostatistics, Department of Pharmacology and Experimental Therapeutics, Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, PA; and ⁸Division of Epidemiology & Community Health, School of Public Health, University of Minnesota, Minneapolis, MN

ABSTRACT

Cluster randomized controlled trials (cRCTs; also known as group randomized trials and community-randomized trials) are multilevel experiments in which units that are randomly assigned to experimental conditions are sets of grouped individuals, whereas outcomes are recorded at the individual level. In human cRCTs, clusters that are randomly assigned are typically families, classrooms, schools, worksites, or counties. With growing interest in community-based, public health, and policy interventions to reduce obesity or improve nutrition, the use of cRCTs has increased. Errors in the design, analysis, and interpretation of cRCTs are unfortunately all too common. This situation seems to stem in part from investigator confusion about how the unit of randomization affects causal inferences and the statistical procedures required for the valid estimation and testing of effects. In this article, we provide a brief introduction and overview of the importance of cRCTs and highlight and explain important considerations for the design, analysis, and reporting of cRCTs by using published examples. *Am J Clin Nutr* 2015;102:241–8.

Keywords: community randomized trial, group randomized trial, intraclass correlation coefficient, power analysis, reporting fidelity

INTRODUCTION

Cluster randomized controlled trials (cRCTs⁹; also known as group randomized trials and community-randomized trials) are multilevel experiments in which units that are randomly assigned to experimental conditions are sets of grouped individuals, whereas outcomes are recorded at the individual level. In human cRCTs, clusters that are randomly assigned are typically families, classrooms, schools, worksites, or counties. Although designs with >2 levels of clustering are possible (1), we restrict our attention in this article to designs with only 2 levels.

With growing interest in community-based, public health, and policy interventions to reduce obesity or improve nutrition, the use of cRCTs has increased. cRCT designs are often appropriate when the nature of the intervention requires that it be imple-

mented for an entire cluster at once. Examples include the assignment of entire schools, medical practices, or communities to interventions related to, for example, school food programs, physician interaction styles, or the building of sidewalks, respectively. These types of studies often meet the general characteristics of interventions, as described by Murray, that “operate at a [cluster] level, manipulate the physical or social environment, or cannot be delivered to individuals” (2). In addition, the practicality of recruiting and delivering weight-loss or nutrition interventions to large numbers often necessitates focused efforts at geographic or health provider levels (3). Therefore, such approaches are sometimes best studied with cRCT designs and, in turn, necessitate appropriate analyses.

Errors in the design, analysis, and interpretation of cRCTs are unfortunately all too common. This situation seems to stem in part from investigator confusion about how the unit of randomization affects causal inferences and the statistical procedures required for the valid estimation and testing of effects. This article aims to serve as a brief introduction and overview to the importance of cRCTs as well as to highlight and explain important considerations for the design, analysis, and reporting of cRCTs.

¹ Supported in part by NIH grants P30DK056336, R25DK099080, and R25HL124208.

² Supplemental Material: Simulation and Supplemental Material: Compliance are available from the “Supplemental data” link in the online posting of the article and from the same link in the online table of contents at <http://ajcn.nutrition.org>.

*To whom correspondence should be addressed. E-mail: awbrown@uab.edu (A Brown), dallison@uab.edu (D Allison).

⁹ Abbreviations used: CONSORT, Consolidated Standards of Reporting Trials; cRCT, cluster randomized controlled trial; ICC, intraclass correlation; RCT, randomized controlled trial.

Received December 9, 2014. Accepted for publication April 21, 2015.

First published online May 27, 2015; doi: 10.3945/ajcn.114.105072.

CLARIFYING WHAT DEFINES A cRCT

Figure 1 illustrates a flow diagram of 4 considerations for deciding whether the statistical principles described herein apply to a particular study. The first consideration is the unit of randomization, which is the unit to which treatments are assigned, but not necessarily the unit to which the treatments are applied. To be a cRCT, the cluster must be assigned to one of the intervention or control conditions, such as a school being assigned to a fruit-consumption intervention. Conversely, designs in which individuals are randomly assigned to different treatments within a cluster are not cRCTs but represent other designs such as split plots or multisite randomized controlled trial (RCT) (Figure 1, bubble 1). Such cases will not be considered here.

Next, it is important to consider the unit of treatment application. Some treatments can be administered to entire clusters, whereas others can be administered only to individuals within clusters. For example, in a school-based intervention where fruit is being provided, a school-wide program could be initiated in

which fruit is provided in kiosks throughout the school (a cluster-level application) or a piece of fruit could be distributed to every student at their desk (an individual-level application). In either case, both the unit of measurement and the unit of analysis must be at the individual level to be considered a cRCT. Cluster-wide measurements (e.g., school-wide vending machine purchases) preclude the study from being multilevel in nature because both units of randomization and measurement are the cluster (e.g., school). Cluster-wide measurement requires a different set of statistical considerations than for cRCTs, such as weighting cluster-level observations on the basis of inverse variance or cluster size. For analysis, if individual-level data (e.g., individual body weights) are collected and summarized for analysis (e.g., mean body weights for each cluster), the multilevel nature of the design is lost as is information related to individual-level variability.

