Journal of Statistical Theory and Practice, 9:685–698, 2015

ISSN: 1559-8608 print / 1559-8616 online
DOI: 10.1080/15598608.2014.992081

Cluster Randomized Trials: Considerations for Design and Analysis

HRISHIKESH CHAKRABORTY AND GENEVIEVE LYONS

Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, USA

Scientists often use randomized controlled trials to compare a newly developed treatment to the existing one, or to a placebo. Patients are randomly assigned to a treatment, and they are compared with respect to the outcome of interest. The cluster randomized trial (CRT) is a type of randomized controlled trial in which the treatments are randomized at the group, rather than individual, level. The intracluster correlation (ICC) measures the degree of similarity between individuals within clusters. CRTs can be designed in several ways; it is essential that researchers carefully plan the study, from sample size calculations to ICC calculation to analysis, in order to get valid and meaningful results. In this article we review and discuss the considerations essential to conducting a successful CRT using both frequentist and Bayesian approaches, and we discuss recent trends in CRT analysis, including highlighting new methodology for both binary and continuous data.

AMS Subject Classification: 62P10.

Keywords: Cluster randomized trials; Intracluster correlation; Statistical design; Sample size estimation; Analysis.

In a typical randomized trial, the treatments are randomly assigned to individuals. However, in a cluster randomized trial (CRT), treatments are randomized by a type of group (e.g., hospitals, clinicians, medical practices, schools, households, villages, communities, or administrative boundaries). The most recently published paper discussing aspects of CRT design and analysis was more than 10 years ago in 2004 (Murray et al. 2004), and some new methods have been developed since then. In this paper we review and discuss recent trends in CRT analysis, including highlighting new methodology for both binary and continuous data and considerations essential to conducting a successful CRT using both frequentist and Bayesian approaches.

Trials where clusters can be formed on the basis of geographic boundaries or a natural grouping have become the gold standard for the evaluation of new health interventions (Hayes and Bennett 1999; Chakraborty 2008b). Depending on these factors, there are several ways to plan and execute the trial. Most CRTs fall into one of three most common design types: completely randomized, stratified, and matched pair designs. According to a 2004 review, in the majority of CRTs, researchers used parallel groups to evaluate an

Received 21 May 2014; accepted 23 November 2014.
Address correspondence to: Hrishikesh Chakraborty, Department of Epidemiology and Biostatistics, Arnold School of Public Health, The University of South Carolina, 915 Green Street, Suite 449, Columbia, SC 29208, USA. Email: rishic@mailbox.sc.edu

intervention or drug effect (Eldridge et al. 2004). In a review of CRTs in primary care, 123 out of 152 trials used this two-arm parallel design (Eldridge et al. 2004; Eldridge and Kerry 2012). If the study includes many clusters, it is fine to use a completely randomized design in which interventions are assigned randomly to the clusters. However, if only a relatively small number of clusters were to be randomized this way, the risk is that the arms of the study would not be balanced with respect to baseline characteristics like sex or age.

The goal of stratified randomization is to ensure that treatment groups are balanced with respect to baseline characteristics or important factors (Kernan et al. 1999). In a stratified design, clusters are placed into groups (strata), which are then randomly assigned to one of the interventions. Kernan et al. (1999) and others assert that stratification is necessary to get valid results in small trials. Two benefits to this type of randomization are that (1) it helps ensure that each arm of the trial contains an approximately equal number of participants, and (2) cluster size could be a surrogate for within-cluster dynamics, which could in turn be associated with outcome. Other common factors for stratification are geographic area or socioeconomic status.

Matched pair design is analogous to a matched case-control study. It is an extreme form of stratification in which each stratum consists of only two clusters. Each one is randomly assigned to a different treatment. The main advantage of this design is that the study is carefully balanced with respect to important baseline risk factors (Donner 1998).

There are several advantages to using a cluster randomized design. Because groups are physically separated from each other, it is highly unlikely that an intervention will spread to the control group, a phenomenon called contamination (Chuang et al. 2002; Killip et al. 2004; Reading et al. 2000). This could potentially be a problem in trials of behavioral, rather than drug or medical, intervention, if participants are able to interact with each other (i.e., if they are not in separate groups). Cluster randomized designs can lower costs, partly because of increased administrative efficiency; they may eliminate potential ethical problems; they can provide a less intrusive randomization; and they may be the most convenient option (Reading et al. 2000; Donner and Klar 2004; Donner and Donald 1987). Finally, CRTs are virtually the only way to evaluate interventions that must be implemented at a group level, such as hospital-based or community-based interventions (Hayes and Bennett 1999).

Consistent with the advantages just described, the use of CRTs is specifically recommended as the best choice in some situations (Hayes and Bennett 1999). First, trials such as educational interventions or hospital procedure interventions should be implemented at the cluster level to avoid procedural or information contamination that could occur if the intervention were provided for some individuals but not others within the same hospital or group. Also, group-level randomization can reduce resentment among participants who may perceive that their benefits are inferior to those of other participants (Hayes and Bennett 1999). Second, without doing a CRT, it is difficult to capture the effect of applying an intervention to a community—for example, the effect of providing advanced training to birth attendants, on early neonatal mortality in a rural setting (Carlo et al. 2010). In vaccine trials, CRTs allow researchers to estimate these indirect group effects (Hayes et al. 2000; Halloran et al. 1997). Third, when the efficacy of an intervention has been established at the individual level, CRTs are useful to measure intervention effectiveness at the community level. Can the intervention successfully be applied in the group setting of interest?

The main disadvantage of CRTs is that researchers cannot assume independence of observations on individual members of a cluster; that is, in each cluster, participants may respond to the intervention similarly due to shared, cluster-level characteristics. This results in two potential sources of correlation: between-cluster correlation, which measures the variation in outcomes between clusters (Campbell et al. 2004), and within-cluster

correlation, which is correlation among members because they share similarities or characteristics, such as age, gender, geographic, socioeconomic, or political factors (Chuang et al. 2002; Killip et al. 2004). The within-cluster correlation is the cause of many of the challenges inherent to the design and analysis of CRTs.

