Cluster Randomized Trials for Health Care Quality Improvement Research

Michelle M. Garrison, PhD; Rita Mangione-Smith, MD, MPH

From the Seattle Children's Research Institute, Center for Child Health, Behavior and Development, Seattle, Wash (Drs Garrison and Mangione-Smith); Department of Health Services, University of Washington, Seattle, Wash (Drs Garrison and Mangione-Smith); and Department of Psychiatry and Behavioral Sciences, Division of Child and Adolescent Psychiatry (Dr Garrison), and Department of Pediatrics, University of Washington, Seattle, Wash (Dr Mangione-Smith)
The views expressed in this report are those of the authors and do not necessarily represent those of the US Department of Health and Human Services, the Agency for Healthcare Research and Quality or the American Board of Pediatrics Foundation.
The authors declare that they have no conflict of interest.
Publication of this article was supported by the Agency for Healthcare Research and Quality and the American Board of Pediatrics Foundation.
Address correspondence to Michelle M. Garrison, PhD, Seattle Children's Research Institute, PO Box 5371, Seattle, WA 98145-5005 (e-mail: michelle.garrison@seattlechildrens.org).

Received for publication March 6, 2013; accepted July 29, 2013.
ACADEMIC PEDIATRICS 2013;13:S31–S37

ALTHOUGH RANDOMIZED CONTROLLED trials have long been the gold standard of clinical research, the needs of quality improvement (QI) research are somewhat different. Here, the randomization of individual patients can create a range of problems—practical, political, statistical—and so increasingly we see the cluster randomized trial (CRT) design utilized instead.[1] In a CRT, groups or clusters are randomized instead of individual patients; the unit of randomization might be provider panels,[2–4] clinics,[5–12] inpatient units,[13] emergency departments,[14] hospitals,[15,16] or even blocks of time.[17] Outcomes are often measured at both the patient and group level, and analyses take the clustering into account. Although CRTs can be the most effective way to evaluate QI interventions in some cases, they present special challenges in design and analysis that require careful planning.[18]

CONSIDERATIONS FOR UTILIZING CRTS

QI interventions often have targets above the patient level. If an intervention is delivered at higher levels such as provider education or modified order sets, it may not be feasible to implement the intervention as randomized at the patient level. In addition to implementation difficulties, cross-contamination can often lead to issues with fidelity in the control group.

EXAMPLE

After developing a decision support tool to increase assessment of asthma severity at each visit, a clinic decides to conduct a randomized trial of the tool. Because patients have been randomized, each participating provider has both intervention and control patients. After using the decision support tool with the intervention patients for 2 weeks, the experience begins to affect how the provider handles asthma visits for control patients as well, even without the decision support tool present.

CRT ADVANTAGE

Instead, the study could randomize provider panels. For any given provider, all of their patients will either be in the intervention group or the control group. As a result, providers do not need to remember which of their patients is or is not in the intervention group and can use the same processes of care across encounters.

Cross-contamination can be an issue within provider panels, or across providers within the same setting of care. The degree to which cross-contamination is likely to be an issue can depend on the intervention itself as well as on cultural aspects of the practice setting. Another significant advantage is the ability to measure group-level outcomes, such as costs or provider satisfaction. Although it is possible to assess group-level outcomes in studies that randomize individual patients, there can be considerable misclassification in the results.

EXAMPLE

A children's hospital has developed a standardized work process for ensuring that patient families have all the information and resources they need before discharge. Patients were randomized to receive either the new (intervention) or old (control) discharge process, and patient-level outcomes (having all prescribed discharge medications, understanding discharge instructions, and parent satisfaction) were indeed better in the group receiving the new standardized process. However, there were concerns among the floor staff that the additional staff time required at discharge might not be sustainable. Because staff time was split across intervention and control patients and some tasks were batched across multiple patients, it was difficult to tease out the additional time demands of the new process.

CRT ADVANTAGE

Instead of randomizing patients, the study could randomize medical and surgical units of the hospital, making it easier to track and compare staff time spent on all discharge-related tasks.

