Review of recent Methodological Developments in group-randomized trials: Part 1 - Design

Mélanie Prague, Elizabeth Turner, Li Fan, Gallis John, Murray David

REVIEW OF RECENT METHODOLOGICAL DEVELOPMENTS IN GROUP-RANDOMIZED TRIALS: PART 1 - DESIGN

ABSTRACT

In 2004, Murray et al. published a review of methodological developments in both the design and analysis of GRTs. In the thirteen years since, there have been many developments in both areas. The goal of the current paper is to focus on developments in design with a companion paper to focus on developments in analysis. As a pair, these papers update the 2004 review. This design paper includes developments in topics included in the earlier review (e.g. clustering, matching, and individually randomized group treatment trials) and new topics including constrained randomization and a range of randomized designs that are alternatives to the standard parallel-arm GRT. These include the stepped wedge GRT, the pseudo-cluster randomized trial and the network-randomized GRT, which, like the parallel-arm GRT, require clustering to be accounted for in both their design and analysis.

INTRODUCTION

A group-randomized trial (GRT) is a randomized controlled trial in which the unit of randomization is a group and outcome measurements are obtained on members of those groups.[1] Also called a cluster randomized trial or community trial,[2-5] a GRT is the best comparative design available if the intervention operates at a group level, manipulates the physical or social environment, cannot be delivered to individual members of the group without substantial risk of contamination across study arms, or if there are other circumstances which warrant the design such as a desire for herd immunity or a need to estimate both the direct and indirect intervention effects in studies of infectious diseases.[1-5]

In GRTs, outcomes on members of the same group are likely to be more similar to each other than to outcomes on members from other groups.[1] Such clustering must be accounted for in the

design of GRTs to avoid under-powering the study and accounted for in the analysis to avoid under-estimated standard errors and inflated type I error for the intervention effect.[1-5]

In 2004, Murray et al.[6] published a review of methodological developments in both the design and analysis of GRTs. In the 13 years since, there have been many developments in both areas. The goal of the current paper is to focus on developments in design with a companion paper to focus on developments in analysis.[7] As a pair, these papers update the 2004 review. With both papers, we seek to provide a broad and comprehensive review to guide the reader to seek out appropriate materials for their own circumstances.

DEVELOPMENTS IN FUNDAMENTALS OF DESIGN

Clustering

In its most basic form, a GRT has a hierarchical structure with groups nested within study arm and members nested within groups. Additional levels of nesting may arise through repeated measures over time or from more complex group structures (e.g., children nested in classrooms nested in schools). When designing and analyzing a GRT, it is necessary to account for the clustering associated with the nested design.[1-5]

The intraclass correlation coefficient (ICC) is the clustering measure most commonly used in power calculations and reported in published studies.[8] Eldridge et al.[9] provide a comprehensive review of ICC definitions and measures in general clustered data for both continuous and binary outcomes, the most commonly reported outcomes in GRTs.[10,11] Whereas the ICC for continuous outcome measures is well-defined and generally well understood,[1-4] Eldridge et al.[9] highlight some of the challenges for binary outcomes and provide several definitions (see **Table 1** for the form most commonly presented in GRT texts).[2,4,5,9] Others compare methods to estimate the ICC

of a binary outcome.[12-17] The ICC is not easily defined for rates based on person-time data.[2,4] Recent publications have defined ICC for time-to-event data.[18,19]

The coefficient of variation (CV) is a measure of clustering that is defined for general clustered data when the distributional parameter of interest is a mean, proportion, or rate.[3,17] The CV and ICC for continuous and binary outcomes are related by a mathematical relationship as a function of the distributional parameter of interest (i.e. mean or proportion) and, for continuous outcomes, of the within-group variance, σ_w^2 (**Table 1**).[2,4] Hayes and Moulton[2] advocate for the CV generally in power calculations; Donner and Klar agree for event data analyzed as rates.[3]

[TABLE 1 ABOUT HERE]

Given the central role of clustering in planning GRTs, imprecision in the estimated level of clustering can lead to an under-powered trial. Multiple authors address imprecision, and all focus on the ICC.[20-26] Simultaneously, there has been an increasing number of publications that report ICCs (for example, Moerbeek and Teerenstra[27] provide a comprehensive list of such papers) to aid the planning of future studies, consistent with the CONSORT statement on GRTs.[28]

Cohort vs. Cross-Sectional GRT Designs

The choice between a cohort and cross-sectional GRT design (or a combination) is driven by the nature of the research question.[1] The cross-sectional design is preferred when the question is about change in a population[1] or when the time to the outcome is so short as to make a cohort study impractical (e.g., studies involving acute conditions).[2] For example, in order to observe enough participants with malaria at 6-monthly follow-up time points and to be able to draw conclusions about population-level behavior related to malaria treatment choices, Laktabai et al.[29] chose a cross-sectional design in which different population samples were obtained at each

follow-up time point. In contrast, when interested in change in specific individuals, or in mediation, the most natural choice is the cohort design in which a cohort of individuals is enrolled and followed up over time.[1] For example, Turner et al.[30] chose such a design to study child outcomes in mothers with prenatal depression. Similarly, the cohort design is usually required to generate event data in individuals.[2] A combination design could be used whereby the cross-sectional design is augmented by subsampling a cohort of individuals who are followed over time, such as in the COMMIT study.[31] A recent review[32] indicated that the cohort design is the most common GRT design (67% of 75 GRTs).

DEVELOPMENTS IN THE DESIGN OF PARALLEL-ARM GROUP-RANDOMIZED TRIALS

Baseline Imbalance of Group Sample Size

Imbalance of group sample size means that group sizes are different across the groups randomized in the study, with implications for statistical efficiency. Donner discussed variation in group size for GRTs for a design stratified by group size.[33] Guittet et al.[34] and Carter[35] studied the impact on power using simulations, which showed the greatest reduction in power with few groups and/or high ICC. Several authors have offered adjustments to the standard sample size formula for a GRT to correct for variability in group size based on the mean and variance of the group size, or the actual size of each group.[36-39] Others have offered adjustments based on relative efficiency.[40-43] Candel et al.[40,41] reported that relative efficiency ranged from 1.0-0.8 across a variety of distributions for group size with lower values for higher ICCs and greater variability in group size; the minimum relative efficiency was usually no worse than 0.9 for continuous outcomes. They recommended dividing the result from standard formulae for

93 balanced designs by the relative efficiency for the expected group-size distribution, which was a
94 function of the ICC and the mean and variance of the group size.[40] For binary outcomes, they
95 suggested an additional correction factor based on the estimation method planned for the
96 analysis.[41] You et al.[42] defined relative efficiency in terms of non-centrality parameters; their
97 measure of relative efficiency was a function of the ICC, the mean and variance of the group
98 size, and the number of groups per study arm. Candel and Van Breukelen[43] considered variability
99 not only in group size but also between arms in error variance and the number of groups per arm.
100 They recommended increasing the number of groups in each arm by the inverse of the relative
101 efficiency minus one. Their estimate of the relative efficiency was a function of the number of
102 groups per study arm, the ICC in each study arm, the ratio of the variances in the two study arms,
103 and the mean and variance of the group size.