A series of recent studies have generated some discussion weighing the benefits and drawbacks of CRTs versus individually randomized trials. In particular, Hewitt et al. (2008) pursued the question of whether contamination (described earlier) can be measured. If it were possible to measure and account for contamination, should researchers choose to do an individually randomized trial? A meta-analysis by Gilbody et al. (2008) sought to determine whether CRTs achieved more reliable results and thus whether the effort of conducting a CRT paid off. Comparing both cluster-randomized and individually randomized trials of a depression treatment intervention, they found that both types of trials achieved similar results in terms of effect size. This result suggests that researchers should design studies with care and thought (a principle that can hardly be disputed) and that more meta-analyses are necessary (Gilbody et al. 2008).

When planning trials, researchers should carefully weigh advantages of a design that randomizes groups, rather than individuals, against the disadvantages in terms of statistical power and cost. One of the main challenges of designing a CRT is that a large number of clusters may be needed, but it may be difficult to recruit enough study participants and clusters.

Inferences from Cluster Randomized Trials

In a CRT, researchers can draw inferences on the group or cluster level and on the individual level. For cluster-level inferences, the unit of analysis is the same as the unit of randomization; therefore, outcomes are assessed only at the level of the cluster. The outcome can be continuous (e.g., the percentage of individuals in the cluster who benefited) or dichotomous (e.g., "success" or "failure"). We give two examples next that illustrate inferences made on both the cluster and individual level.

The Guideline trial randomly assigned 19 hospitals in three urban districts of Argentina and Uruguay to intervention or a control group (Althabe et al. 2005). The intervention group received multifaceted behavioral training on new guidelines for episiotomy use during labor and delivery. The control group continued with standard in-service training activities. The outcomes assessed were use of oxytocin and episiotomies during the third stage of labor. In this trial, the unit of randomization and unit of analysis are the same.

The FIRST BREATH trial tested an education intervention given to birth attendants, randomized by community (Carlo et al. 2010). Birth attendants in communities in six different countries received the revised Essential Newborn Care training; however, those in the intervention clusters also received training in the American Academy of Pediatrics Neonatal Resuscitation Program. The outcome was 7-day neonatal mortality. Each community was a geographical area defined by physical or political boundaries, and each had an average of 500 births per year. Although the primary outcome (7-day mortality) was individual level, the researchers wanted to draw inferences at cluster and individual levels. Thus, the design and analysis accounted for both within- and between-cluster correlation.

Intracluster Correlation

The variance within a cluster is often less than what would be expected among randomly assigned individuals, because cluster members tend to share similarities. For example,

residents of the same community probably use many of the same resources, such as education, health care, and access to nutrition; therefore, their outcomes are more likely to be similar than for residents who do not share these resources. The intracluster correlation (ICC) is used to measure the degree of similarity in response from members of the same cluster. ICCs in most human studies are usually small and positive (between 0 and 1, but often less than 0.2), indicating the expected similarity among group members (Killip et al. 2004; Baskerville et al. 2001). The ICC is almost never negative, although this is theoretically possible. If it is negative, the researcher usually assumes a value of zero, which means there is no correlation. In this case the data can be analyzed using the methods for simple (individually randomized) trials.

The ICC, usually denoted by ρ (rho), can be calculated in several different ways. Fleiss's method uses a formula derived from mean square values in a one-way analysis of variance (ANOVA); this method has become quite popular (Fleiss 1986). Another popular method uses mixed models. Using Fleiss's method, ICC is calculated by dividing between-cluster variance by the total variance. An ICC of 1 indicates perfect agreement in the responses of individuals within the same cluster (Chuang et al. 2002). An ICC of 0 indicates no intraclass correlation, as described earlier.

Several other methods to estimate the ICC for binary data have been proposed and refined. These include estimators based on direct calculation of correlation within each cluster (Donner 1986; Karlin et al. 1981; Lipsitz et al. 1994), moment estimators (Kleinman 1973; Tamura and Young 1987; Williams 1982; Yamamoto and Yanagimoto 1992), extended quasi-likelihood and pseudo-likelihood estimators (Carrol and Ruppert 1988; McCullagh and Nelder 1989; Nelder and Pregibon 1987), and estimators with direct probabilistic interpretation (Fleiss and Cuzick 1979; Mak 1988). Chakraborty and Sen (2013) present a new method for estimating ICC, based on resampling with replacement and U-statistics. The method was tested in simulation studies using binary outcome data. When compared to the ANOVA method and an approach based on method of moments (Yamamoto and Yanagimoto 1992), Chakraborty's method estimated ICC values more precisely (Chakraborty and Sen 2013). It is true that there are disadvantages to resampling-based methods, which are discussed in that paper. In an extensive simulation comparison of several of these methods, researchers found that that the ANOVA estimators, some of the moment estimators, and one of the probabilistic estimators performed well in terms of bias and standard deviations (Ridout et al. 1999). Additional correlation models have been published, where the authors assumed beta binomial distribution (Kupper and Haseman 1978; Prentice 1986), a correlated binomial distribution (Kupper and Haseman 1978; Altham 1978), or a correlated probit distribution (Ochi and Prentice 1984). Wu et al. (2012) compare (via simulation) several methods for estimating ICC for binary clustered data. Then, using real data from cancer screening intervention trials, they concluded that researchers cannot assume the ICC is the same across clusters. They also demonstrated a method of estimating ICC from generalized estimating equations (GEE) (Wu et al. 2012).