CRTs also have the advantage of more closely approximating real-world implementation of an intervention, which means that the results may be more generalizable and useful in planning widespread rollout.

EXAMPLE

A large urban general pediatrics clinic decides to test using phone calls from a nurse educator to families of children not receiving indicated well-child care visits; the phone calls will be used to address parent outcome expectations about the visits and problem solve around barriers to well-child visits. Because they are not sure whether the change will work well for the clinic as a whole, they decide to randomly assign 10% of eligible families to receive the intervention. Although families respond positively to the intervention, the nurse manager raises the concern that although the nurse educators were able to incorporate calling 10% of eligible families into their workload, it was not clear how feasible that would be when the change was scaled up.

CRT ADVANTAGE

Instead of randomizing patients, the study could randomize provider–nurse clusters. Because each nurse would be either delivering the intervention or not to their entire usual patient panel, feasibility issues would be more transparent. The clinic could even choose to survey nurses in both intervention and control groups to assess perceived time pressure, unachieved tasks, and job satisfaction.

Last but not least, CRTs can simply be more feasible to conduct, both practically and politically. In addition to the complications in trying to roll out a QI intervention to only a randomized subset of patients within a panel or clinic, political concerns can arise when patients or providers perceive inequities in care.

EXAMPLE

The school-based health clinics in a large school district have developed a case management program for adolescents with depression. The case manager provides brief, problem-focused counseling and talks to the patient weekly via phone to discuss medication compliance, side effects, and changes in symptoms; updates are sent to the prescribing provider, who is also contacted if a change in medication or dose appears indicated. Newly diagnosed teens have been randomized to receive the case management program or usual care. However, providers are concerned about the lack of equity in care, and there have been complaints from the adolescents assigned to the usual-care arm who have talked to peers receiving case management.

CRT ADVANTAGE

Instead of randomizing patients, the study could randomize school-based health clinics if there are a sufficient number. Because treatment would be the same for all newly diagnosed students seeking services within each clinic, conversations and comparisons between intervention and control adolescents are less likely; even if disparities between schools are noticed, institution-level program differences can be perceived more favorably by patients, families, and providers than inequalities in care within a clinic.

So although patient-level randomization clearly makes sense for studies such as pharmaceutical trials, cluster randomization may be a better model in many cases for evaluating the effectiveness of QI interventions.

The primary disadvantage of a CRT design is the decreased statistical power. A CRT will almost always require more subjects than a trial randomized at the patient level in order to have comparable power to detect the same difference in effect. In the case of QI research, however, this disadvantage is generally more than made up for by the many advantages.

SCENARIOS OF CRT UTILITY

Although CRTs can provide a rigorous method to evaluate QI interventions, they are a better choice in some situations than others. A CRT design is most often going to be a good fit in situations where either the intervention is delivered at a level above the patient or there are salient outcomes to be measured at the group level.

EXAMPLE 1

The intervention provides training for providers in motivational interviewing skills, aimed at increasing protocol adherence in adolescents with diabetes. The study will enroll adolescents with diabetes enrolled in a staff-model HMO who see any of the study providers for their diabetes care and will measure self-reported compliance, A1c values, and the amount of provider time spent. In this case, intervention delivery would be at the provider level, and outcomes are at both the patient and provider levels. Simply randomizing patients would mean that participating providers have both intervention and control patients, and cross-contamination would occur. Depending on continuity of care and the degree to which providers collaborate, randomization may need to be at the clinic level rather than at the provider level. However, randomizing at the provider level would preserve more statistical power.

EXAMPLE 2

The intervention coaches pediatric primary care providers on communicating with families who are refusing immunizations for their children and gives providers family education resources to use. The study will enroll families who report on screening before well-child visits that they are planning or considering declination of the standard immunization schedule and who receive care at a pediatric provider within the research network. Outcomes measured will include the proportion of eligible vaccines received by the child and provider satisfaction. In this case, the intervention is delivered at the provider level, and outcomes are measured at both the provider and patient levels. Randomization will occur

at the clinic level, given the significant potential for cross-contamination across providers within a given clinic and the concern that families may seek out and switch to a control group provider within the clinic if they plan to decline immunizations.