In this article, we restrict our discussion to cRCT designs in which 1) random assignment occurs at the cluster level, 2) every participating individual within the cluster is subject to the same treatment whether the treatment is applied directly to individuals or the whole cluster, 3) measurements are taken at the individual level, and 4) the analysis is conducted at the individual level. Furthermore, when considering longitudinal designs, we restrict our discussion to designs in which repeated measurements are taken on the same individual. An alternative design of repeated cross-sections within clusters is also possible (2) but will not be discussed.

SPECIAL ISSUES IN DESIGNING, ANALYZING, AND REPORTING cRCTs

Power analysis and sample-size calculations depend on intraclass correlations

Because individuals in the same cluster usually are more similar to one another, on average, than are individuals from different clusters, cRCTs are statistically less efficient (i.e., have less power and precision per individual studied) than are non-clustered RCTs. The intraclass correlation (ICC), which is often denoted as ρ , can be interpreted as the proportion of total variance attributable to clustering (**Appendix A**, Additional details: ICC). Therefore, the ICC is bounded by 0 and 1 and is used to express how similar individuals within clusters tend to be. In extremes, an ICC approaching 0 indicates that individuals within clusters do not tend to be similar, and an ICC approaching 1 indicates nearly identical measurements within clusters. In most cRCTs, ICCs are between 0.001 and 0.05 (4); although such values may seem small, their effects on power and type I error rates can be very large. Therefore, ignoring ICCs in the sample-size calculation may result in underpowered designs, whereas ignoring ICCs in the resulting analysis may result in inflated type I error rates (also known as anticonservative tests or increased false-positive results) as we discuss in Analyzing cRCTs.

There are the following 2 sample-size choices in cRCTs: the total number of clusters (K) and the number of individuals per cluster (m). Although m can vary between clusters, we assumed equal cluster size for simplicity. As in the sample-size determination for nonclustered RCTs, the choices of K and m may sometimes depend on feasibility and cost. For instance, only a limited number of classrooms may be available, which limits K , whereas the number of students in a given class may limit m for each K . Increasing K where possible is almost always

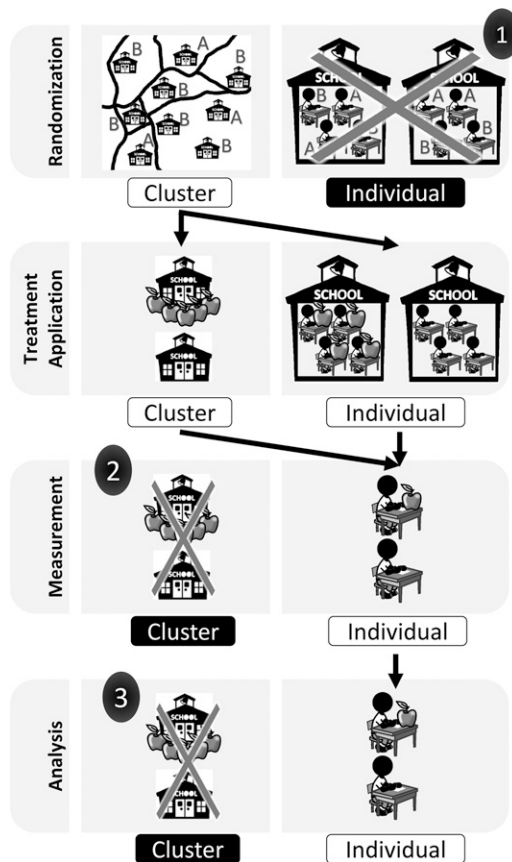


FIGURE 1 Flow diagram to illustrate an example process for a cRCT. Each of the 4 panels from top to bottom highlights different levels of design and analysis considerations. White text boxes indicate design characteristics consistent with cRCTs, whereas black text boxes denote characteristics inconsistent with cRCTs (coupled with the unit being crossed out with an X). A crossed out design does not indicate an invalid study design per se but, rather, indicates a design that is beyond the scope of this article. “A” and “B” in the randomization row represent intervention allocations that would later be assigned to an intervention or control condition. Bubble 1: Non-cRCT design, such as a multisite RCT. Bubble 2: Observation of cluster-level outcomes makes the design no longer a multilevel design. Bubble 3: Analysis of individual-level observations as cluster-level outcomes ignores the multilevel nature of data and requires a different set of considerations for analyzing data than addressed herein. cRCT, cluster randomized controlled trial; RCT, randomized controlled trial.

a more-efficient way to increase power than increasing m . For a 2-armed cRCT, the sample size required for adequate power can be estimated as described in Appendix A (Additional details: power analysis and sample-size calculation). Power curves under different K and m are illustrated in **Figure 2** for a cRCT that assumes a mean difference (d) of 0.25 with an SD of 1 (i.e., a standardized mean difference of 0.25) with an ICC (ρ) of 0.02, equal number of clusters per treatment ($K \div 2$), and equal m in each cluster. In this case, it would take 30 classrooms with 28 children per classroom (840 children total) to reach 80% power. If only 18 classrooms were available, each classroom would need 85 children (1530 children total) to achieve 80% power. For 12 classrooms, 80% power is impossible with power approaching 79% as m goes to infinity. Clearly, increasing the number of clusters is more efficient than including more individuals per cluster and sometimes is the only feasible way of attaining sufficient statistical power for a study.