More and more researchers are beginning to publish observed, stable ICC values, especially for sample-size calculation and study design purposes (Hedges and Hedberg 2007). In addition, registries or databases of ICC values, organized around particular types of data, are being made available. Two examples are a surgical ICC database (Cook et al. 2012), and a database established at the Northwestern Institute for Policy Research to aid in the design of education policy research (Online intraclass correlation database n.d.). This practice assumes a standard acceptable range of values for ICC, given a set of circumstances. For example one group provides their ICC estimates, both crude and adjusted, for several types of cancer screening, noting that their ICC values agree with previously published

cancer screening ICCs (Hade et al. 2010). However, another study summarized the ICCs from many CRTs in primary care settings with 10 or more clusters, and found that different datasets give a wide range of ICC values for the same variables (Adams et al. 2004). They found that using either individual or cluster characteristics to calculate an adjusted ICC will give a smaller ICC (Adams et al. 2004). Because of the wide range of ICCs, the researchers suggest that ICC is inherently uncertain at least in the primary care setting (Adams et al. 2004).

Recent advances in computing techniques have led to advances in Bayesian methods, which sometimes utilize a large amount of computing power. Bansal and Bhandary (2000) present a general Bayesian approach to estimation of ICC using a prior distribution of the eigenvectors of the covariance matrix. Turner, Omar, and Thompson (2001) consider the CRT situation with a binary outcome, comparing the results of using informative vs. non-informative priors. The potential benefit of using informative priors is that researchers may consider non-normal underlying distributions for the random effects underlying the clustering (Turner et al. 2001). They find that prior choice can affect both the bias (precision) and the variability of ICC estimates: using informative priors resulted in narrower credible intervals for the ICC. Ahmed and Shoukri (2010) demonstrate an alternative approach that is less computer-intensive than the Markov-chain Monte Carlo (MCMC) method and that works with both informative and noninformative priors. One drawback to their process is that it assumes equal cluster sizes, requiring additional computations if the clusters are uneven in size (Ahmed and Shoukri 2010). Spiegelhalter (2001) explores how prior choice affects results in randomized trials with continuous outcomes, finding that the use of non-informative priors has an impact, and calls for continued work in this area to determine the best practices. He recommends using an informative prior based on known information for the ICC. With regard to MCMC methods, this approach is promising but can be slow, and there is still a need for more research to confirm which steps yield convincing, reliable results (Spiegelhalter 2001). Bansal et al. (2013) compare a Bayesian approach to a non-Bayesian method presented by Srivastava (1984) for the case of unequal cluster (family) sizes. The Bayesian posterior distribution obtained for the ICC provided estimates that performed better in that they had a smaller MSE.

This review of Bayesian methods shows that researchers must plan carefully in order to get meaningful and valid results when using a Bayesian analysis. In particular, researchers should take care when choosing a prior distribution, understanding what the implications may be in terms of variance and bias. Misspecified priors could impact both accuracy and precision of the estimate. This is an area ripe for further methodological work.

Confidence Intervals for Intracluster Correlation

Confidence intervals for ICCs can be calculated in a number of ways. Several common methods use an approximation to the *F*-distribution: Thomas and Hultquist's procedure (Donner 1979; Thomas and Hultquist 1978), Fisher's transformation procedure (Fisher 1925), procedures by Smith (1956) and Swiger et al. (1964), and a maximum likelihood-based confidence limit (Donner and Koval 1980; Smith 1980a; Smith 1980b). For moderately large sample sizes, Donner and Koval (1983) showed that Fisher's transformation procedure approximates the true variance of ICC estimates with a great degree of accuracy and robustness. Donner and Wells (2013) compared several confidence interval estimators, using Monte Carlo simulation with a one-way random-effect model. They recommended Smith's method, which uses large-sample standard error of sample ICC (Smith 1980a; Smith 1980b), because it provides consistently good coverage for all ICC values.

Chakraborty's simulation-based technique for estimating ICC and its confidence interval is useful in planning trials because it gives a reasonable range of possible ICC values, and it can be used effectively when possible ICC values or number of clusters are not known in advance (Chakraborty et al. 2009a). It can be used for both binary and continuous outcomes.

Bayesians usually present credible intervals rather than confidence intervals. A credible interval uses an upper and a lower percentile of the posterior distribution as boundaries for the value of the parameter of interest, in this case the ICC. Rather than being "95% confident" of the true value of the parameter of interest, Bayesians can confidently say that there is a 95% probability that the ICC is between these values, because the credible interval is based on a probability distribution for the ICC.

Sample Size Estimation for Cluster Randomized Trials

Sample size calculation for a CRT is a little different than for individually randomized trials because the researcher must account for both within- and between-cluster variation. Because the sample size is directly proportional to the variance and the clustered design affects the variance, researchers can use standard sample size estimation methods, then multiply the result by the variance inflation factor (also called the design effect), $[1+ (m - 1) \rho]$. Here, m represents the cluster size and ρ represents the estimated ICC. Sometimes researchers use an upper 95% confidence limit for the estimated ICC instead of the estimate itself; this is a very conservative approach that leads to a larger sample size. For trials that will have varying cluster sizes, using the average cluster size in place of m will give a slight underestimate of sample size, and using the maximum cluster size will provide a conservative estimate (Donner and Klar 2000). If using Fleiss's method to estimate ICC and the average cluster size approximation, the sample size will be slightly underestimated; however, if the total number of participants is large, the effect will be negligible. Again, an alternative is to replace the average cluster size but using the largest expected cluster size; this gives a conservative sample size estimate (Donner et al. 1981). In this setting, a conservative estimate errs on the side of having more participants per cluster, to ensure sufficient power to detect a statistical difference (should one exist).

The design effect tells us the ratio of the number of participants required using cluster randomization to the number that would be required using individual randomization (Killip et al. 2004). This ratio will always be greater than one, although it can be very close to one. We can see from the design effect formula that there is a direct relationship between the ICC, the design effect, and sample size: a larger the ICC means that there is a larger design effect, therefore, more participants are needed (Kerry and Bland 1998).

If the CRT is to have a binary outcome, consideration should be given to the prevalence of the outcome, because it has an impact on the ICC (Gulliford et al. 2005). In recent work, it was found that higher prevalence is associated with higher ICC and sample variance (Gulliford et al. 2005). Turner, Prevost, and Thompson (2004) present a Bayesian method that can handle imprecision of ICC; this method provides an option for researchers to accurately calculate power and sample size. This method can be used when designing Bayesian CRTs.