EXAMPLE 3

The intervention aims to improve the sleep environment for both children and parents in the inpatient setting because adequate sleep is critical for healing, pain management, and ability to engage in treatment and discussions with providers. The study will enroll families with a child admitted to any of the hospital's medical or surgical units and will use actigraphy and sleep diaries to track child and parent sleep in addition to clinical outcomes. Although intervention delivery will be at the level of the family, many of the components of the intervention will be more feasible to implement consistently if they are used across an entire unit because the intervention involves changing the settings on clinical monitoring equipment and changes in nursing protocols. Randomizing at the unit level would be the most feasible strategy with the least cross-contamination.

On the other hand, there are some red flags to watch out for that can suggest that a CRT design may not be the right fit for a situation. If there are fewer than 6 to 8 groups or clusters (whether these are provider panels, clinics, or hospitals) that can be randomized, it is unlikely that the study can be sufficiently powered as a CRT; depending on the intervention and outcomes involved, even more groups may be required. If there is significant diversity across the groups (eg, if the study will include both medical and surgical units within a hospital), the total number of groups required will be even greater so as to ensure a balance across study arms. If there is frequent crossover in patients, providers, residents, or nursing staff between groups, cross-contamination would also be a concern. In such a situation, the likelihood is high that at least some patients in control groups would also receive clinical care that was influenced by the intervention. As a result, the difference in outcomes between intervention and control group would be diluted, and the study may not find an effect even if the underlying intervention was successful.

EXAMPLE 4

The intervention involves a follow-up phone call from a nurse several days after discharge from an asthma emergency department (ED) or urgent-care visit to encourage outpatient follow-up and medication adherence. The study will enroll children discharged from 5 different area emergency departments and urgent-care centers. The primary outcome will be returns to ED or urgent care within 3 months. There are several reasons why a CRT design would not be a good fit for this study. Cluster randomization would significantly dilute power without any real gains, especially because both intervention delivery and outcome are at the patient level, and there are relatively few concerns about cross-contamination across patients. The total number of groups available to randomize would not be sufficient, especially given that the effects of the intervention may be different for those being discharged from an ED versus an urgent-care clinic. A better design would be patient-level randomization, stratified by site.

EXAMPLE 5

The intervention involves the implementation of a bronchiolitis care pathway among children under 2 years hospitalized with bronchiolitis in any of the 4 medical units at a single academic hospital. Outcomes measured would include provider compliance with pathway protocol, patient length of stay, 30-day ED or inpatient returns, and parent satisfaction with care. There are several reasons why a CRT design would not be a good fit for this study. There would not be sufficient power to randomize at the unit level, and although cluster randomization by attending provider is technically an option, successful attempts to implement a care pathway in the intervention group would almost certainly have significant cross-contamination to attending providers in the control group, especially because this is an academic setting. A stepped wedge design might be the most straightforward way to answer the question about the care pathway.

DESIGNING A CRT

Assuming that the QI intervention has already been developed, there are a number of steps to be followed in designing the CRT. The CONSORT-CRT criteria can be useful here; not only do the standards lay out the information about a CRT that should be reported at publication, but they can also be used as a guide during study design to ensure that critical issues are addressed intentionally and in advance.[19]

Some of the most critical decisions revolve around sample selection and randomization. Once the target population for the intervention has been identified, the unit of randomization can be ascertained. This decision often hinges on several factors, including implementation feasibility and the need to minimize cross-contamination. Although some CRTs will randomize provider panels, others may have 3 or more levels, such as patients nested within providers nested within clinics.[6,20] In some cases, the pool of potential clusters may already be defined: medical units within a hospital or primary care clinics within a staff-model HMO. In other cases, however, clusters need to be actively recruited for participation. Networking within organizations such as Children's Hospital Association and the Pediatric Research in Office Settings national practice-based research network has helped some QI researchers identify potential clusters. Once some interested providers or care settings have been identified, they may provide connections to additional potential clusters, in a snowball effect.