Other design issues can increase or decrease power in cRCTs as well. Unequal m in each cluster or an unequal number of clusters assigned to each treatment will decrease power (5) just as an unbalanced allocation will decrease power in a nonclustered RCT (6). In contrast, similar to the analysis of nonclustered RCTs (7), incorporating covariates into the design and analysis of cRCTs can sometimes improve statistical power and reduce the sample size needed. Teerenstra et al. (8) showed that, if an ANCOVA (i.e., adjusting for baseline values of the dependent variable) is applied to the analysis of cRCTs, the sample size can be reduced by a factor of r^2 , where r is the correlation of the cluster mean between baseline and follow-up. Any baseline covariate that has a nontrivial correlation with the outcome variable, which is conditional on other covariates in the model, should increase power.

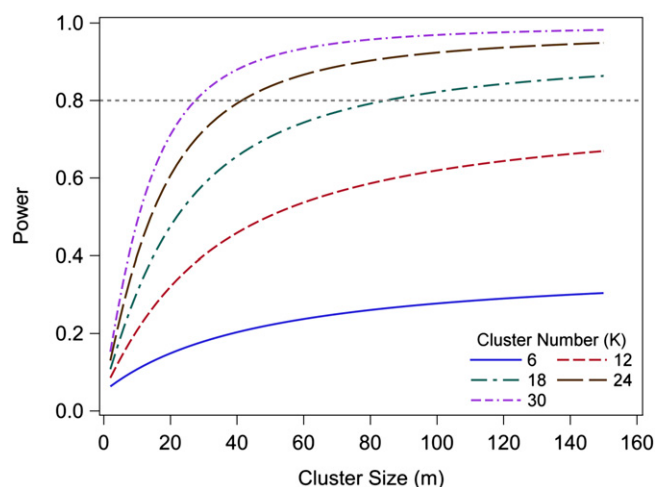


FIGURE 2 Power curves of an example, 2-armed cRCT as a function of number of clusters and individuals within cluster. Curves show the power to detect a 0.25-unit difference (d) of the outcome with an SD of 1 (i.e., a standardized mean difference of 0.25), ICC of 0.02, and $\alpha = 0.05$. The calculations assume an equal cluster size (m) and equal allocation ($K \div 2$) between treatments and use a t distribution. Although other values of ICC, d , and SDs may be used, the general shape of the lines will be qualitatively similar under different conditions, with greater K resulting in higher power than with lesser K for the same m . See Appendix A (Additional details: power analysis and sample-size calculation) for sample-size equations. ICC, intraclass correlation.

Implementation

Defining clusters

Clusters are composed of individuals belonging to only one of multiple, discrete groups such as classrooms in schools, schools in districts, and districts in cities. Clusters must have identifiable characteristics to be able to serve as useful units of random assignment such as geographic, structural, or jurisdictional boundaries. The ability for researchers to separate clusters from a higher level of aggregation is essential for implementing a cRCT.

A particularly pernicious and invalid design that requires recognition is the inclusion of only one cluster per condition (e.g., see reference 9) or even the inclusion of 3 control clusters but only one treatment cluster (e.g., see reference 10). If all clusters (e.g., schools) in a larger level of aggregation (e.g., a city) are assigned to the same treatment, the larger level of aggregation (in this case, the city) is the true level of clustering. For instance, Tarro et al. (10) assigned all schools (the purported cluster) in a single city to an education intervention and all schools in 3 other cities to the control condition. Although there were 24 intervention schools and 14 control schools, the actual level of clustering was the city, which meant that one cluster was assigned the treatment and 3 clusters were assigned to the control. Such designs are unable to support any valid analysis for an intervention effect, absent strong and untestable assumptions (11, 12). In such designs, the variation that is due to the cluster is not identifiable apart from the variation due to the condition. A one-cluster-per-condition design is analogous to assigning one person to the treatment and one person to the control in an ordinary (nonclustered) RCT, measuring each person's outcome multiple times, treating the multiple observations per person like independent observations, and interpreting the results like a valid RCT. In such a situation, the observations on person A can be tested as to whether they are significantly different from those on person B but cannot support an inference about the effect of treatment per se.

Selecting individuals within clusters

How individuals are selected within clusters for outcome measurements is dependent on the study design and outcomes of interest. For clusters in which all individuals within a cluster participate [sometimes referred to as "complete enumeration" (13)], all participants contribute data to the analysis. In contrast to complete enumeration, subsampling is sometimes used, particularly when measurements are expensive. Subsamples should ideally be a representative, random sample of the cluster unless researchers deliberately stratify by factors such as sex or race.

Informed consent

Participant consent and ethical considerations in cRCTs can sometimes differ from in nonclustered RCTs (14). Consent can be sought at the individual level, the cluster level, or both levels depending on local review board requirements and the study design and should be sought before random assignment to discourage selective participation on the basis of the treatment assignment. Sometimes when complete enumeration is attempted, individuals within clusters may refuse or neglect to provide written consent. For instance, Grydeland et al. (15) invited all sixth graders in 37 schools to participate but only received

parent-signed consent forms from 73% of pupils. Depending on the design, cluster-level interventions may still be able to be applied across consenting and nonconsenting individuals (e.g., a classroom teaching strategy), without the collection of individual-level data from nonconsenting individuals. The process of obtaining human subject approval from an ethics board (e.g., and institutional review board) before beginning a cRCT should help clarify data-collection and consenting-related issues (16).