Several methods for sample size estimation for CRTs must also be considered carefully. Lee and Durbin (1994) proposed a method for clustered binary data in which equal weights are assigned to all clusters (although in actuality cluster size may vary); this simplifies the derivation. A more recent formula, presented by Donner and Klar (2000), is used when cluster sizes are constant.

Using standard (nonclustered) sample size methods for a cluster randomized design may lead to seriously underpowered studies. This problem gets more severe as ICC increases and as average cluster size increases. To gain statistical power, it is more efficient to add clusters, rather than to increase the number of participants by adding to each cluster. Although it may be more feasible to add participants to existing clusters, this approach can only increase statistical power to a point (Donner 1998). As discussed earlier, this is one of the main challenges in CRT design.

The simulation method described earlier by Chakraborty et al. (2009a) can be used to estimate the sample size during the design phase of a CRT with dichotomous outcome. The method simulates the ICC estimate and its 95% confidence interval. It works for different cluster sizes and number of clusters. In general, during the design phase of cluster randomized, a common design effect across intervention groups is usually assumed; however, after the intervention period ends, the design effect can no longer be assumed. To compensate for this situation, Chakraborty et al. (2009b) demonstrated that the ICC value depends on three factors: effect size distribution, cluster size, and number of clusters. Under the assumption that the effect size changes overall at the end of the intervention period, the authors also showed how to adjust for the ICC value during the design phase. In a Bayesian analysis of a CRT, when specifying a prior distribution for ICC, researchers are essentially stating what they believe to be true about the ICC and describing their uncertainty about it (Spiegelhalter 2001). Based on this statement, researchers can calculate a reasonable corresponding sample size.

Analysis of Cluster Randomized Trial Data

Statistical methods for CRT analysis are not as well established as methods for individually randomized trials. Fisher's classical theory of experimental design assumes that the unit of analysis will be the same as the unit of randomization (McKinlay et al. 1989), so for cluster randomized data, standard (individually-randomized) analytic methods cannot be used (Donner 1998), and the analysis must take the clustering approach into account appropriately. Unit of analysis error occurs if an investigator incorrectly analyzes trial data as though the unit of randomization had been the individual participant, rather than the cluster, without accounting for clustering (Whiting-O'Keefe et al. 1984). Cornfield (1978) brought to the attention of the health research community the analytic implications of cluster randomization in 1978. Ignoring cluster randomization during data analysis will cause the within-cluster variance and between-cluster variance to be mixed, resulting in an underestimate of the overall variance. The p values will be inaccurately small and confidence intervals will be inaccurately narrow, potentially leading to false inferences of statistical significance (Donner and Klar 2000; Puffer et al. 2003; Schulz 1995). The likelihood of false statistical significance increases as ICC increases and as average cluster size increases (Donner and Klar 2000). Several studies have reported problems with the analysis and reporting of CRTs, especially related to the ICC (Donner and Klar 1994b; Chakraborty 2008a). Additional review articles have been published every several years as the methodology continues to advance (Donner 1998; Donner and Klar 2000; Donner and Klar 1996; Bland 2004; Campbell et al. 2007).

Researchers have developed statistical methods to analyze data for CRTs and to draw inferences both at the cluster level and at the individual level. For cluster-level analysis, just one summary measure (e.g., the cluster mean or proportion) is calculated from individual observations within a cluster; then standard statistical methods are used to analyze this summary measure as a primary observation. Of course, since the sample size of this test is now

equal to the number of clusters, the statistical power is lower and the degrees of freedom are reduced. Cluster-level CRT analysis is fully efficient if the clusters are all equal in size. To draw cluster-level inferences, researchers can use most of the standard statistical analysis techniques, for example, simple *t*-tests, weighted and unweighted linear regression, and random-effects meta-analysis. For example, two groups in a CRT can be compared with a *t*-test applied to a cluster-specific measure like the mean, weighted by the cluster size (Kerry and Bland 1998; Donner and Klar 1994a). Similarly, a paired *t*-test can be used to analyze continuous outcomes from a paired CRT. Since the paired *t*-test requires normality and homogeneity assumptions, some researchers prefer to use permutation tests(Gail et al. 1992; Maritz and Jarrett 1983); however, other researchers found that the *t*-test is remarkably robust when these assumptions are violated, even for fairly small samples (Donner and Klar 1996; Hereren and D'Agostino 1987). For stratified CRTs we need to adjust for the stratification factor in the analysis. However, the paired *t*-test accounting for design effect can only be used to test unadjusted primary outcome, so other methods such as GEE or mixed models can be used to adjust for covariates.

CRT analysis is analogous to meta-analysis (DerSimonian and Laird 1986; Thompson et al. 1997). In a meta-analysis, information from different groups (i.e., trials) is combined. In random effect meta-analysis, maximum likelihood estimation is used, and summary statistics are pooled across clusters defined by similar characteristics, rather than across studies (DerSimonian and Laird 1986; Thompson et al. 1997). The parameter estimates from different analyses are not hugely different unless the cluster sizes and/or outcome proportions are also hugely different. When drawing conclusions at the individual level, it is essential that researchers account for clustering. The simple analysis strategy is to treat individual data as independent observations (ignoring the clustering) and apply a standard statistical approach, then use the variance inflation factor (described previously), to adjust the variance before performing hypothesis tests (Donner and Klar 2000). The number of clusters, not the total number of individual participants, is the basis for determining the degrees of freedom of the revised test statistics.

Several statistical methods account for clustering while also allowing inference on individuals. For binary variables, the "sandwich" variance estimator may be used to estimate robust standard errors (Huber 1981; White 1980), which are modified as described by Rogers (1993) to allow for clustering. This means that the analysis might be based on a standard logistic regression with these robust standard errors; that is, the model adjusts the estimated standard errors to allow for the clustered study design. The regression coefficient estimators (e.g., log odds ratios) are the same as those in a standard logistic regression model because they are not affected by the procedure.