104 Consistent across these papers was the recommendation that expectations for variation in group
105 sample size be considered during both the planning stages and the analysis stage. Failure in
106 planning can result in an underpowered study[40-43] while failure in analysis can result in type I
107 error rate inflation.[44]

108 **Baseline Imbalance of Covariates**

109 Imbalance of covariates at baseline threatens the internal validity of the trial. Yet GRTs often
110 randomize a limited number of groups that are heterogeneous in baseline covariates and in
111 baseline outcome measurements. As a result, there is a good chance of baseline covariate
112 imbalance.[6,45] Restricted randomization strategies such as stratification, matching or constrained
113 randomization can be implemented in the design phase to address this issue. However,
114 stratification may have limited use in GRTs if there are more than a handful of covariates to

balance, due to the small number of groups in most trials.[46] Pair-matching also comes with several disadvantages[46] as it affects the proper calculation of ICC[47] and complicates the significance testing of individual-level risk factors.[48] More recently, Imai et al. presented a design-based estimator,[49] which led them to advocate for the use of pair-matching based on the unbiasedness and efficiency of their estimator. Several others highlighted features of this work,[50-52] including the authors' power calculation that does not depend on the ICC, thus avoiding the known ICC problem.[53] Despite efficiency gains of pair-matching over stratification, a simulation study conducted by Imbens led him to conclude that stratified randomization would generally be preferred to pair-matching.[54] We note that strata of size four provide virtually all the advantages of pair-matching while avoiding the disadvantages, and may be preferred over pair-matching for that reason.

To overcome challenges when trying to balance on multiple, possibly continuous, covariates, Raab and Butcher[55] proposed constrained randomization. It is based on a balancing criterion calculated by a weighted sum of squared differences between the study arm means on any group-level or individual-level covariate and seeks to offer better internal validity than both pair-matching and stratification. The approach randomly selects one allocation scheme from a subset of schemes that achieve acceptable balance, identified based on having the smallest values of the balancing criterion. Carter and Hood[56] extended this work to randomize multiple blocks of groups and provided an efficient computer program for public use. The "best balance" score was proposed to measure imbalance of group-level factors under constrained randomization.[57] In simulations with 4 to 20 groups, constrained randomization with the "best balance" score was shown to optimally reduce quadratic imbalances compared with simple randomization, matching and minimization.

Li et al.[58] systematically studied the design parameters of constrained randomization for continuous outcomes, including choice of balancing criterion, candidate set size, and number of covariates to balance. With extensive simulations, they demonstrated that constrained randomization with a balanced candidate subset could improve study power while maintaining the nominal type I error rate, both for a model-based analysis and for a permutation test, as long as the analysis adjusted for potential confounding. Moulton[59] proposed to check for overly constrained designs by counting the number of times each pair of groups received the same study arm allocation. He revealed the risk of inflated type I error in overly constrained designs using a simulation example with 10 groups per study arm. Li et al. further noticed the limitation of overly constrained designs in that they may fail to support a permutation test with a fixed size.[58] In practice, if covariate imbalance is present even after using one of the design strategies described, such imbalance can be accounted for using adjusted analysis that is either pre-planned in the protocol or through post-hoc sensitivity analysis.[7] In summary, constrained randomization seeks to provide both internal validity and efficiency.

Methods and Software for Power and Sample Size

If the ICC is positive, not accounting for it in the analysis will inflate the type I error rate, and the power of the trial will be unknown. If the ICC is estimated as negative, as it can be when the true value is close to zero and sampling error leads to a negative estimate or when there is competition within groups,[1-4,9,60] not accounting for it will reduce the type I error rate so that the test is more conservative, and the power of the trial will be lower than planned.[61] Thus, a good estimate of the ICC is essential for sample size calculation for all GRTs.

One of the simplest power analysis methods often offered for a standard parallel-arm GRT with a single follow-up measurement is to compute the power for an individually randomized trial using the standard formula, and to then inflate this by the design effect,[62] given by $1 + (m - 1)\rho$. In this formula, m is the number of subjects per group and ρ is the ICC. Unfortunately, this approach only addresses the first of the two penalties associated with group-randomization that were identified by Cornfield almost 40 years ago:[63] extra variation and limited degrees of freedom for the test of the intervention effect. In order to accurately estimate sample size and power for a GRT, it is necessary to also account for the limited degrees of freedom that can arise due to having few groups to randomize. This can be achieved by using appropriate methods detailed in one the GRT texts rather than using the naïve approach of simply inflating the individually randomized trial sample size by the design effect .[1-5,61] In general, appropriate methods calculate sample size using a variance estimate inflated based on the expected ICC and use a t-test rather than a z-test to reflect the desired power and type I error rate, with degrees of freedom based on the number of groups to be randomized.

In practice, both cross-sectional and cohort GRTs are commonly powered based on a comparison between study arms at a single point in time. Then, for GRTs with cohort designs, the analysis section of the study protocol may state that power will be gained by accounting for the repeated measures design in the analysis. However, methods exist for directly computing power in the case of repeated measures in the context of both cross-sectional and cohort designs.[1,27] Authors have noted that regression adjustment for covariates often reduces both the ICC and the residual variance, thereby improving power.[1,64] Heo et al.[65] and Murray et al.[66] provide methods that utilize data from across the entire course of the study, rather than just comparing two means at the end of the study. In practice, the user would require estimates of the variance reduction

expected from repeated measures or from regression adjustment for covariates, which could be obtained from prior studies or pilot data.

Methods exist to power GRTs with additional layers of clustering, whether from additional structural hierarchies[1,67-69] or from the repeated measures in the cohort design.[1,27,64,66,70-73] Konstantopoulos describes how to incorporate cost into the power calculation for three-level GRTs.[74] Hemming et al. discuss approaches to take when the number of groups is fixed ahead of time.[75] Two recent papers focus specifically on binary outcome variables.[13,76] Candel et al. examine the effects of varying group sizes in the context of a two-arm GRT.[77] Durán Pacheco et al. focus on power methods for overdispersed counts.[78]

Rutterford et al. and Gao et al. summarize a wide array of methods for sample size calculations in GRTs,[79,80] including for GRT designs involving 1-2 measurements per member or per group and for designs involving 3 or more measurements per member or per group. A new textbook on power analysis for studies with multilevel data also provides a thorough treatment.[27] Previous textbooks on the design and analysis of GRTs devoted at least a chapter to methods for power and sample size.[1-5]

[TABLE 2 ABOUT HERE]

DEVELOPMENTS IN THE DESIGN OF ALTERNATIVES TO THE PARALLEL-ARM GRT

We discuss four alternatives that can be used in place of a traditional parallel-arm GRT (**Figure 1A, Table 3**). All of these four designs involve randomization and some form of clustering that must be appropriately accounted for in both the design and analysis. As such, they share key

features of the standard parallel-arm GRT yet all have distinct and different features that are important to understand. In practice, some of these designs are still poorly understood.