Intervention contamination

One popular justification for the use of a cRCT design is to try to decrease contamination across experimental conditions. By contamination, we mean situations in which there is an experimentally unwanted cross-exposure in interventions such as one intervention group receiving or being exposed to all or part of another intervention. One infamous example of such contamination is the report in 1931 of the Lanarkshire milk experiment (17). In this experiment, 67 schools with 20,000 students were tasked with randomly assigning students to receive a milk intervention or serve as a control within each school. There is evidence that the ill-nourished children were selected (as opposed to randomly assigned) for the milk condition, whereas the well-fed children were distributed to the control condition. In this way, milk and control conditions were contaminated. Greater control over the implementation may have prevented some such contamination, but as Gossett [writing as Student (17)] pointed out, the allocation was likely “swayed by the very human feeling that the poorer children needed the milk.” The randomization of each school to either the control or intervention might have created a greater barrier to this type of contamination.

In other cases, attempts at preventing contamination may make little sense and may reduce the power of the study design. For instance, Mucklebauer et al. (9) installed water fountains in schools in one of 2 cities, thus making the cluster level the city (i.e., a one-cluster-per-condition design). The authors stated, “Randomization was performed at the city level to minimize contamination...” (9). In this case, contamination between schools from installing drinking fountains was likely minimal, whereas the assignment of all schools in only one city per intervention made causal statistical inferences impossible and the study invalid.

Allocation concealment and random assignment

Methods for random assignment for cRCTs can be similar to those for nonclustered RCTs except that random assignment occurs at the cluster rather than individual level. Similar to an analysis of individuals in nonclustered RCTs, clusters can be paired or stratified to try to account for particularly important cluster-level covariates or confounders a priori. The concealment of treatments from participants in cRCTs can also be similar to that in nonclustered RCTs, running the gamut from completely concealed treatments (e.g., placebo pills and cluster-level consent) to completely unconcealed (e.g., most dietary interventions). In most cases, treatments applied to clusters should be concealed or obscured, if possible. In situations where concealment from participants is not possible, it is often still possible to conceal treatments from the investigators assigning or delivering the treatments, those taking measurements, or those analyzing data. As described for the case of the Lanarkshire milk

experiment, allocation concealment may help prevent the introduction of biases into the experimental implementation or analysis.

Analyzing cRCTs

Impact of ignoring clusters

One of the strengths of cRCTs is the ability to obtain individual-level data despite random assignment at the cluster level. Although the calculation of summary statistics for each cluster instead of the use of individual-level data can simplify analyses, it comes at the cost of power and individual-level inference. As described briefly with respect to Figure 1, there are also different statistical considerations that are beyond our scope here. For the remainder of this discussion, we will assume data are analyzed at the individual level, keeping the nested multilevel structure of clustered individuals intact.

One inappropriate but common method of simplifying cRCT analyses is to ignore clustering and compare all individuals in intervention clusters with all those in control clusters. This approach ignores the fact that individuals within a particular cluster tend to be more similar to each other than to members of other clusters (i.e., ICC). Such analyses can underestimate variance and overestimate the significance of differences, as we show through simulation (Figure 3; see **Supplemental Material: Simulation for methods**). In one example (15), the authors stated that “As only 2% of the variance in BMI and waist circumference was explained by [cluster], we did not adjust for clustering in our analysis.” However, our simulations of significance tests from their design showed that, with an ICC of only 0.005, the type I error rate was significantly greater than the nominal 5% for their tests. At an ICC of 0.02 (the 2% variance previously described), the type I error rate for their analysis of only girls inflated to ~9%, which was nearly twice the nominal 5% the authors assumed. This effect was exacerbated when the full sample (boys and girls) was considered, where the type I error rate was inflated by nearly a factor of 3 to ~14%. Similarly, in a second example, Bere et al. (18) did not report the results of adjusting for clusters in their analysis and concluded that “free school fruit might contribute to prevent future excessive weight gain” on the basis of a *P* value of 0.04. Our simulated analysis of this study showed that at an ICC of only 0.02 resulted in an inflated type I error rate of ~6.5% for the reported test, whereas an ICC of 0.05 raised the error rate to almost 10%. Although methods to adjust statistical tests for the anticipated type I error rate inflation have been proposed (e.g., see reference 19), other authors have shown this approach does not necessarily correct the problem (20). Moreover, there can be complex issues involving df and whether asymptotic *z* or finite sample *t* distributions as described in Appendix A (Additional details: power analysis and sample-size calculation) can or should be relied on.