It is common to use mixed effect linear models with the generalized least-square method to analyze continuous outcome data from a completely randomized or a stratified CRT (Stiratelli et al. 1984; Ware 1985; Wolfinger and O'Connell 1993). Additionally, GEE extend the use of standard logistic regression to account for clustering. Researchers can specify different types of correlation matrices, and both the regression coefficients and their standard errors are accurate in the sense that they are consistently estimated (assuming a large enough sample size, and if robust standard errors are specified) (Liang and Zeger 1986). The parameter estimates from GEE do not correspond to the parameter estimates from random effects models, instead GEE parameter estimates are considered "population-averaged" interpretations because they are averaged across the cluster-level random effect values in the context of longitudinal data analysis (Williams 1982; Zeger and Liang 1992). Other analysis methods, such as multilevel modeling (Goldstein 1995), hierarchical linear

modeling (Bryk and Raudenbush 1992; Raudenbush 1997), and variance components analysis to analyze CRT data. Finally, there exist various sensitivity analysis methods, including the presentation of results using different analysis methods. Sensitivity analysis was used to assess the chosen analysis technique, or to test whether a non-cluster-based technique could have been used (Thabane et al. 2013). A recent CRT aimed at reducing prescription of unnecessary antibiotics tested two different interventions given to groups of general practitioners (Cals et al. 2009). The goal of the sensitivity analysis was to determine if the interventions were more effective among sub-groups of physicians, suggesting that the results were not due to the intervention but to latent characteristics (Cals et al. 2009). Thabane et al. (2013) found that only 26% of incorporated sensitivity analyses in a survey of January 2012 editions, but their paper is an excellent primer on the application of sensitivity analysis to clinical trials, and this may begin to change.

ICCs are seldom reported in published literature. For example, recent reviews found that ICC values were reported by only 2.0% to 8.6% of trials (6 of 149, 13 of 152, and 1 of 51, respectively) (Elddrige et al. 2004; MacLennan et al. 2003; Isaakidis and Ioannidis 2003). In a review of 54 papers on physicians' behavior published in a broad selection of journals, 70% used the wrong unit of analysis (Divine et al. 1992). In a study of 21 published reports of primary prevention trials, only 57% accounted for clustering in the analysis (Simpson et al. 1995). Other investigators have reported similar findings (Donner et al. 1990; Ennett et al. 1994). However, there is evidence that there is improvement in CRT reporting. Perhaps because CRTs are becoming more common, researchers may be more familiar with the methodology and are, accordingly, doing more accurate work (Bland 2004; Campbell et al. 2007). Bland (2004) suggests that 1998 is a benchmark year after which improvements became apparent. Handlos, Chakraborty, and Sen (2009) found that between 1998 and 2008, in CRTs related to maternal and child health in developing countries 29% did not adjust for clustering in their sample size calculations, and 20% did not adjust for clustering in their analysis.

A Bayesian approach to the analysis of CRTs is beneficial because it does not require assumptions of normality to model the random effects (Turner et al. 2001). In general, the Bayesian approach involves specifying prior distributions for parameters (such as the variance components), then either directly deriving posterior distributions analytically, or indirectly deriving them using MCMC methods (Turner et al. 2001). It is often easier, and much more common, to use MCMC methods rather than to solve it via integration (Spiegelhalter 2001).

Discussion

In this paper we have discussed many of the issues related to the design and analysis of a CRT, including the nature of clusters and the inherent similarity among cluster members. To summarize, the ICC measures this similarity and thus plays an important role in the design and analysis of CRTs. Although it may seem small enough not to have much impact, failing to account for ICC can result in underestimated standard errors for intervention effects, confidence intervals that are too narrow, and p-values that are too small. This, of course, means that researchers could make a false conclusion of statistical significance.

As discussed in the preceding, various methods are available to calculate ICC. Different software packages may use different method(s) and may find different values for ICC. Therefore, it is important for researchers to report the method used for estimation (Ukoumunne et al. 1999). Another factor for consideration is whether to adjust for covariates. If researchers decide to do this, they may likely get smaller ICCs because some of

the between-cluster variation is explained by cluster-level factors (Feng et al. 1999). To facilitate more accurate estimation of ICC for study design, researchers should continue to report ICC values. This information could help other groups design studies and calculate sample sizes.

Although many principles for RCT design and analysis have been established, there is still lack of consensus about the relative merits of the different methods. More methodological research is required to compare alternative methods and their performance in different settings. Because researchers have a variety of options and tools at their disposal, it is important that they carefully make informed decisions about the analysis and the variables included. These considerations can have an impact on the conclusions of the study. Different procedures generate different parameters, and it is important to know how each procedure performs in different circumstances (Evans et al. 2001).

So far, there is plenty still to be learned through methodological studies. The fruits of this labor have implications in trial design and in analyzing CRT data for individual-level inference. For example, researchers need methods to (1) address adjustment for other covariates when length of subject follow-up varies; (2) analyze ordinal, multinomial, and time-to-event data; (3) deal with missing values at both the individual and the cluster level; and (4) further develop and establish the reliability of Bayesian methods for estimating ICC and CRTs, for which particular attention needs to be paid to prior distribution choice. Without sufficient methodological guidance, researchers must check their model assumptions and the sensitivity of their conclusions carefully before interpreting CRT results. Murray et al. (2004) reviewed the methodological developments regarding the design and analysis of CRTs and concluded that the methods required are not as simple as those for individually randomized trials; still, several methods are readily available for the design and analysis of such trials.

Acknowledgment

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Adams, G., M. C. Gulliford, O.C. Ukoumunne, S. Eldridge, S. Chinn, and M. J. Campbell. 2004. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J. Clin. Epidemiol*., 57(8), 785–94.

Ahmed, M., and M. Shoukri. 2010. A Bayesian estimator of the intracluster correlation coefficient from correlated binary responses. *J. Data Sci*., 8, 127–137.