Also important to take into account at this stage are cluster-level covariates that may be significant confounders or effect modifiers. For example, a QI intervention aimed at decreasing unnecessary ED utilization might have

6 participating EDs with relatively low-income patient populations and 4 participating EDs with significantly higher median incomes. Although randomization of larger sample sizes generally does a good job of balancing covariates across study arms, CRTs are often randomizing a small enough number of clusters that steps may need to be taken to balance allocation.[21,22]

One option for balancing allocation of groups across study arms is matching, where each group is part of a pair that is matched on the covariates of concern. During randomization, the first group of the pair is randomly allocated and the remaining group of the pair is automatically assigned to the opposite arm. For example, a CRT that had recruited 8 primary care providers might have a pair of family practice doctors, a pair of nurse practitioners, and 2 pairs of pediatricians. In this case, matching would ensure that each arm had an equal balance of provider types. In some situations, studies will recruit clusters and then decide which pairs to match with each other; in others, they will conduct a targeted recruitment for a cluster that will be a good pair match for a currently enrolled cluster. However, there are some downsides of using matched pairs. Because the analytic approach to a matched pair study directly compares each group to their matched pair, having 1 group drop out of the study results in their matched pair dropping out of the analysis as well. Matching can also make rolling allocation difficult or impossible, which means that all clusters may need to be recruited and enrolled before matching and random assignment can occur. When the covariate of concern is a continuous variable, the matching may be relatively rough—for example, a pair matched on percentage of Medicaid may have 70% vs 40% if the other participating groups all have lower values.

If a matched approach is used unnecessarily, the effective analytic power of the study can be reduced. An option for balancing allocation that falls between matched and unrestricted approaches is a stratified randomization.[23] Like matching, stratification groups clusters by covariate values; instead of using pairs, however, clusters are part of a larger group. For example, a study of a QI intervention targeting a resident continuity clinic might stratify residents by year, then use block randomization within those strata. This would ensure that each study arm had a balance of patient panels from R1, R2, and R3 residents. With stratified randomization, having 1 cluster drop out of the study does not result in the loss of the whole strata from the analysis as long as at least 1 cluster from each study arm remains in the strata.

Outcome Measures

The selection of appropriate outcomes for measurement and analysis is critical in all studies, and CRTs are no exception. As with other types of clinical trials, collecting baseline measures of outcome variables can help increase the precision of effect estimates and preserve valuable statistical power. Because it is not uncommon for QI efforts (as with other interventions) to result in a transient benefit before the outcome regresses to the mean, collecting long-term follow-up data can be incredibly valuable. However, there are unique factors to consider as well. A CRT may collect and analyze outcomes at multiple levels, such as patient-level functional status, provider-level time spent, or clinic-level costs.[24,25] Additionally, it may not be necessary to collect outcome data on all patients within a cluster. Depending on the outcome measure and the population, the marginal value for statistical power of each additional patient within a cluster can sometimes mean that collecting data on a randomly selected sample from each cluster is the most efficient method.

Statistical Issues With CRTs: Sample Size and Analysis

Determining how many clusters are needed, how many subjects are needed per cluster, and what effect size will be detectable for the outcomes are questions that all come down to power (sample size) calculations. Unlike the power calculations for a simple randomized clinical trial, the power calculations for CRTs tend to start with complex and move up from there. As a result, the involvement of an experienced biostatistician at this stage can be crucial to ensuring that a study will be capable of truly evaluating the QI intervention. A project team that waits to contact a biostatistician until after their CRT has finished collecting data may unfortunately learn that they have too few clusters to allow for any useful analysis of the results.

In a standard clinical trial, patients are randomized, and the number of patients enrolled is the critical factor in determining the statistical power to detect an effect. A CRT, on the other hand, will need a larger sample size than would a clinical trial randomizing patients to examine the same intervention, population, and outcomes—and the number of clusters can often be more influential in power calculations than the number of subjects per cluster. The more similarity between subjects within a cluster compared to subjects in other clusters for the outcome, the greater the intraclass correlation coefficient (ICC) and the larger the sample size needed. Failing to adequately take this into account can result in a study that has grossly underestimated the needed sample size.[26] Researchers who plan ahead, however, can often make study design choices that can make the post of power in the CRT. For example, it is possible that a study could require 100 subjects per cluster if there are 10 clusters, but only 50 subjects per cluster if an additional 4 clusters are recruited for 14 total clusters—a difference of 1000 total subjects versus 700. Having a similar number of subjects within each cluster can also help maximize power; on the other hand, a study with cluster sizes that range from 20 to 100 will have less statistical power than a study with the same number of total subjects but with a consistent cluster size of 60.[27] Another strategy sometimes used is to select clusters that have similar clinical and demographic characteristics because it can result in a decreased ICC, as patients may be just as similar across clusters as they are within clusters.