Brief description of statistical procedures for accounting for clustering

To account for ICCs, the following 2 types of statistical models are widely used for individual-level analyses of cRCTs: 1) conditional or subject-specific models, also called mixed-effects models, and 2) marginal or population-averaged models that use the generalized estimating equations approach. These models

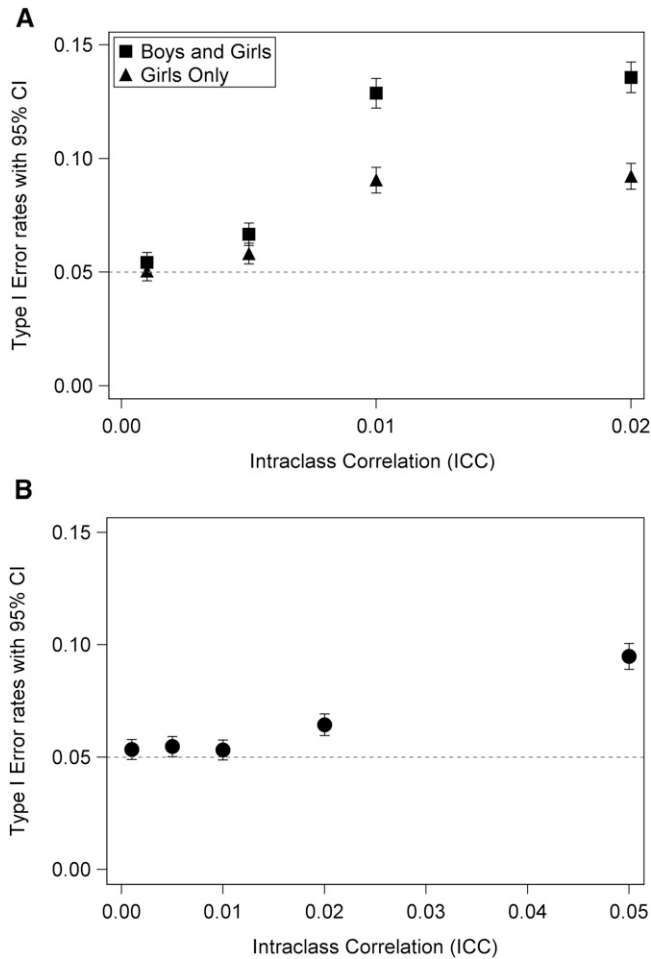


FIGURE 3 Simulation studies showed that ignoring ICC in analyses can affect the underlying type I error rate. Horizontal dotted lines demarcate the desired nominal $\alpha = 0.05$ significance level. A: Grydeland et al. (15) observed an ICC of 0.02 but excluded it from their analyses. They reported a significant difference in the total sample and in a subanalysis of girls only. ICCs ≤ 0.02 may still affect the type I error. B: Bere et al. (18) did not report the ICC of their clusters; nor did they report an analysis that accounted for clustering. A reasonable range of ICCs from 0.001 to 0.05 were selected, and type I error rates were estimated from 10,000 independent simulated datasets per ICC and illustrated with 95% CIs. See Supplemental Material: Simulation for methods. ICC, intraclass correlation.

differ in the way that the correlation of measurements within clusters is incorporated in the model and also in the interpretation of estimated model variables. In the population-averaged model, the treatment-effect estimate is the average change across the entire population under different intervention conditions. In the subject-specific model, the treatment-effect estimate is specific to a given cluster because the treatment-effect estimate is conditional on the specific characteristics of that cluster. For large numbers of clusters (i.e., as K goes to infinity), both models converge to yield similar results, and in practice, we showed the 2 approaches provided similar estimates. Cluster- and individual-level covariates, such as baseline measures before random assignment, can be incorporated in either model. In most cases, it is advisable to adjust for baseline covariates for which there is expected to be a nontrivial correlation with the outcome variable conditional on other covariates, but this should be decided a priori and not on the basis of data observed after

collection. The validity of inferences from a given analysis is dependent on sufficiently large K (per the central limit theorem) or under the usual parametric modeling assumptions, including the normality of residuals. Again, there are special issues when the number of clusters is small (e.g., <10), and not all statistical methods will perform well in such circumstances (21, 22).

Despite these special circumstances, researchers can use the statistical methods we describe to improve the accuracy and precision of effect estimation as well as design and analysis decision making in the context of cRCTs. In the common case where the outcome variable is continuous and measured both at baseline before random assignment and at the end of the cRCT protocol, an ANCOVA model is an apt choice as long as it also accounts for the ICC of the outcome of interest. In such cases, it is recommended that the baseline measure be used as a covariate to appropriately estimate the conditional intervention effect, and then, a random effect is estimated for the cluster-level indicator variable.

Accounting for treatment compliance and missing data

One challenge during analysis is deciding what to do when participants or entire clusters were noncompliant or partially compliant. At least 3 common strategies exist. One strategy is to exclude noncompliers, which is also called a per protocol analysis, which eliminates any existing data for noncompliant individuals. A second method is called an as-treated analysis, where the data are analyzed on the basis of what treatment a participant or cluster actually received rather than on the basis of random assignment. For instance, consider a school that was randomly assigned to the treatment condition but refused to implement it. If the school still contributed outcome measurements, it would be analyzed with the control group. Although this method retains all observations, it breaks the randomization. Both of these methods present an additional challenge in cRCTs compared with non-cRCTs because researchers will need to decide whether the analysis plan relates to compliance at the individual level, cluster level, or both. A third method is known as an intent-to-treat analysis, which analyzes observations on the basis of the randomized treatment assignment regardless of compliance. Intent-to-treat analyses tend to minimize type I errors while making the final outcome estimates relevant to what the expected impact of the treatment may be across a population instead of just in subjects who are able to comply with treatment.