[TABLE 3 ABOUT HERE]

[FIGURE 1 ABOUT HERE]

Stepped Wedge GRT

The stepped wedge GRT (SW-GRT) is a one-directional crossover GRT in which time is divided into intervals and in which all groups eventually receive the intervention (**Figure 1B**).[81] Systematic reviews indicate increasing popularity.[82-84] Both Trials (2015) and the Journal of Clinical Epidemiology (2013) recently published special issues focused on the design and analysis of SW-GRTs.

The rationale for this alternative is primarily logistical, i.e., it may not be possible to roll out the intervention in all groups simultaneously,[85-88] although a staggered parallel-arm GRT design could alternatively be used in which blocks of groups were randomized to intervention or control instead of all groups eventually receiving the intervention as in the SW-GRT.[89-91] Others propose a SW-GRT for ethical and acceptability reasons because all groups eventually receive the intervention.[82] This second argument has been discounted as the intervention could be delivered to all control groups at the end of a parallel-arm GRT design,[88,92] often earlier than would be the case in a SW-GRT.[93] When SW-GRTs are conducted in low incidence settings, Hayes et al. emphasized that the order and period of intervention allocation is crucial.[94]

As for the parallel-arm GRT, design choices include cross-sectional[82] vs. cohort[95] with most SW-GRT methodological literature focused on cross-sectional designs whereas most published SW-

GRTs are cohort designs.[96] An additional variation is that of complete vs. incomplete SW-GRTs defined according to whether each group is measured at every time point.[90] Regardless of the specifics of the SW-GRT design, it is important to consider the possible confounding and moderating effects of time in the analysis.[85,90,97-99] Failure to account for both, if they exist, will threaten the internal validity of the study.

Cross-sectional SW-GRT sample size formulae are available for complete and incomplete designs.[90,100-103] Hemming et al. provide a unified approach for the design of both parallel-arm and SW-GRTs and allow for multiple layers of clustering.[90] Cohort SW-GRT sample size calculation relies on simulation.[97,104] Recent work on optimal designs shows that, for large studies, the optimal design is a mixture of a stepped wedge trial embedded in a parallel-arm trial.[105,106] Moerbeek & Teerenstra devote a chapter to sample size methods for SW-GRTs.[27]

Network-Randomized GRT

GRTs have historically been used to minimize the contamination between study arms; such contamination is also called interference.[107] This contamination may give rise to a network of connections between individuals both within- and between-study arms. The latter is of particular relevance to GRT design because it leads to reduced power, although sample size methods exist to preserve power and efficiency.[108]

The network-randomized GRT is a novel design that uses network information to address the challenge of potential contamination in GRTs of infectious diseases.[109-111] In such a design, groups are defined as the network contacts of a disease (index) case and those groups are randomized to study arms. Examples include the snowball trial and the ring trial, each with a distinct way in which the intervention is delivered. In the snowball trial, only the index case

directly receives the intervention, which he is encouraged to share with his contacts (e.g. see Latkin et al.[109] for such a trial of HIV prevention in injection drug users). In the ring trial, 'rings' of contacts of the index case are randomized to receive the intervention (**Figure 1C**). This design has been used to study foot-and-mouth,[112] smallpox,[113] and Ebola.[114] For the same sample size, ring trials are more powerful than classical GRTs when the incidence of the infection is low.[115]

Pseudo-Cluster Randomized Trial

In GRTs where all members of the selected groups are recruited to the study, study participants are expected to be representative of the underlying population and, as a result, selection bias is expected to be minimal. In contrast, GRTs with unblinded recruitment after randomization are at risk of selection bias. For example, consider a GRT used to evaluate the effect of a behavioral intervention delivered by providers in the primary care setting. If a provider is first randomized to study arm and then prospectively recruits participants, he may differentially select participants depending on whether he is randomized to the intervention or control arm.[116]

To reduce the risk of such selection bias, Borm et al. introduced the pseudo-cluster randomized trial (PCRT) to allocate intervention to participants in a two-stage process.[117] In the first stage, providers are randomized to a patient allocation-mix (e.g., patients predominantly randomized to intervention vs. patients predominantly randomized to control). In the second stage, patients recruited to the PCRT are individually randomized to intervention or control according to the allocation probability of their provider (e.g., 80% to intervention vs. 20% to intervention) (**Figure 1D**).

An obvious threat to a PCRT design is that the same providers are asked to implement both the intervention and the control arms, depending on which patient they are seeing. Concerns about

contamination are a common reason to randomize providers (i.e. group randomization) so that they deliver either the intervention or the control but not both. The PCRT design would not be appropriate if there are concerns about contamination, and if they exceed concerns about selection bias.

In two published cases, providers were blinded to the two-stage form of randomization and instead assumed that patients were individually randomized to the intervention arm with equal probability.[118,119] Later publications indicate that the PCRT design did well at balancing contamination and selection bias in both studies.[120-122]

Borm et al. provide sample size calculations for continuous outcomes.[117] The clustering by provider (or unit of first stage randomization) must be accounted for in both the design and analysis. No explicit sample size methods are known to be available for non-continuous outcomes. Moerbeek & Teerenstra devote a chapter to sample size methods for PCRTs.[27]

Individually Randomized Group Treatment Trial

Pals et al.[123] identified studies that randomize individuals to study arms but deliver interventions in small groups or through a common change agent as individually randomized group-treatment (IRGT) trials, also called partially clustered or partially nested designs (**Figure 1E**).[72,124] Examples include studies of psychotherapy,[125] weight loss,[126] reduction in sun exposure,[127] and many other outcomes. Clustering associated with these small groups or change agents must be accounted for in the analysis to avoid type I error rate inflation.[72,123,124,128,129] Even so, this accounting appears to be rare in practice.[123,130-133]

Recent papers have reported sample size formulae for IRGT trials with clustering in only one study arm, both for balanced[72,123,128,134] and unbalanced designs.[77,128] Moerbeek & Teerenstra

devote a chapter to sample size methods for IRGT trials focused on methods with clustering in either one or both arms.[27] Roberts addresses sample size methods for IRGT trials in which members belong to more than one small group at the same time or change small groups over the course of the study.[135] Both features have been shown to increase the type I error rate if ignored in the analysis.[135,136]

DISCUSSION

We have summarized many of the most important advances in the design of GRTs during the 13 years since the publication of the earlier review by Murray et al.[6] Many of these developments have focused on alternatives to the standard parallel-arm GRT design, as well as those related to the nature of clustering and its features in all of the designs presented. Space limitations have prevented us from including recent developments involving pilot and feasibility GRTs and group designs such as cutoff designs and regression discontinuity applied to groups. Interested readers are directed to the recently launched *Pilot and Feasibility Studies* peer-reviewed journal and related references [4,137] and to cutoff design references by Pennell et al.[138] and by Schochet.[139]

Through this review, we have sought to ensure that the reader is reminded of the value of good design and gains knowledge in the fundamental principles of a range of recent and potentially beneficial design strategies. Pairing this knowledge with our companion review of developments in the analysis of GRTs,[7] we hope that our work leads to continued improvements in the design and analysis of GRTs.