Additionally, some outcomes will have a greater ICC than others, even within the same populations; as a result, outcomes that are less dependent on cluster characteristics may have greater statistical power.[28,29]

The first step a biostatistician will need to take in power calculations for a CRT is to estimate the ICC for the chosen outcome measures.[30] An ICC of 0 would mean that there are no differences in how the outcome is distributed within versus between clusters, and the power calculations would be exactly as they would be for a patient-randomized trial. As the ICC rises, the number of clusters becomes increasingly more important than the number of subjects per cluster. Ideally, data are available to directly measure the ICC in a similar population using ANOVA; previous research, administrative data, and national survey data sets are often useful for this.[31] In reality, it is often necessary to use simulation models to estimate a range of possible ICCs or to simply perform sample size calculations with a range of realistic ICCs.[32] Once calculated or estimated, the ICC can be used to compute the sample size needed for a study—or to compute the effect size that can be detected from a fixed sample size. Although the simplest method for doing this involves using the ICC to calculate the variance inflation factor and adjusting potential sample sizes accordingly, more complex study designs can require more sophisticated methods, especially for 3-level CRTs (such as patients nested within providers nested within clinics).[33,34]

At the analysis stage, again, the assistance of an experienced biostatistician is indicated. Just as with the power calculations, the analysis also needs to take cluster levels into account.[35] An analysis that fails to do this correctly may overestimate the significance of the intervention effect quite considerably.[26] There are a wide range of analytic options available to achieve this, depending on factors such as the outcome type, number of measurement time points (both before and after the intervention), levels of clusters, whether matching or stratified randomization was used, and the need to control for potential confounders not equally distributed across study arms.

LOGISTICAL AND ETHICAL ISSUES IN CRTS

Given that statistical power in CRTs can be so dependent on the number of clusters, it may not be surprising that one of the greatest challenges in a CRT can be the loss of a cluster before outcome data are collected. The loss of a cluster generally has a more detrimental impact on statistical power than would the loss of the same number of subjects spread out across clusters, and it can sometimes be enough to change an otherwise meaningful study into one that simply does not have adequate power to answer the study questions. As a result, it is important to plan for cluster loss in the initial study design. If a matched pair design is being used to randomize clusters, it is important to realize that the loss of 1 cluster also drops the entire matched pair from the analysis. In some situations, outcome data on the patients in a cluster may still be available even if a cluster may choose to drop out of the implementation of a QI intervention, in which case an intent-to-treat analysis can be performed.[36]

Even if all clusters remain in the study until the end, there may be considerable variability across or within clusters in how the QI intervention is delivered and adhered to.[37,38] After all, the providers involved may not be the ones who volunteered the cluster—and yet they are most often in QI CRTs as either the targets of the intervention and/or the ones tasked with delivering the intervention to the patients. If engagement of all providers in a cluster is necessary, it can be well worth the effort to recruit and obtain consent for each as if they were signing up for a study individually. There may also be provider characteristics, differences in patient populations, or social and physical environmental factors of the practice setting that can influence the fidelity of the QI intervention implementation. Where possible, these factors ought to be measured so they may be taken into account in analyzing and interpreting the results.[39,40]