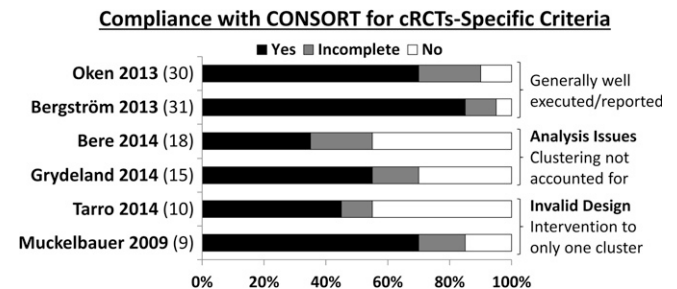


FIGURE 4 Ratings of compliance with specific criteria for the CONSORT for cRCTs. For each criterion, articles were assigned “Yes” if it was adequately reported; “Other” if it was partially, incompletely, unclearly, or implicitly reported; or “No” if it was not reported. Only items specific to cRCTs were rated. See Supplemental Material: Compliance for methods. CONSORT, Consolidated Standards of Reporting Trials; cRCT, cluster randomized controlled trial.

Another challenge during analysis is deciding how to account for missing data. The attempt to account for missing data is a better option than the alternative of the use of a completers-only analysis (23). A completers-only analysis eliminates individuals or even entire clusters from the analysis and, therefore, reduces statistical power and tends to bias estimates. To maintain the level of statistical power and unbiased estimates, a number of methods have been developed to estimate or account for missing values. Methods have improved markedly compared with limited methods such as the “last observation carried forward” or “baseline observation carried forward,” in which missing data are set to previous values (24–27). Two such improved methods include mixed models and multiple imputation, which use nonmissing data to inform the model but in different ways. In either case, missing data patterns should be evaluated for the type of missingness, including 1) missing completely at random, such as the loss of laboratory measurements because of instrument malfunction where observations are simply missing randomly across participants and clusters; 2) missing at random, such as more missing BMI measurements in female than male subjects in a survey, and after conditioning on sex, BMI measurements are missing randomly across participants and clusters; or 3) missing not at random, such as participants dropping out nonrandomly from a weight-loss trial for a reason that cannot be accounted for by any study variable measured. Methods to account for missing data are typically predicated on the data being missing completely at random or missing at random but not missing not at random and need to be carefully considered (28). For an example of a data-analysis plan for missing data in an cRCT, see reference 29.

Reporting cRCTs

One of the principles of scientific communication is that research should be reported in a transparent and thorough manner such that it can be replicated. The clustered nature of cRCTs requires additional information to be reported compared with for nonclustered RCTs. The Consolidated Standards of Reporting Trials (CONSORT) guidelines comprise a list of best reporting practices for randomized trials, and in 2012, the guidelines were extended for cRCTs (4). Several key components, many of which we previously described, are different for cRCTs. For instance, item 7a of the extended guidelines specifies researcher should provide the “Method of calculation, number of cluster(s) (and whether equal or unequal cluster sizes are assumed), cluster size, a coefficient of intracluster correlation (ICC or ρ), and an indication of its uncertainty [e.g., the standard error of the ICC].” Item 7a for the original CONSORT simply directs the reporting of “How sample size was determined.”

Guidelines such as the CONSORT guidelines are useful in 2 ways as follows: first, as a guide for authors who want to be thorough and transparent in their reporting; and second, as a tool for peer reviewers and journal editors to ensure the complete reporting of all information necessary to interpret and reproduce the research. In **Figure 4**, we summarize our ratings of compliance with the CONSORT guidelines for a cRCT for the 4 publication examples previously discussed (9, 10, 15, 18) plus 2 more examples that generally did a sufficient job designing, implementing, and reporting cRCTs (30, 31; see **Supplemental Material: Compliance for methods**). Even fatally flawed

studies can be reported in a manner compliant with many of the reporting guidelines, although it is debatable whether they merit being reported. Although a priori study registration is one way of communicating items in the CONSORT for cRCT guidelines, reporting the items along with the results in a publication contemporaneously provides readers with the information needed to understand and fully evaluate the results.

In conclusion, cRCTs are an increasingly popular design choice used to address nutrition, obesity, and other research questions. Building on the theory and methods applied to non-clustered RCTs can make conducting appropriately powered cRCTs feasible, but care is needed in the design, analysis, and reporting of cRCTs to make them valuable to science and helpful to inform policy.

We thank John A Dawson for his input and feedback.

The authors' responsibilities were as follows—all authors: contributed to the content of the article, critical review of the content, editing of drafts, and approval of final draft. The opinions in this article are those of the authors and not necessarily those of the NIH or any other organization.

None of the authors reported a conflict of interest.