APPENDIX: GLOSSARY

Balanced candidate subset: In constrained randomization, where a subset of randomization schemes is chosen that has "sufficient balance across potentially confounding covariates" according to "some pre-specified balance metric."[58]

Baseline covariate balance: The group-level and individual-level covariate distributions are similar in all study arms.[55]

Candidate set size: "The number of possible randomization schemes in a specific implementation."[58] "Simple randomization draws from the complete set of candidate schemes, while constrained randomization considers a subset of schemes."[58]

Choice of balancing criterion: Li et al. describe several balancing criteria to assess how well a GRT is balanced across covariates. These include the "best balance" (BB) metric of de Hoop et al.,[57] the balance criterion (B) of Raab and Butcher,[55] and the total balance score introduced by Li et al.[58]

Coefficient of variation: A measure of between-group variation, defined in Table 1.

Cohort GRT design: A cohort of individuals is enrolled at baseline and those same individuals are followed up over time.

Constrained randomization: Refers "to those designs that go beyond the basic design constraints to specify classes of randomization outcomes that satisfy certain balancing criteria, while retaining validity of the design."[59]

Cross-sectional GRT design: A different set of individuals is obtained at each time point.

Designed balance at the group level: When there are equal numbers of groups randomized to each study arm.

Equivalence: Assessing whether the new intervention is equivalent to the comparison intervention.

Individually Randomized Group Treatment Trials: Studies that randomize individuals to study arms but deliver treatments in small groups or through a common change agent.[123]

Intraclass correlation: A measure of between-group variation, defined in Table 1.

Minimization in GRTs: When the researchers allocate groups to intervention arms based on groups-specific characteristics in order to achieve a high degree of balance by minimizing the differences between intervention arms.[57] May be performed sequentially or all at once when group characteristics are known at the beginning of the study.

Network-Randomized GRT: The network-randomized GRT is a novel design that uses network information to address the challenge of potential contamination in GRTs of infectious diseases.[109-111]

Non-inferiority: When a trial is designed to show that the new intervention is not worse than the comparison intervention.

Pair-matching: At randomization, when groups are matched based on factors thought to be related to the outcome. Then within each pair of groups, one is allocated at random to one study arm and the other to the comparison study arm.[140]

Pseudo-cluster randomized trial: Intervention is allocated to individuals in a two-stage process. In the first stage, providers are randomized to a patient allocation-mix. In the second stage, patients recruited to the PCRT are individually randomized to intervention or control according to the allocation probability of their provider.

Selection bias: In some GRTs, groups are randomized before participant recruitment. This can lead to selection bias if researchers (either consciously or unconsciously) recruit specific participants for inclusion in treatment and exclude others based on certain participant characteristics, even when the aforementioned participants are all eligible for participation in the trial (see Farrin et al.[116]).

Stepped Wedge GRT: A one-directional crossover GRT in which time is divided into intervals and in which all groups eventually receive the intervention (**Figure 1B**).[81]

Stratification: At randomization, when groups are placed into strata based on factors thought to be related to the outcome.[141] Then groups are separately randomized within each strata.

Superiority: When a trial is designed to establish whether a new intervention is superior to the comparison intervention (e.g., another drug, a placebo, enhanced usual care). However, the statistical test is still two-sided, allowing for the possibility that the new intervention is actually worse than the comparison.

ACKNOWLEDGEMENTS

To be added later to avoid unblinding during the review process.

CONTRIBUTORS

To be added later to avoid unblinding during the review process.

HUMAN PARTICIPANT PROTECTION

No human subjects participated in this research therefore no IRB approval was sought.

Figure 1 The Parallel-Arm GRT and Alternative Group Designs

Abbreviation: GRT – Group-randomized trial.

Each pictorial representation is an example of the specific design in which baseline measurements are taken. Other versions of each design exist. All examples show 5 individuals per group.

*The stepped wedge group-randomized trial is a one-directional crossover GRT in which time is divided into intervals and in which all groups eventually receive the intervention, indicated by the shading of the boxes in the figure. The design shown in this figure is known as a "complete design"—that is, every group is measured at every time point. Like parallel-arm GRTs, SW-GRTs can either be cross-sectional or cohort.

†In the PCRT, a group randomized to "intervention" contains a larger proportion of group members receiving the intervention than a group randomized to control.

REFERENCES

1. Murray DM. *Design and Analysis of Group-Randomized Trials.* New York, NY: Oxford University Press; 1998.
2. Hayes RJ, Moulton LH. *Cluster Randomised Trials.* Boca Raton: CRC Press; 2009.
3. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research.* London: Arnold; 2000.
4. Eldridge S, Kerry S. *A Practical Guide to Cluster Randomised Trials in Health Services Research.* Vol 120: John Wiley & Sons; 2012.
5. Campbell MJ, Walters SJ. *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research.* Chichester, West Sussex: John Wiley & Sons; 2014.
6. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health.* 2004;94(3):423-432.
7. Turner EL, Prague M, Gallis JA, Li F, Murray DM. Review of Recent Methodological Developments in Group-Randomized Trials: Part 2 - Analysis. *Am J Public Health.* Submitted.
8. 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(2):99-107.
9. Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. *Int Stat Rev.* 2009;77(3):378-394.
10. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials.* 2016;17(1):72.
11. Rutterford C, Taljaard M, Dixon S, Copas A, Eldridge S. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. *J Clin Epidemiol.* 2015;68(6):716-723.
12. Ridout MS, Demetrio CG, Firth D. Estimating intraclass correlation for binary data. *Biometrics.* 1999;55(1):137-148.
13. Chakraborty H, Moore J, Hartwell TD. Intracluster correlation adjustments to maintain power in cluster trials for binary outcomes. *Contemp Clin Trials.* 2009;30(5):473-480.
14. Thomson A, Hayes R, Cousens S. Measures of between-cluster variability in cluster randomized trials with binary outcomes. *Stat Med.* 2009;28(12):1739-1751.
15. Yelland LN, Salter AB, Ryan P. Performance of the modified Poisson regression approach for estimating relative risks from clustered prospective data. *Am J Epidemiol.* 2011;174(8):984-992.
16. Crespi CM, Wong WK, Wu S. A new dependence parameter approach to improve the design of cluster randomized trials with binary outcomes. *Clin Trials.* 2011;8(6):687-698.
17. Wu S, Crespi CM, Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemp Clin Trials.* 2012;33(5):869-880.
18. Jahn-Eimermacher A, Ingel K, Schneider A. Sample size in cluster-randomized trials with time to event as the primary endpoint. *Stat Med.* 2013;32(5):739-751.
19. Oliveira IR, Molenberghs G, Demétrio CG, Dias CT, Giolo SR, Andrade MC. Quantifying intraclass correlations for count and time-to-event data. *Biom J.* 2016;58(4):852-867.
20. Ukoumunne OC, Davison AC, Gulliford MC, Chinn S. Non-parametric bootstrap confidence intervals for the intraclass correlation coefficient. *Stat Med.* 2003;22(24):3805-3821.
21. Zou G, Donner A. Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. *Biometrics.* 2004;60(3):807-811.
22. Turner RM, Toby Prevost A, Thompson SG. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Stat Med.* 2004;23(8):1195-1214.