Althabe, F., P. Buekens, E. Bergel, J. M. Belizán, N. Kropp, L. Wright, et al. 2005. A cluster randomized controlled trial of a behavioral intervention to facilitate the development and implementation of clinical practice guidelines in Latin American maternity hospitals: The Guidelines Trial: Study protocol. *BMC Womens Health*, 5(1), 4.

Altham, P. M. E. 1978. Two generalizations of the binomial distribution. *J. R. Stat. Soc. Ser. C*, 27(2), 162–167.

Bansal, N. K., and M. K. Bhandary. 2000. Bayes estimation of intraclass correlation coefficient. *Commun. Stat.Theory Methods*, 29(1), 79–93.

Bansal, N. K., M. Bhandary, and K. Fujiwara. 2013. Bayes estimation of intraclass correlation coefficients under unequal family sizes. *Commun. Stat. Simul. Comput*., 42(2), 294–302.

Baskerville, N. B., W. Hogg, and J. Lemelin. 2001. The effect of cluster randomization on sample size in prevention research. *J. Family Pract*., 50(3), 242.

Bland, J. M. 2004. Cluster randomised trials in the medical literature: Two bibliometric surveys. *BMC Med. Res. Methodol*., 13(4), 21.

Bryk, A., and S. Raudenbush. 1992. *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Cals, J. W. L., C. C. Butler, R. M. Hopstaken, K. Hood, and G.-J. Dinant. 2009. Effect of point of care testing for C reactive protein and training in communication skills on antibiotic use in lower respiratory tract infections: Cluster randomised trial. *Br. Med J*., 338, b1374.

Campbell, M. J., A. Donner, and N. Klar. 2007. Developments in cluster randomized trials and Statistics in Medicine. *Stat. Med*., 26, 2–19.

Campbell, M. K., J. M. Grimshaw, and D. R. Elbourne. 2004. Intracluster correlation coefficients in cluster randomized trials: Empirical insights into how should they be reported. *BMC Med. Res. Methodol*., 4(9), 702–708.

Carlo, W. A., S. S. Goudar, I. Jehan, E. Chomba, A. Tshefu, A. Garces, et al. 2010. Newborn-care training and perinatal mortality in developing countries. *N. Engl. J. Med*., 362(7), 614–623.

Carrol, R., and D. Ruppert. 1988. *Transformation and weighting in regression*. London, UK: Chapman and Hall.

Chakraborty, H. 2008a. Cluster-randomized trial of a community-based intervention. *Lancet*, 372(9649), 1541.

Chakraborty, H. 2008b. The design and analysis aspects of cluster randomized trials. In *Statistical advances in the biomedical sciences: Clinical trials, epidemiology, survival analysis, and bioinformatics*, ed. A. Biswas, S. Datta, J. Fine, and M. Segal, 67–75. New York, NY: John Wiley and Sons.

Chakraborty, H., J. Moore, W. A. Carlo, T. D. Hartwell, and L. L. Wright. 2009. A simulation based technique to estimate intracluster correlation for a binary variable. *Contemp. Clin. Trials*, 30(1), 71–80.

Chakraborty, H., J. Moore, and T. D. Hartwell. 2009. Intracluster correlation adjustments to maintain power in cluster trials for binary outcomes. *Contemp. Clin. Trials*, 30(5), 473–480.

Chakraborty, H., and P. K. Sen. 2013. Resampling method to estimate intracluster correlation for clustered binary data. *Commun. Stat. Theory Methods*, (forthcoming).

Chuang, J.-H., G. Hripcsak, and D. Heitjan. 2002. Design and analysis of controlled trials in naturally clustered environments. *J. Am. Med. Inform. Assoc*., 9(3), 230–239.

Cook, J., T. Bruckner, and G. S. MacLennan. 2012. Clustering in surgical trials—Database of intracluster correlations. *Trials*, 13(2), 1–8.

Cornfield, J. 1978. Randomization by group: A formal analysis. *Am. J. Epidemiol*., 108, 100–102.

DerSimonian, R., and N. Laird. 1986. Meta-analysis in clinical trials. *Control Clin. Trials*, 7(3), 177–88.

Divine, G. W., J. T. Brown, and L. M. Frazier. 1992. The unit of analysis error in studies about physicians' patient care behavior. *J. Gen. Intern. Med*., 7(6): 623–629.

Donner, A. 1979. The use of correlation and regression in the analysis of family resemblance. *Am. J. Epidemiol*., 1979;110, 335–342.

Donner, A. 1986. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *Int. Stat. Rev*., 54(1), 67–82.

Donner, A. 1998. Some aspects of the design and analysis of cluster randomization trials. *J. R. Stat. Soc. Ser. C (Appl. Stat.)*, 47(1), 95–113.

Donner, A., N. Birkett, and C. Buck. 1981. Randomization by cluster: Sample size requirements and analysis. *Am. J. Epidemiol*., 114(6), 906–914.

Donner, A., K. Brown, and P. Brasher. 1990. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. *Int. J. Epidemiol*., 19, 795–800.

Donner, A., and A. Donald. 1987. Analysis of data arising from a stratified design with the cluster as unit of randomization. *Stat. Med*., 6, 43–52.

Donner, A., and N. Klar. 1994a. Cluster randomisation trials in epidemiology: Theory and application. *J. Stat. Plan. Inference*, 42, 37–56.

Donner, A., and N. Klar. 1994b. Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. *Am. J. Epidemiol*., 140, 279–289.

Donner, A., and N. Klar. 1996. Statistical considerations in the design and analysis of community intervention trials. *J. Clin. Epidemiol*., 49(4), 435–439.

Donner, A., and N. Klar. 2000. *Design and analysis of cluster randomisation trials in health research*. London, UK: Hodder Arnold.

Donner, A., and N. Klar. 2004. Pitfalls of and controversies in cluster randomization trials. *Am. J. Public Health*, 94(3), 416–422.

Donner, A., and J. J. Koval. 1980. The estimation of intraclass correlation in the analysis of family data. *Biometrics*, 36(1), 19–25.