REFERENCES

- Teerenstra S, Moerbeek M, van Achterberg T, Pelzer BJ, Borm GF. Sample size calculations for 3-level cluster randomized trials. *Clin Trials* 2008;5:486–95.
- Murray DM. Design and analysis of group-randomized trials. Oxford (United Kingdom): Oxford University Press; 1998.
- Sim LJ, Parker L, Kumanyika SK. Bridging the evidence gap in obesity prevention: a framework to inform decision making. National Academies Press; 2010.
- Campbell MK, Piaggio G, Elbourne DR, Altman DG, Group C. Consort 2010 statement: extension to cluster randomised trials. *BMJ* 2012; 345:e5661.
- 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.
- Allison DB, Allison RL, Faith MS, Paultre F, Pi-Sunyer FX. Power and money: designing statistically powerful studies while minimizing financial costs. *Psychol Methods* 1997;2:20.
- Allison DB. When is it worth measuring a covariate in a randomized clinical trial? *J Consult Clin Psychol* 1995;63:339–43.
- 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.
- Muckelbauer R, Libuda L, Clausen K, Toschke AM, Reinehr T, Kersting M. Promotion and provision of drinking water in schools for overweight prevention: randomized, controlled cluster trial. *Pediatrics* 2009;123:e661–7.
- Tarro L, Llauro E, Albaladejo R, Morina D, Arija V, Sola R, Giralt M. A primary-school-based study to reduce the prevalence of childhood obesity—the EdAl (Educacio En Alimentacio) study: a randomized controlled trial. *Trials* 2014;15:58.
- Varnell SP, Murray DM, Baker WL. An evaluation of analysis options for the one-group-per-condition design. can any of the alternatives overcome the problems inherent in this design? *Eval Rev* 2001;25: 440–53.
- Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health* 2004;94:423–32.
- Campbell MJ. Cluster randomized trials in general (family) practice research. *Stat Methods Med Res* 2000;9:81–94.
- Hutton JL. Are distinctive ethical principles required for cluster randomized controlled trials? *Stat Med* 2001;20:473–88.
- Grydeland M, Bjelland M, Anderssen SA, Klepp KI, Bergh IH, Andersen LF, Ommundsen Y, Lien N. Effects of a 20-month cluster randomised controlled school-based intervention trial on BMI of school-aged boys and girls: the HEIA study. *Br J Sports Med* 2014;48: 768–73.

16. Oakes JM. Risks and wrongs in social science research. An evaluator's guide to the IRB. *Eval Rev* 2002;26:443–79.
17. Student. The Lanarkshire milk experiment. *Biometrika* 1931;23:398–406.
18. Bere E, Klepp KI, Overby NC. Free school fruit: can an extra piece of fruit every school day contribute to the prevention of future weight gain? A cluster randomized trial. *Food Nutr Res* 2014;58.
19. Kish L. Survey sampling. New York: Wiley; 1965.
20. Musca SC, Kamiejski R, Nugier A, Meot A, Er-Rafiy A, Brauer M. Data with hierarchical structure: impact of intraclass correlation and sample size on type-i error. *Front Psychol* 2011;2:74.
21. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997;53:983–97.
22. 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.
23. Díaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clin Trials* 2014;11:590–600.
24. Cresswell L, Mander AP. How to use published complete case results from weight loss studies in a missing data sensitivity analysis. *Obesity (Silver Spring)* 2014;22:996–1001.
25. Elobeid MA, Padilla MA, McVie T, Thomas O, Brock DW, Musser B, Lu K, Coffey CS, Desmond RA, St-Onge MP, et al. Missing data in randomized clinical trials for weight loss: scope of the problem, state of the field, and performance of statistical methods. *PLoS ONE* 2009;4:e6624.
26. Gadbury GL, Coffey CS, Allison DB. Modern statistical methods for handling missing repeated measurements in obesity trial data: beyond LOCF. *Obes Rev* 2003;4:175–84.
27. Kaiser KA, Affuso O, Beasley TM, Allison DB. Getting carried away: a note showing baseline observation carried forward (BOCF) results can be calculated from published complete-cases results. *Int J Obes (Lond)* 2012;36:886–9.
28. White IR, Carpenter J, Horton NJ. Including all individuals is not enough: lessons for intention-to-treat analysis. *Clin Trials* 2012;9:396–407.
29. Hunsberger S, Murray D, Davis CE, Fabsitz RR. Imputation strategies for missing data in a school-based multi-centre study: the Pathways study. *Stat Med* 2001;20:305–16.
30. Oken E, Patel R, Guthrie LB, Vilchuck K, Bogdanovich N, Sergeichick N, Palmer TM, Kramer MS, Martin RM. Effects of an intervention to promote breastfeeding on maternal adiposity and blood pressure at 11.5 y post-partum: results from the Promotion of Breastfeeding Intervention Trial, a cluster-randomized controlled trial. *Am J Clin Nutr* 2013;98:1048–56.
31. Bergström H, Hagstromer M, Hagberg J, Elinder LS. A multi-component universal intervention to improve diet and physical activity among adults with intellectual disabilities in community residences: a cluster randomised controlled trial. *Res Dev Disabil* 2013;34:347–57.

APPENDIX A

Additional details: ICC

There are ≥ 6 ways to estimate ICCs, with one method that is applicable to cRCTs. Shrout and Fleiss (1) wrote a seminal article regarding ICCs by using judges and their targets as an example. In cRCTs, instead of judges, we have clusters, and instead of targets, we have individuals. With the use of the terminology of Shrout and Fleiss (1), ICCs in cRCTs are classified as ICC(1) and particularly ICC(1,k). With the use a variance component notation, the ICC in cRCTs can be expressed as

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2} \quad (A1)$$

where σ_b^2 represents between- (or among-) cluster variance, and σ_e^2 represents within-cluster variance.