23. 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(2):108-118.
24. Turner RM, Omar RZ, Thompson SG. Constructing intervals for the intracluster correlation coefficient using Bayesian modelling, and application in cluster randomized trials. *Stat Med.* 2006;25(9):1443-1456.
25. Braschel MC, Svec I, Darlington GA, Donner A. A comparison of confidence interval methods for the intraclass correlation coefficient in community-based cluster randomization trials with a binary outcome. *Clin Trials.* 2016;13(2):180-187.
26. Shoukri MM, Donner A, El-Dali A. Covariate-adjusted confidence interval for the intraclass correlation coefficient. *Contemp Clin Trials.* 2013;36(1):244-253.
27. Moerbeek M, Teerenstra S. *Power Analysis of Trials with Multilevel Data.* Boca Raton: CRC Press; 2016.
28. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. *Br Med J.* 2004;328(7441):702-708.
29. Laktabai J, Lesser A, Platt A, et al. An innovative public-private partnership to target subsidized antimalarials: a study protocol for a cluster randomized controlled trial to evaluate a community intervention in Western Kenya. In press at BMJ Open.
30. Turner EL, Sikander S, Bangash O, et al. The effectiveness of the peer delivered Thinking Healthy Plus (THPP+) Programme for maternal depression and child socio-emotional development in Pakistan: study protocol for a three-year cluster randomized controlled trial. *Trials.* 2016;17(1):442.
31. Community Intervention Trial for Smoking Cessation (COMMIT): summary of design and intervention. *J Natl Cancer Inst.* 1991;83(22):1620-1628.
32. Murray DM, Pals SP, Blitstein JL, Alfano CM, Lehman J. Design and analysis of group-randomized trials in cancer: a review of current practices. *J Natl Cancer Inst.* 2008;100(7):483-491.
33. Donner A. Sample size requirements for stratified cluster randomization designs. *Stat Med.* 1992;11(6):743-750.
34. Guittet L, Ravaud P, Giraudeau B. Planning a cluster randomized trial with unequal cluster sizes: practical issues involving continuous outcomes. *BMC Med Res Methodol.* 2006;6:17.
35. Carter B. Cluster size variability and imbalance in cluster randomized controlled trials. *Stat Med.* 2010;29(29):2984-2993.
36. Lake S, Kaumann E, Klar N, Betensky R. Sample size re-estimation in cluster randomization trials. *Stat Med.* 2002;21(10):1337-1350.
37. Manatunga AK, Hudgens MG, Chen SD. Sample size estimation in cluster randomized studies with varying cluster size. *Biom J.* 2001;43(1):75-86.
38. Kerry SM, Bland JM. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Stat Med.* 2001;20(3):377-390.
39. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol.* 2006;35(5):1292-1300.
40. van Breukelen G, Candel M, Berger M. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Stat Med.* 2007;26(13):2589-2603.
41. Candel MJ, Van Breukelen GJ. Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. *Stat Med.* 2010;29(14):1488-1501.
42. You Z, Williams OD, Aban I, Kabagambe EK, Tiwari HK, Cutter G. Relative efficiency and sample size for cluster randomized trials with variable cluster sizes. *Clin Trials.* 2011;8(1):27-36.

43. Candel MJ, Van Breukelen GJ. Repairing the efficiency loss due to varying cluster sizes in two-level two-armed randomized trials with heterogeneous clustering. *Stat Med.* 2016;35(12):2000-2015.
44. Johnson JL, Kreidler SM, Catellier DJ, Murray DM, Muller KE, Glueck DH. Recommendations for choosing an analysis method that controls Type I error for unbalanced cluster sample designs with Gaussian outcomes. *Stat Med.* 2015;34(27):3531-3545.
45. Wright N, Ivers N, Eldridge S, Taljaard M, Bremner S. A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice. *J Clin Epidemiol.* 2015;68(6):603-609.
46. 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.
47. Donner A, Klar N. Pitfalls of and controversies in cluster randomized trials. *Am J Public Health.* 2004;26(1):2-19.
48. Donner A, Taljaard M, Klar N. The merits of breaking the matches: a cautionary tale. *Stat Med.* 2007;26(9):2036-2051.
49. Imai K, King G, Nall C. The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Stat Sci.* 2009;24(1):29-53.
50. Hill J, Scott M. Comment: The Essential Role of Pair Matching. *Stat Sci.* 2009;24(1):54-58.
51. Zhang K, Small DS. Comment: The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation. *Stat Sci.* 2009;24(1):59-64.
52. Imai K, King G, Nall C. Rejoinder: Matched Pairs and the Future of Cluster-Randomized Experiments. *Stat Sci.* 2009;24(1):65-72.
53. Klar N, Donner A. The merits of matching in community intervention trials: a cautionary tale. *Stat Med.* 1997;16(15):1753-1764.
54. Imbens GW. Experimental design for unit and cluster randomized trials. Paper presented at: Initiative for Impact Evaluation 2011; Cuernavaca, Mexico.
55. Raab GM, Butcher I. Balance in cluster randomized trials. *Stat Med.* 2001;20(3):351-365.
56. Carter BR, Hood K. Balance algorithm for cluster randomized trials. *BMC Med Res Methodol.* 2008;8:65.
57. de Hoop E, Teerenstra S, van Gaal BG, Moerbeek M, Borm GF. The "best balance" allocation led to optimal balance in cluster-controlled trials. *J Clin Epidemiol.* 2012;65(2):132-137.
58. Li F, Lokhnygina Y, Murray DM, Heagerty PJ, DeLong ER. An evaluation of constrained randomization for the design and analysis of group-randomized trials. *Stat Med.* 2015;35(10):1565-1579.
59. Moulton LH. Covariate-based constrained randomization of group-randomized trials. *Clin Trials.* 2004;1(3):297-305.
60. Snedecor GW, Cochran WG. *Statistical methods.* 8th ed. Ames: Iowa State University Press; 1989.
61. Murray DM, Hannan PJ, Baker WL. A Monte Carlo Study of Alternative Responses To Intraclass Correlation in Community Trials Is It Ever Possible to Avoid Cornfield's Penalties? *Eval Rev.* 1996;20(3):313-337.
62. Donner A, Birkett N, Buck C. Randomization by cluster sample size requirements and analysis. *Am J Epidemiol.* 1981;114(6):906-914.
63. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol.* 1978;108(2):100-102.
64. Teerenstra S, Eldridge S, Graff M, Hoop E, Borm GF. A simple sample size formula for analysis of covariance in cluster randomized trials. *Stat Med.* 2012;31(20):2169-2178.