Donner, A., and J. J. Koval. 1983. A note on the accuracy of Fisher's approximation to the large-sample variance of an intraclass correlation. *Commun. Stat. Simul. Comput*., 12, 443–449.

Donner, A., and G. Wells. 2013. A comparison of confidence interval methods for the intraclass correlation coefficient. *Biometrics*, 42(2), 401–412.

Eldridge, S., and S. Kerry. 2012. *A practical guide to cluster randomised trials in health services research*. New York, NY: John Wiley and Sons.

Eldridge, S. M., D. Ashby, G. S. Feder, A. R. Rudnicka, and O. C. Ukoumunne. 2004. Lessons for cluster randomized trials in the twenty-first century: A systematic review of trials in primary care. *Clin. Trials*, 1(1), 80–90.

Ennett, S. T., N. S. Tobler, C. L. Ringwalt, and R. L. Flewelling. 1994. How effective is drug abuse resistance education? A meta-analysis of Project DARE outcome evaluations. *Am. J. Public Health*, 84(9), 1394–401.

Evans, B. A., Z. Feng, and A. V. Peterson. 2001. A comparison of generalized linear mixed model procedures with estimating equations for variance and covariance parameter estimation in longitudinal studies and group randomized trials. *Stat. Med*., 20, 3353–3373.

Feng, Z., P. Diehr, Y. Yasui, B. Evans, S. Beresford, and T. D. Koepsell. 1999. Explaining community-level variance in group randomized trials. *Stat. Med*., 18(5), 539–56.

Fisher, R. 1925. *Statistical methods for research workers*. Edinburgh, UK: Oliver and Boyd.

Fleiss, J. L. 1986. Reliability of measurement. In *The design and analysis of clinical experiments*, 1–32. New York, NY: John Wiley and Sons.

Fleiss, J. L., and J. Cuzick. 1979. The reliability of dichotomous judgments: Unequal numbers of judges per subject. *Appl. Psychol. Meas*., 3(4), 537–542.

Gail, M., D. Byar, T. Pechaceck, and D. Corle. 1992. Aspects of statistical design for the community intervention trial for smoking cessation (COMMIT). *Control Clin. Trials*, 13, 6–21.

Gilbody, S., P. Bower, D. Torgerson, and D. Richards. 2008. Cluster randomized trials produced similar results to individually randomized trials in a meta-analysis of enhanced care for depression. *J. Clin. Epidemiol*., 61(2), 160–168.

Goldstein, H. 1995. *Multilevel statistical models*. New York, NY: Edward Arnold; Halstead Press.

Gulliford, M. C., G. Adams, O. C. Ukoumunne, R. Latinovic, S. Chinn, and M. J. Campbell. 2005. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *J. Clin. Epidemiol*., 58(3), 246–251.

Hade, E. M., D. M. Murray, M. L. Pennell, D. Rhoda, E. D. Paskett, V. L. Champion, et al. 2010. Intraclass correlation estimates for cancer screening outcomes: Estimates and applications in the design of group-randomized cancer screening studies. *J. Natl. Cancer Inst. Monogr*., 2010(40), 97–103.

Halloran, M. E., C. J. Struchiner, and I. M. Longini. 1997. Study designs for evaluating different efficacy and effectiveness aspects of vaccines. *Am. J. Epidemiol*., 146(10), 789–803.

Handlos, L., H. Chakraborty, and P. K. Sen. 2009. Evaluation of cluster randomized trials on maternal and child health research in developing countries. *Trop. Med. Int. Health*, 14(8), 947–956.

Hayes, R. J., N. D. E. Alexander, S. Bennett, and S. N. Cousens. 2000. Design and analysis issues in cluster-randomized trials of interventions against infectious diseases. *Stat. Methods Med. Res*., 9, 95–116.

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

Hedges, L. V., and E. C. Hedberg. 2007. Intraclass correlation values for planning group-randomized trials in education. *Educ. Eval. Policy Anal*., 29(1), 60–87.

Hereren, T., and R. D'Agostino. 1987. Robustness of the two independent samples *t*-test when applied to ordinal scaled data. *Stat. Med*., 6, 79–90.

Hewitt, C. E., D. J. Torgerson, and J. N. Miles. 2008. Individual allocation had an advantage over cluster randomization in statistical efficiency in some circumstances. *J. Clin. Epidemiol*., 61(10), 1004–1008.

Huber, P. 1981. *Robust statistics*. New York, NY: Wiley.

Isaakidis, P., and J. P. A. Ioannidis. 2003. Evaluation of cluster randomized controlled trials in Sub-Saharan Africa. *Am. J. Epidemiol*., 158(9), 921–926.

Karlin, S., E. Cameron, and P. Williams. 1981. Sibling and parent-offspring correlation estimation with variable family size. *Proc. Natl. Acad. Sci. USA*, 78(5), 2664–2668.

Kernan, W. N., C. M. Viscoli, R. W. Makuch, L. M. Brass, and R. I. Horwitz. 1999. Stratified randomization for clinical trials. *J. Clin. Epidemiol*., 52(1), 19–26.

Kerry, S. M., and J. M. Bland. 1998. Trials which randomize practices 1: How should they be analysed? *Family Pract*., 15(1), 80–83.

Killip, S., Z. Mahfoud, and K. Pearce. 2004. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *Ann. Family Med*., 2(3), 204–208.

Kleinman, J. C. 1973. Proportions with extraneous variance: Single and independent samples. *J. Am. Stat. Assoc*., 68(341), 46–54.

Kupper, L. L., and J. K. Haseman. 1978. The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics*, 34(1), 69–76.

Lee, E., and N. Durbin. 1994. Estimation and sample size considerations for clustered binary responses. *Stat. Med*., 13, 1241–1252.

Liang, K. Y., and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.

Lipsitz, S. R., N. M. Laird, and T. A. Brennan. 1994. Simple moment estimates of the k-coefficient and its variance. *J. R. Stat. Soc. Ser. C*, 43(2), 309–323.