Last, there are several ethical issues to consider in conducting a CRT.[41–44] Although a clinic manager may be eager in some cases to sign up the entire clinic for participating in the CRT, there can be other considerations.[45–47] If providers are the target of the QI intervention and outcome measurement, obtaining informed consent from each participating provider may be necessary for ethical and institutional purposes, above and beyond the positive effect that doing so may have on engagement.[48] If patient-level information will be obtained, both Health Insurance Portability and Accountability Act and human subjects regulations can come into play, and require consent of patients or parents.[49] Because many CRTs require access to patient-level information for recruitment purposes before consent, it is important to work carefully with the institutional review board in advance.[50,51] If the intervention is being delivered at the provider or clinic level, informed consent at the patient level may be solely for data collection, rather than consenting to randomization and the intervention itself.[52]

Although control groups and blinding can create ethical challenges in any clinical trial, this can be especially so in CRTs of QI efforts, where the study is often changing the basic process of receiving health care. Especially if the QI effort is in response to a problem with patient safety or inadequate care, there may be resistance from providers and administrators to having a control group at all, despite the clinical equipoise regarding the effectiveness of the proposed QI intervention to address the issue. One solution to this is the use of a stepped wedge CRT design, in which all clusters receive the QI intervention and only the timing of the implantation is randomized.[53] If clusters are randomized before recruitment of providers or patients, this may result in biased sampling, as clusters randomized to the control group may find it more difficult to recruit than clusters randomized to the intervention.[26,54,55]

CONCLUSION

CRTs can be an efficient and powerful way to demonstrate the effectiveness of QI interventions in pediatric settings. However, the design and execution of a CRT

can involve considerable challenges; interdisciplinary teamwork is critical to the successful planning, implementation, and analysis of a CRT.