65. Heo M, Kim Y, Xue X, Kim MY. Sample size requirement to detect an intervention effect at the end of follow-up in a longitudinal cluster randomized trial. *Stat Med.* 2010;29(3):382-390.
66. Murray DM, Blitstein JL, Hannan PJ, Baker WL, Lytle LA. Sizing a trial to alter the trajectory of health behaviours: methods, parameter estimates, and their application. *Stat Med.* 2007;26(11):2297-2316.
67. Teerenstra S, Lu B, Preisser JS, van Achterberg T, Borm GF. Sample size considerations for GEE analyses of three-level cluster randomized trials. *Biometrics.* 2010;66(4):1230-1237.
68. Heo M, Leon AC. Statistical power and sample size requirements for three level hierarchical cluster randomized trials. *Biometrics.* 2008;64(4):1256-1262.
69. Teerenstra S, Moerbeek M, van Achterberg T, Pelzer BJ, Borm GF. Sample size calculations for 3-level cluster randomized trials. *Clin Trials.* 2008;5(5):486-495.
70. Heo M. Impact of subject attrition on sample size determinations for longitudinal cluster randomized clinical trials. *J Biopharm Stat.* 2014;24(3):507-522.
71. Heo M, Leon AC. Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials. *Stat Med.* 2009;28(6):1017-1027.
72. Heo M, Litwin AH, Blackstock O, Kim N, Arnsten JH. Sample size determinations for group-based randomized clinical trials with different levels of data hierarchy between experimental and control arms. *Stat Methods Med Res.* 2014;26(1):399-413.
73. Heo M, Xue X, Kim MY. Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials with random slopes. *Comput Stat Data Anal.* 2013;60:169-178.
74. Konstantopoulos S. Incorporating cost in power analysis for three-level cluster-randomized designs. *Eval Rev.* 2009;33(4):335-357.
75. Hemming K, Girling AJ, Sitch AJ, Marsh J, Lilford RJ. Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. *BMC Med Res Methodol.* 2011;11:102.
76. Ahn C, Hu F, Skinner CS, Ahn D. Effect of imbalance and intracluster correlation coefficient in cluster randomization trials with binary outcomes when the available number of clusters is fixed in advance. *Contemp Clin Trials.* 2009;30(4):317-320.
77. Candel MJ, Van Breukelen GJ. Varying cluster sizes in trials with clusters in one treatment arm: Sample size adjustments when testing treatment effects with linear mixed models. *Stat Med.* 2009;28(18):2307-2324.
78. Durán Pacheco G, Hattendorf J, Colford JM, Mäusezahl D, Smith T. Performance of analytical methods for overdispersed counts in cluster randomized trials: Sample size, degree of clustering and imbalance. *Stat Med.* 2009;28(24):2989-3011.
79. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. *Int J Epidemiol.* 2015;44(3):1051-1067.
80. Gao F, Earnest A, Matchar DB, Campbell MJ, Machin D. Sample size calculations for the design of cluster randomized trials: A summary of methodology. *Contemp Clin Trials.* 2015;42:41-50.
81. Spiegelman D. Evaluating public health interventions: 2. Stepping up to routine public health evaluation with the stepped wedge design. *Am J Public Health.* 2016;106(3):453-457.
82. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol.* 2006;6(1):1.
83. Mdege ND, Man M-S, Taylor CA, Torgerson DJ. 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(9):936-948.
84. Beard E, Lewis JJ, Copas A, et al. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials.* 2015;16(1):1-14.

85. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials.* 2007;28(2):182-191.
86. Hargreaves JR, Copas AJ, Beard E, et al. Five questions to consider before conducting a stepped wedge trial. *Trials.* 2015;16(1):350.
87. Moulton LH, Golub JE, Durovni B, et al. Statistical design of THRio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. *Clin Trials.* 2007;4(2):190-199.
88. Prost A, Binik A, Abubakar I, et al. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies. *Trials.* 2015;16(1):351.
89. Shah More N, Das S, Bapat U, et al. Community resource centres to improve the health of women and children in Mumbai slums: study protocol for a cluster randomized controlled trial. *Trials.* 2013;14:132.
90. Hemming K, Lilford R, Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Stat Med.* 2015;34(2):181-196.
91. Kotz D, Spigt M, Arts ICW, Crutzen R, Viechtbauer W. Use of the stepped wedge design cannot be recommended: A critical appraisal and comparison with the classic cluster randomized controlled trial design. *J Clin Epidemiol.* 2012;65(12):1249-1252.
92. Kotz D, Spigt M, Arts IC, Crutzen R, Viechtbauer W. Researchers should convince policy makers to perform a classic cluster randomized controlled trial instead of a stepped wedge design when an intervention is rolled out. *J Clin Epidemiol.* 2012;65(12):1255.
93. Murray DM, Pennell M, Rhoda D, Hade EM, Paskett ED. Designing studies that would address the multilayered nature of health care. *J Natl Cancer Inst Monogr.* 2010(40):90-96.
94. Hayes RJ, Alexander ND, Bennett S, Cousens SN. Design and analysis issues in cluster-randomized trials of interventions against infectious diseases. *Stat Methods Med Res.* 2000;9(2):95-116.
95. Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G, Hargreaves JR. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials.* 2015;16(1):352.
96. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ.* 2015;350:h391.
97. Baio G, Copas A, Ambler G, Hargreaves J, Beard E, Omar RZ. Sample size calculation for a stepped wedge trial. *Trials.* 2015;16(1):354.
98. Handley MA, Schillinger D, Shiboski S. Quasi-experimental designs in practice-based research settings: design and implementation considerations. *The Journal of the American Board of Family Medicine.* 2011;24(5):589-596.
99. Liao X, Zhou X, Spiegelman D. A note on "Design and analysis of stepped wedge cluster randomized trials". *Contemp Clin Trials.* 2015;45(Pt B):338-339.
100. Hemming K, Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *J Clin Epidemiol.* 2016;69:137-146.
101. Hemming K, Girling A. A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. *Stata J.* 2014;14(2):363-380.
102. Hughes J. Calculation of power for stepped wedge design. http://tinyurl.com/hwp5dgr. Accessed January 12, 2017.
103. Hughes J. Calculation of power for stepped wedge design (means). http://tinyurl.com/jvcr5bu. Accessed January 12, 2017.
104. Baio G. SWSamp: Simulation-based sample size calculations for a Stepped Wedge Trial (and more). 2016; https://sites.google.com/a/statistica.it/gianluca/swsamp.