The unique challenge in a cRCT power analysis (see Additional details: power analysis and sample-size calculation) lies in

the estimation of the ICC. The ICC can sometimes be estimated from previously reported studies. However, the imprecision of the ICC in previous studies, especially in small studies, may result in overpowered or underpowered designs. A practical approach is to create a curve of possible ICCs on the basis of the most-relevant previously published studies or pilot data when available. The selection of more-conservative (i.e., larger) ICCs for the power calculation can help prevent the underestimation of sample sizes. As Turner et al. (2) pointed out, a design with a large number of clusters has less chance of having low power caused by the imprecision of the ICC estimation. Many modern statistical software packages can calculate ICCs.

Additional details: power analysis and sample-size calculation

The total sample size in a balanced cluster design (K clusters with m individuals per cluster and $K \div 2$ clusters per treatment) to be able to detect a difference between control and intervention groups (d) in a continuous outcome (such as BMI) with an estimated variance σ^2 at a 2-tailed α significance level with power $1 - \beta$ calculated as

$$Km = \frac{4 \left(t_{\alpha/2, K-2} + t_{\beta, K-2} \right)^2 \sigma^2 [1 + (m-1)\rho]}{d^2} \quad (A2)$$

where $t_{\alpha/2, K-2}$ and $t_{\beta, K-2}$ denote the upper 100 $\frac{\alpha}{2}$ and 100 β t distribution with $K - 2$ df, respectively. When K is small, the df available to estimate the cluster-level effects are also small. Equation A2 is valid with sufficiently large K (per the central limit theorem) or under usual statistical assumptions, including the normality of residuals. In the most-extreme case of one-cluster-per-treatment, no df are available for hypothesis testing and, therefore, represent invalid designs.

Another method is to begin with a nonclustered estimate of sample size (N_I) and account for the influence of the clustered design effect (DE). For clusters of equal size m

$$DE = 1 + (m - 1)\rho \quad (A3)$$

Therefore, DE accounts for the similarity of observations within clusters. The required total sample size for a cRCT (N_C) can be estimated by inflating the sample size required for a nonclustered RCT (N_I) with DE as follows:

$$N_C = N_I \times [1 + (m - 1)\rho] = N_I \times DE \quad (A4)$$

In these examples, m was assumed to be constant across clusters. In cRCTs that have variable cluster sizes, DE can be modified as follows (3):

$$DE = 1 + [(cv^2 + 1)\bar{m} - 1]\rho \quad (A5)$$

where cv is the CV for sample size and can be calculated as

$$s_m \div \bar{m} \quad (A6)$$

where s_m and $\bar{m}$ are the SD and mean of the cluster size, respectively. In practice, the cv can be approximated by using

estimates of the maximal ($m_{\max}$), minimal ($m_{\min}$), and average ($\bar{m}$) cluster sizes as

$$(m_{\max} - m_{\min}) \div 4\bar{m} \quad (A7)$$

if the actual cluster sizes are not available during the study design. It follows that larger sample sizes are required for unbalanced designs. However, the effect of adjustment for the variable cluster size on the sample size is negligible when the cv is less than 0.23 (3). A rule of thumb is that the power will not increase appreciably once m or $\bar{m}$ exceeds $1 \div \rho$ when K is held constant (4). Other estimation methods exist, such as empirical simulation-based estimates, which provide many ways to ensure that an appropriately powered study is conducted.

Additional reading

For more information about cRCTs, consider reading: general overviews (4–8), ethical considerations (6), analysis of cRCTs (8–10), accounting for missing data in cRCTs (11, 12), ICCs (1, 13), or reporting on cRCTs (14).

APPENDIX A REFERENCES

1. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86(2):420–8.
2. Turner RM, Prevost AT, Thompson SG. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Stat Med* 2004;23(8):1195–214.
3. 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(5):1292–300.
4. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and statistics in medicine. *Stat Med* 2007;26(1):2–19.
5. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: A review of recent methodological developments. *Am J Public Health* 2004;94(3):423–32.
6. Hutton JL. Are distinctive ethical principles required for cluster randomized controlled trials? *Stat Med* 2001;20(3):473–88.
7. Hannan PJ. Experimental social epidemiology—controlled community trials. 1st ed. In: Oakes JM, Kaufman KS, eds. *Methods in Social Epidemiology*. San Francisco (CA): Jossey-Bass; 2006:341–69.
8. Eldridge S, Kerry S. *A practical guide to cluster randomised trials in health services research*. West Sussex (United Kingdom): John Wiley & Sons; 2012.
9. Preisser JS, Young ML, Zaccaro DJ, Wolfson M. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Stat Med* 2003;22(8):1235–54.
10. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MH, White JS. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol* 2009;24(3):127–35.
11. White IR, Carpenter J, Horton NJ. Including all individuals is not enough: lessons for intention-to-treat analysis. *Clinical Trials* 2012;9(4):396–407.
12. Diaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clin Trials* 2014;11(5):590–600.
13. 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.
14. Campbell MK, Piaggio G, Elbourne DR, Altman DG, Group C. Consort 2010 statement: extension to cluster randomised trials. *BMJ* 2012; 345:e5661.