MacLennan, G., C. Ramsay, J. Mollison, K. Campbell, J. Grimshaw, and R. Thomas. 2003. *Room for improvement in the reporting of cluster randomised trials in behaviour change research*. New York, NY: Elsevier Science.

Mak, T. K. 1988. Analysing Intraclass Correlation for Dichotomous Variables. *J. R. Stat. Soc. Ser. C*, 37(3), 344–352.

Maritz, T., and R. Jarrett. 1983. The use of statistics to examine the assiciation between fluoride in drinking water and cancer death rates. *Appl. Stat*., 32(2), 97–101.

McCullagh, P., and J. A. Nelder. 1989. *Generalized linear models*, 2nd ed. London, UK: Chapman and Hall.

McKinlay, S. M., E. J. Stone, and D. M. Zucker. 1989. Research design and analysis issues. *Health Educ. Behav*., 16(2), 307–313.

Murray, D. M., S. P. Varnell, and J. L. Blitstein. 2004. Design and analysis of group-randomized trials: A review of recent methodological developments. *Am. J. Public Health*, 94(3), 423–32.

Nelder, J. A. 1987. Pregibon D. An extended quasi-likelihood function. *Biometrika*, 74(2), 221–232.

Ochi, Y., and R. L. Prentice. 1984. Likelihood inference in a correlated probit regression model. *Biometrika*, 71(3), 531–543.

Online intraclass correlation database. n.d. http://stateva.ci.northwestern.edu

Prentice, R. L. 1986. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J. Am. Stat. Assoc*., 81(394), 321–327.

Puffer, S., D. J. Torgerson, and J. Watson. 2003. Evidence for risk of bias in cluster randomised trials: Review of recent trials published in three general medical journals. *Br. Med. J*., 327, 1–5.

Raudenbush, S. W. 1997. Statistical analysis and optimal design for cluster randomized trials. *Psychol. Methods*, 2(2), 173–185.

Reading, R., I. Harvey, and M. Mclean. 2000. Cluster randomised trials in maternal and child health: Implications for power and sample size. *Arch. Dis. Child*., 82(1), 79–83.

Ridout, M. S., C. G. B. Demetrio, and D. Firth. 1999. Estimating intraclass correlation for binary data. *Biometrics*, 55(March), 137–148.

Rogers W. 1993. Regression standard errors in clustered samples. *Stata Tech. Bull*., 13, 19–23.

Schulz, K. F. 1995. Subverting randomization in controlled trials. *J. Am. Med. Assoc*., 274(18), 1456–1458.

Simpson, J. M., N. Klar, and A. Donner. 1995. Accounting for cluster randomization: A review of primary prevention trials, 1990 through 1993. *Am. J. Public Health*, 85(10), 1378–1383.

Smith, C. 1956. On the estimation of intraclass correlation. *Ann. Hum. Genet*., 21, 363–373.

Smith, C. 1980a. Estimating genetic correlations. *Ann. Hum. Genet*., 44, 265–284.

Smith, C. 1980b. Further Remarks on estimating genetic correlations. *Ann. Hum. Genet*., 44, 95–105.

Spiegelhalter, D. J. 2001. Bayesian methods for cluster randomized trials with continuous responses. *Stat. Med*., 20(3), 435–52.

Srivastava, M. S. 1984. Estimation of interclass correlations in familial data. *Biometrika*, 71(1), 177.

Stiratelli, R., N. Laird, and J. H. Ware. 1984. Random-effects models for serial observations with binary response. *Biometrics*, 40(4), 961–971.

Swiger, L. A., W. R. Harvey, D. O. Everson, and K. E. Gregory. 1964. The variance of intraclass correlation involving groups with one observation. *Biometrics*, 20(4), 818–826.

Tamura, R. N., and S. S. Young. 1987. A stabilized moment estimator for the beta-binomial distribution. *Biometrics*, 43(4), 813–824.

Thabane, L., L. Mbuagbaw, S. Zhang, Z. Samaan, M. Marcucci, C. Ye, et al. 2013. A tutorial on sensitivity analyses in clinical trials: The what, why, when and how. *BMC Med. Res. Methodol*., 13(1), 92.

Thomas, J., and R. Hultquist. 1978. Interval estimation for the unbalanced case of the one-way random effects model. *Ann. Stat*., 6, 582–587.

Thompson, S. G., S. D. Pyke, and R. J. Hardy. 1997. The design and analysis of paired cluster randomized trials: An application of meta-analysis techniques. *Stat. Med*., 16(18), 2063–2079.

Turner, R. M., R. Z. Omar, and S. G. Thompson. 2001. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Stat. Med*., 20(3), 453–72.

Turner, R., A. Prevost, and S. G. Thompson. 2004. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Stat. Med*., 23, 1195–1214.

Ukoumunne, O. C., M. C. Gulliford, S. Chinn, J. Sterne, and P. Burney. 1999. Methods for evaluating area-wide and organization-based interventions in health and health care: A systematic review. *Health Technol. Assess*., 3(5), iii–92.

Ware, J. H. 1985. Linear models for the analysis of longitudinal studies. *Am. Stat*., 39(2), 95–101.

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817–838.

Whiting-O'Keefe, Q. E., C. Henke, and D. W. Simborg. 1984. Choosing the correct unit of analysis in medical care experiments. *Med. Care*. 22(12), 1101–1114.

Williams, D. A. 1982. Extra-binomial variation in logistic linear models. *J. R. Stat. Soc. Ser. C*, 31(2), 144–148.

Wolfinger, R., and M. O'Connell. 1993. Generalized linear mixed models: A pseudo-likelihood approach. *J. Stat. Comput. Simul*., 48, 233–243.

Wu, S., C. M. Crespi, and W. K. Wong. 2012. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemp. Clin. Trials*, 33(5), 869–880.

Yamamoto, E., and T. Yanagimoto. 1992. Moment estimators for the binomial distribution. *J. Appl. Stat*., 19, 273–283.

Zeger, S., and K. Liang. 1992. An overview of methods for the analysis of longitudinal data. *Stat. Med*., 11, 825–839.