105. Lawrie J, Carlin JB, Forbes AB. Optimal stepped wedge designs. *Stat Probab Lett.* 2015;99:210-214.
106. Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Stat Med.* 2016;35(13):2149-2166.
107. Hudgens MG, Halloran ME. Toward causal inference with interference. *J Am Stat Assoc.* 2008;103(482):832-842.
108. Wang R, Goyal R, Lei Q, Essex M, De Gruttola V. Sample size considerations in the design of cluster randomized trials of combination HIV prevention. *Clin Trials.* 2014;11(3):309-318.
109. Latkin C, Donnell D, Liu TY, Davey-Rothwell M, Celentano D, Metzger D. The dynamic relationship between social norms and behaviors: the results of an HIV prevention network intervention for injection drug users. *Addiction.* 2013;108(5):934-943.
110. Staples PC, Ogburn EL, Onnela J-P. Incorporating Contact Network Structure in Cluster Randomized Trials. *Sci Rep.* 2015;5:17581.
111. Harling G, Wang R, Onnela J, De Gruttola V. Leveraging contact network structure in the design of cluster randomized trials. *Clin Trials.* 2016 [Epub ahead of print].
112. Keeling M, Woolhouse M, May R, Davies G, Grenfell B. Modelling vaccination strategies against foot-and-mouth disease. *Nature.* 2003;421(6919):136-142.
113. Kretzschmar M, Van den Hof S, Wallinga J, Van Wijngaarden J. Ring vaccination and smallpox control. *Emerg Infect Dis.* 2004;10(5):832-841.
114. Enserink M. High hopes for Guinean vaccine trial. *Science.* 2015;347(6219):219-220.
115. Henao-Restrepo AM. The ring vaccination trial: a novel cluster randomised controlled trial design to evaluate vaccine efficacy and effectiveness during outbreaks, with special reference to Ebola. *BMJ.* 2015;351.
116. Farrin A, Russell I, Torgerson D, Underwood M. Differential recruitment in a cluster randomized trial in primary care: the experience of the UK back pain, exercise, active management and manipulation (UK BEAM) feasibility study. *Clin Trials.* 2005;2(2):119-124.
117. Borm GF, Melis RJ, Teerenstra S, Peer PG. Pseudo cluster randomization: a treatment allocation method to minimize contamination and selection bias. *Stat Med.* 2005;24(23):3535-3547.
118. Melis RJ, van Eijken MI, Borm GF, et al. The design of the Dutch EASYcare study: a randomised controlled trial on the effectiveness of a problem-based community intervention model for frail elderly people. *BMC Health Serv Res.* 2005;5:65.
119. Pence BW, Gaynes BN, Thielman NM, et al. Balancing contamination and referral bias in a randomized clinical trial: An application of pseudo-cluster randomization. *Am J Epidemiol.* 2015;182(12):1039-1046.
120. Melis RJ, Teerenstra S, Rikkert MG, Borm GF. Pseudo cluster randomization performed well when used in practice. *J Clin Epidemiol.* 2008;61(11):1169-1175.
121. Pence BW, Gaynes BN, Adams JL, et al. The effect of antidepressant treatment on HIV and depression outcomes: results from a randomized trial. *AIDS.* 2015;29(15):1975-1986.
122. Teerenstra S, Melis R, Peer P, Borm G. Pseudo cluster randomization dealt with selection bias and contamination in clinical trials. *J Clin Epidemiol.* 2006;59(4):381-386.
123. Pals SP, Murray DM, Alfano CM, Shadish WR, Hannan PJ, Baker WL. Individually randomized group treatment trials: a critical appraisal of frequently used design and analytic approaches. *Am J Public Health.* 2008;98(8):1418-1424.
124. Baldwin SA, Bauer DJ, Stice E, Rohde P. Evaluating models for partially clustered designs. *Psychol Methods.* 2011;16(2):149-165.
125. Carlbring P, Bohman S, Brunt S, et al. Remote treatment of panic disorder: a randomized trial of internet-based cognitive behavior therapy supplemented with telephone calls. *Am J Psychiatry.* 2006;163(12):2119-2125.

126. Jeffery RW, Linde JA, Finch EA, Rothman AJ, King CM. A Satisfaction Enhancement Intervention for Long-Term Weight Loss. *Obesity.* 2006;14(5):863-869.
127. Jackson KM, Aiken LS. Evaluation of a multicomponent appearance-based sun-protective intervention for young women: uncovering the mechanisms of program efficacy. *Health Psychol.* 2006;25(1):34.
128. Roberts C, Roberts SA. Design and analysis of clinical trials with clustering effects due to treatment. *Clin Trials.* 2005;2(2):152-162.
129. Kahan BC, Morris TP. Assessing potential sources of clustering in individually randomised trials. *BMC Med Res Methodol.* 2013;13:58.
130. Pals SL, Wiegand RE, Murray DM. Ignoring the group in group-level HIV/AIDS intervention trials: a review of reported design and analytic methods. *AIDS.* 2011;25(7):989-996.
131. Lee KJ, Thompson SG. Clustering by health professional in individually randomised trials. *BMJ.* 2005;330(7483):142-144.
132. Biau DJ, Porcher R, Boutron I. The account for provider and center effects in multicenter interventional and surgical randomized controlled trials is in need of improvement: a review. *J Clin Epidemiol.* 2008;61(5):435-439.
133. Oltean H, Gagnier JJ. Use of clustering analysis in randomized controlled trials in orthopaedic surgery. *BMC Med Res Methodol.* 2015;15:17.
134. Moerbeek M, Wong WK. Sample size formulae for trials comparing group and individual treatments in a multilevel model. *Stat Med.* 2008;27(15):2850-2864.
135. Roberts C, Walwyn R. Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. *Stat Med.* 2013;32(1):81-98.
136. Andridge RR, Shoben AB, Muller KE, Murray DM. Analytic methods for individually randomized group treatment trials and group-randomized trials when subjects belong to multiple groups. *Stat Med.* 2014;33(13):2178-2190.
137. Eldridge SM, Costelloe CE, Kahan BC, Lancaster GA, Kerry SM. How big should the pilot study for my cluster randomised trial be? *Stat Methods Med Res.* 2016;25(3):1039-1056.
138. Pennell ML, Hade EM, Murray DM, Rhoda DA. Cutoff designs for community-based intervention studies. *Stat Med.* 2011;30(15):1865-1882.
139. Schochet PZ. Statistical power for regression discontinuity designs in education evaluations. *J Educ Behav Stat.* 2009;34(2):238-266.
140. Campbell M, Donner A, Klar N. Developments in cluster randomized trials and Statistics in Medicine. *Stat Med.* 2007;26(1):2-19.
141. Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. *J Clin Epidemiol.* 1999;52(1):19-26.

Table 1. Two Common Measures of Clustering for General Clustered Data for Two Common Types of Outcome

Outcome measure	Intra-cluster correlation coefficient (ICC, ρ)[a]	Coefficient of Variation (CV, k)	Relationship of ICC to CV[b]
Continuous	$\sigma_B^2/(\sigma_B^2+\sigma_W^2)$	σ_B/μ	$1/\left(1+\frac{\sigma_w^2}{k^2\mu^2}\right)$
Binary	$\sigma_B^2/\pi(1-\pi)$	σ_B/π	$k^2\pi/(1-\pi)$

Note: μ = overall mean for continuous outcome data; π = overall proportion for binary outcome data; σ_B^2 = between-group variance; σ_W^2 = within-group variance (i.e. residual error variance). As is common practice, the two clustering measures are for general clustered data and do not focus on the GRT design in which the intervention effect is of primary interest (e.g. see Chapter 2 of Hayes and Moulton[2] for more details). The intervention parameter of interest in GRT is typically: difference of means for continuous outcomes; difference of proportions, ratio of proportions or odds ratio for binary outcomes; rate difference or rate ratio for event outcomes.

[a] There are multiple definitions of the ICC for binary outcomes (see [12-17]). The specific formulation provided here is one of the simplest and most commonly used (see, for example, equation (2.4) of Hayes and Moulton[2] and equation (8) of Eldridge et al.[9]).