REFERENCES

1. Ukoumunne OC, Gulliford MC, Chinn S, et al. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technol Assess*. 1999;3: iii-92.
2. van Agt HME, van der Stege HA, de Ridder-Sluiter H, et al. A cluster-randomized trial of screening for language delay in toddlers: effects on school performance and language development at age 8. *Pediatrics*. 2007;120:1317–1325.
3. Brown R, Bratton SL, Cabana MD, et al. Physician asthma education program improves outcomes for children of low-income families. *Chest*. 2004;126:369–374.
4. Davis RL, Wright J, Chalmers F, et al. A cluster randomized clinical trial to improve prescribing patterns in ambulatory pediatrics. *PLoS Clin Trials*. 2007;2:e25.
5. Bell LM, Grundmeier R, Localio R, et al. Electronic health record–based decision support to improve asthma care: a cluster-randomized trial. *Pediatrics*. 2010;125:e770–e777.
6. Epstein JN, Langberg JM, Lichtenstein PK, et al. Use of an Internet portal to improve community-based pediatric ADHD care: a cluster randomized trial. *Pediatrics*. 2011;128:e1201–e1208.
7. Epstein JN, Rabiner D, Johnson DE, et al. Improving attention-deficit/hyperactivity disorder treatment outcomes through use of a collaborative consultation treatment service by community-based pediatricians: a cluster randomized trial. *Arch Pediatr Adolesc Med*. 2007; 161:835–840.
8. Hiscock H, Bayer JK, Price A, et al. Universal parenting programme to prevent early childhood behavioural problems: cluster randomised trial. *BMJ*. 2008;336:318–321.
9. Hiscock H, Bayer JK, Hampton A, et al. Long-term mother and child mental health effects of a population-based infant sleep intervention: cluster-randomized, controlled trial. *Pediatrics*. 2008;122:e621–e627.
10. Stout JW, Smith K, Zhou C, et al. Learning from a distance: effectiveness of online spirometry training in improving asthma care. *Acad Pediatr*. 2012;12:88–95.
11. Taveras EM, Gortmaker SL, Hohman KH, et al. Randomized controlled trial to improve primary care to prevent and manage childhood obesity: the High Five for Kids study. *Arch Pediatr Adolesc Med*. 2011;165:714–722.
12. Weiss K, Lozano P, Finkelstein J, et al. A randomized controlled clinical trial to improve asthma care for children through provider education and health systems change: a description of the Pediatric Asthma Care Patient Outcome Research Team (PAC-PORT II) study design. *Health Serv Outcomes Res Methodol*. 2003;4:265–282.
13. Tamblyn R, Huang AR, Meguerditchian AN, et al. Using novel Canadian resources to improve medication reconciliation at discharge: study protocol for a randomized controlled trial. *Trials*. 2012;13:150.
14. Stiell IG, Clement CM, Grimshaw JM, et al. A prospective cluster-randomized trial to implement the Canadian CT head rule in emergency departments. *CMAJ*. 2010;182:1527–1532.
15. Platt R, Takvorian SU, Septimus E, et al. Cluster randomized trials in comparative effectiveness research: randomizing hospitals to test methods for prevention of healthcare-associated infections. *Med Care*. 2010;48(6 suppl):S52–S57.
16. Walsh M, Laptook A, Kazzi SN, et al., National Institute of Child Health and Human Development Neonatal Research Network. A cluster-randomized trial of benchmarking and multimodal quality improvement to improve rates of survival free of bronchopulmonary dysplasia for infants with birth weights of less than 1250 grams. *Pediatrics*. 2007;119:876–890.
17. Cunningham S, Logan C, Lockerbie L, et al. Effect of an integrated care pathway on acute asthma/wheeze in children attending hospital: cluster randomized trial. *J Pediatr*. 2008;152:315–320.
18. Eldridge SM, Ashby D, Feder GS, et al. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials*. 2004;1:80–90.
19. Walleser S, Hill SR, Bero LA. Characteristics and quality of reporting of cluster randomized trials in children: reporting needs improvement. *J Clin Epidemiol*. 2011;64:1331–1340.
20. Lozano P, Finkelstein JA, Carey VJ, et al. A multisite randomized trial of the effects of physician education and organizational change in chronic-asthma care: health outcomes of the Pediatric Asthma Care Patient Outcomes Research Team II Study. *Arch Pediatr Adolesc Med*. 2004;158:875–883.
21. Glynn RJ, Brookhart MA, Stedman M, et al. Design of cluster-randomized trials of quality improvement interventions aimed at medical care providers. *Med Care*. 2007;45(10 suppl 2):S38–S43.
22. Ivers NM, Halperin IJ, Barnsley J, et al. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials*. 2012;13:120.
23. Lewsey JD. Comparing completely and stratified randomized designs in cluster randomized trials when the stratifying factor is cluster size: a simulation study. *Stat Med*. 2004;23:897–905.
24. Finkelstein JA, Lozano P, Fuhlbrigge AL, et al., Pediatric Asthma Care Patient Outcomes Research Team. Practice-level effects of interventions to improve asthma care in primary care settings: the Pediatric Asthma Care Patient Outcomes Research Team. *Health Serv Res*. 2005;40(6 pt 1):1737–1757.
25. Sullivan SD, Lee TA, Blough DK, et al. A multisite randomized trial of the effects of physician education and organizational change in chronic asthma care: cost-effectiveness analysis of the Pediatric Asthma Care Patient Outcomes Research Team II (PAC-PORT II). *Arch Pediatr Adolesc Med*. 2005;159:428–434.
26. Eldridge S, Ashby D, Bennett C, et al. Internal and external validity of cluster randomised trials: systematic review of recent trials. *BMJ*. 2008;336:876–880.
27. You Z, Williams OD, Aban I, et al. Relative efficiency and sample size for cluster randomized trials with variable cluster sizes. *Clin Trials*. 2011;8:27–36.
28. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clin Trials*. 2005;2:99–107.
29. Diehr P, Chen L, Patrick D, et al. Reliability, effect size, and responsiveness of health status measures in the design of randomized and cluster-randomized trials. *Contemp Clin Trials*. 2005;26:45–58.
30. Campbell MJ. Cluster randomized trials in general (family) practice research. *Stat Methods Med Res*. 2000;9:81–94.
31. Turner RM, Thompson SG, Spiegelhalter DJ. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clin Trials*. 2005;2:108–118.
32. van Breukelen GJP, Candel MJJM. Calculating sample sizes for cluster randomized trials: we can keep it simple and efficient! *J Clin Epidemiol*. 2012;65:1212–1218.
33. Heo M, Leon AC. Statistical power and sample size requirements for three level hierarchical cluster randomized trials. *Biometrics*. 2008; 64:1256–1262.
34. Teerenstra S, Moerbeek M, van Achterberg T, et al. Sample size calculations for 3-level cluster randomized trials. *Clin Trials*. 2008; 5:486–495.
35. Campbell MK, Mollison J, Steen N, et al. Analysis of cluster randomized trials in primary care: a practical approach. *Fam Pract*. 2000;17: 192–196.
36. Jo B, Asparouhov T, Muthén BO. Intention-to-treat analysis in cluster randomized trials with noncompliance. *Stat Med*. 2008;27: 5565–5577.
37. Jo B, Asparouhov T, Muthén BO, et al. Cluster randomized trials with treatment noncompliance. *Psychol Methods*. 2008;13:1–18.
38. Tjia J, Mazor KM, Field T, et al. Predicting nursing home adherence to a clinical trial intervention: lessons for the conduct of cluster randomized trials. *J Am Geriatr Soc*. 2011;59:2332–2336.
39. Moerbeek M. Power and money in cluster randomized trials: when is it worth measuring a covariate? *Stat Med*. 2006;25:2607–2617.