[b] Note that, while the relationship for binary outcomes is only a function of k and the distributional parameter of interest (π), the relationship for continuous outcomes is a function of both the distributional parameter of interest (μ) and σ_W^2.

716 **Table 2. Software for Sample Size Calculations in Parallel-Arm GRTs**

Software	Functionality
PASS	Sample size calculations for GRTs comparing two means (non-inferiority, equivalence, or superiority), two proportions (non-inferiority, equivalence, or superiority), two Poisson rates, and for a logrank test.
nQuery	Comparison of two means, proportions, and rates.
Stata	User-provided command clustersampsi. Can compute sample size for continuous, binary, and rate outcomes for two-sided tests in equal-sized arms.
R	Package CRTSize for comparing two means or two binary proportions.
SAS	No built-in functionality at this time.
Calculator	For some simple designs, parameter values can be plugged in to formulas provided in textbooks and online.

717

718

719 **Table 3. Characteristics of the Parallel-Arm Group Randomized Trial (GRT) and of Alternative Group Designs**

		One-stage randomization		Two-stage randomization	Type of follow-up possible	720
Design	**Acronym**	**By Group**	**By Individual**		**Cross-sectional**	**Cohort**
Parallel-Arm GRT	GRT	✓	-	-	✓	✓
Stepped Wedge GRT	SW-GRT	✓	-	-	✓	✓
Network-Randomized GRT	NR-GRT	✓	-	-	-	✓[1]
Pseudo-Cluster Randomized Trial	PCRT	-	-	✓	-	✓[2]
Individually Randomized Group Treatment Trial	IRGT trial	-	✓	-	-	✓[3]

[1] In the network-randomized GRT, the index case and its network is usually defined at baseline and therefore the design is expected to use a cohort design and not allow a cross-sectional design

[2] In the pseudo-cluster randomized trial, because randomization is undertaken in two stages with individuals randomized to intervention or control in the second stage, the design requires that a cohort of individuals be enrolled at study baseline in order to be followed over time

[3] In the individually randomized group treatment trial, individual randomization is performed and therefore, like the pseudo-cluster randomized trial, a cohort of individuals is enrolled and followed over time.

Figure

Click here to download Figure 2017-01-28 Figure 1 GRT Design Review FINAL.ppt

Baseline Follow-up

Legend

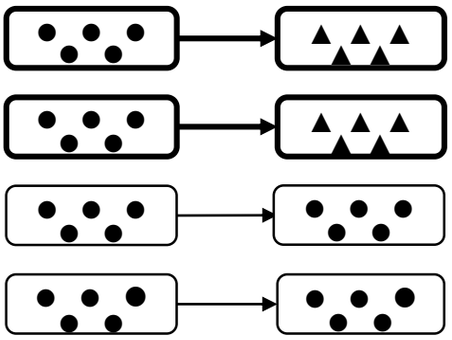
▲ Individual measured under intervention

● Individual measured under no intervention

Group randomized to intervention

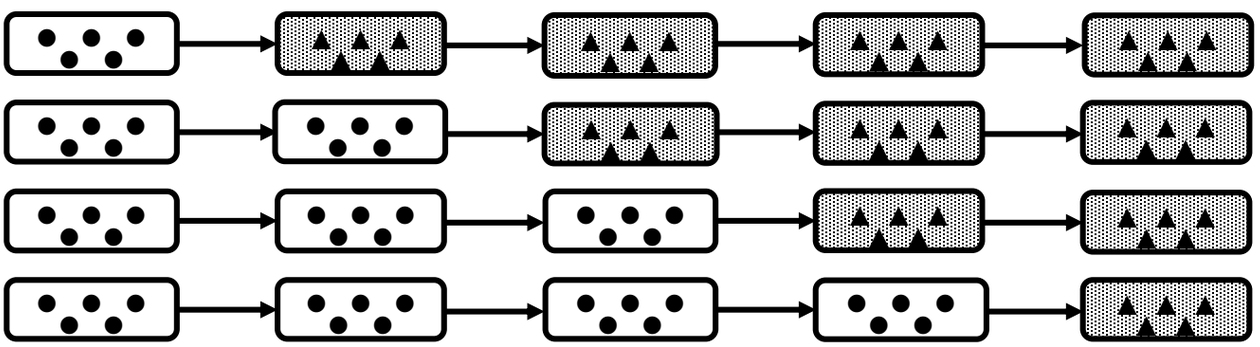
Group randomized to control

A: Parallel GRT

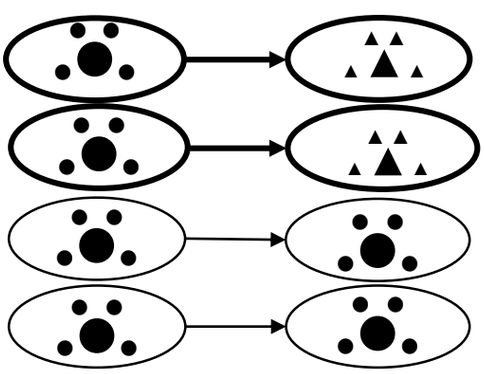


Groups are randomized to intervention or control at baseline, then either the same individuals are followed up over time (cohort GRT) or different individuals in the same group are sampled at different time points (cross-sectional GRT).

B: Stepped Wedge GRT (SW-GRT)*

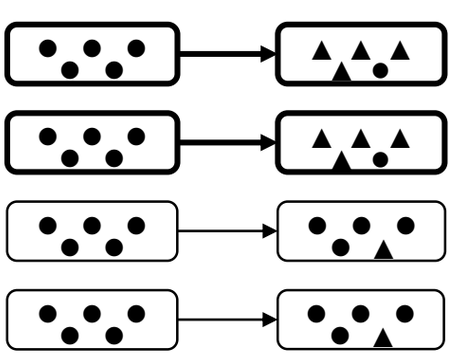


C: Network-Randomized GRT (NR-GRT)



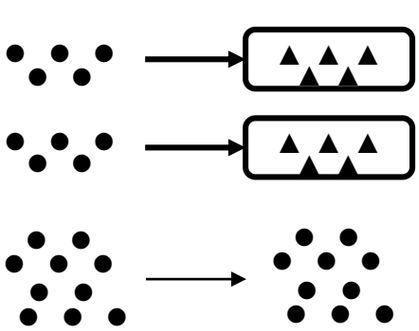
Groups are defined as the network contacts of an index disease case and those groups of contacts are then randomized to intervention or control. The larger symbols represent the index case in each group.

D: Pseudo-Cluster Randomized Trial (PCRT)†



Assignment to intervention is based on a two-stage process. In the first stage, groups (e.g., providers) are randomized to a patient allocation-mix, here shown as predominantly (80%) intervention vs. predominantly (80%) control. In the second stage, patients recruited to the PCRT are individually randomized to intervention or control.

E: Individually Randomized Group Treatment (IGRT) Trial



Individuals are randomized to intervention or control but treatments are delivered in small groups or through a common change agent.