40. Stephens AJ, Tchetgen Tchetgen EJ, De Gruttola V. Augmented generalized estimating equations for improving efficiency and validity of estimation in cluster randomized trials by leveraging cluster-level and individual-level covariates. *Stat Med*. 2012;31: 915–930.
41. McRae AD, Bennett C, Brown JB, et al. Researchers' perceptions of ethical challenges in cluster randomized trials: a qualitative analysis. *Trials*. 2013;14:1.
42. Taljaard M, Weijer C, Grimshaw JM, et al. Ethical and policy issues in cluster randomized trials: rationale and design of a mixed methods research study. *Trials*. 2009;10:61.
43. Weijer C, Grimshaw JM, Eccles MP, et al., Ottawa Ethics of Cluster Randomized Trials Consensus Group. The Ottawa statement on the ethical design and conduct of cluster randomized trials. *PLoS Med*. 2012;9:e1001346.
44. Weijer C, Grimshaw JM, Taljaard M, et al. Ethical issues posed by cluster randomized trials in health research. *Trials*. 2011;12:100.
45. Gallo A, Weijer C, White A, et al. What is the role and authority of gatekeepers in cluster randomized trials in health research? *Trials*. 2012;13:116.
46. Mazor KM, Sabin JE, Goff SL, et al. Cluster randomized trials to study the comparative effectiveness of therapeutics: stakeholders' concerns and recommendations. *Pharmacoepidemiol Drug Saf*. 2009;18:554–561.
47. Mazor KM, Sabin JE, Boudreau D, et al. Cluster randomized trials: opportunities and barriers identified by leaders of eight health plans. *Med Care*. 2007;45(10 suppl 2):S29–S37.
48. McRae AD, Weijer C, Binik A, et al. Who is the research subject in cluster randomized trials in health research? *Trials*. 2011;12:183.
49. Eldridge SM, Ashby D, Feder GS. Informed patient consent to participation in cluster randomized trials: an empirical exploration of trials in primary care. *Clin Trials*. 2005;2:91–98.
50. McRae AD, Weijer C, Binik A, et al. When is informed consent required in cluster randomized trials in health research? *Trials*. 2011;12:202.
51. Sim J, Dawson A. Informed consent and cluster-randomized trials. *Am J Public Health*. 2012;102:480–485.
52. Giraudeau B, Caille A, Le Gouge A, et al. Participant informed consent in cluster randomized trials: review. *PLoS One*. 2012;7:e40436.
53. Mdege ND, Man MS, Taylor Nee Brown CA, et al. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *J Clin Epidemiol*. 2011;64:936–948.
54. Brierley G, Brabyn S, Torgerson D, et al. Bias in recruitment to cluster randomized trials: a review of recent publications. *J Eval Clin Pract*. 2012;18:878–886.
55. Jordhøy MS, Fayers PM, Ahlner-Elmqvist M, et al. Lack of concealment may lead to selection bias in cluster randomized trials of palliative care. *Palliat Med*. 2002;16:43–49.