Review of Recent Methodological Developments in Group-Randomized Trials: Part 2—Analysis

In 2004, Murray et al. reviewed methodological developments in the design and analysis of group-randomized trials (GRTs). We have updated that review with developments in analysis of the past 13 years, with a companion article to focus on developments in design.

We discuss developments in the topics of the earlier review (e.g., methods for parallel-arm GRTs, individually randomized group-treatment trials, and missing data) and in new topics, including methods to account for multiple-level clustering and alternative estimation methods (e.g., augmented generalized estimating equations, targeted maximum likelihood, and quadratic inference functions).

In addition, we describe developments in analysis of alternative group designs (including stepped-wedge GRTs, network-randomized trials, and pseudocluster randomized trials), which require clustering to be accounted for in their design and analysis. (*Am J Public Health*. 2017;107:1078–1086. doi:10.2105/AJPH.2017.303707)

Elizabeth L. Turner, PhD, Melanie Prague, PhD, John A. Gallis, ScM, Fan Li, MSc, and David M. Murray, PhD

In a group-randomized trial (GRT), the unit of randomization is a group, and outcome measurements are obtained for 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, or cannot be delivered to individual members of the group without substantial risk of contamination; it is also the best available design in other circumstances such as a desire for herd immunity in studies of infectious disease.[1–5]

In GRTs, outcomes for members of the same group are likely to be more similar to each other than to outcomes for members from other groups.[1] Such clustering must be accounted for in the design to avoid an underpowered study and in the analysis to avoid underestimated standard errors and inflated type I error for the intervention effect.[1–5] In analyses, regression modeling approaches are generally preferred and most commonly used because of their ease of implementation.[6] Several textbooks now address these and other issues.[1–5]

In 2004, Murray et al.[7] 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 each area. Here we focus on developments in analytic methods, including those relevant to our companion article that focuses on developments in GRT design.[8] (The glossary of terms is available as a supplement to the online version of this article at http://www.ajph.org.) As a pair, these articles update the 2004 review. In both, our goal is to provide a broad and comprehensive review to guide readers in seeking out appropriate materials for their own circumstances.

ANALYSIS OF PARALLEL-ARM GRTS

In GRTs, superiority trials are more common than equivalence or noninferiority trials: a PubMed search by one of the authors (D. M. M.) of studies published in 2015 identified 562 superiority GRTs but only 1 equivalence GRT and 2 noninferiority GRTs. Similarly, developments in the methods literature have focused on superiority GRTs, with developments for equivalence and noninferiority GRTs limited to small sections in 2 of the more recent textbooks[2,5] and a review article on sample size methods.[9] As a consequence, we focus here on superiority GRTs.

Methods for Intervention Effects

In GRTs, protocol violations can lead to noncompliance at either the group or member level.[5] As a means of minimizing bias, intention-to-treat principles are recommended at both levels rather than are "on-treatment" and "per-protocol" analyses.[2,4,5] Although group-level protocol violations are usually easy to identify, member-level compliance may be more difficult to ascertain in practice.[2] Jo et al. demonstrated that analyses ignoring compliance information may be underpowered to detect an intention-to-treat effect, and they proposed a multilevel model combined with a mixture model.[10] The implications of group-level

ABOUT THE AUTHORS

Elizabeth L. Turner and John A. Gallis are with the Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, and the Duke Global Health Institute, Duke University. Melanie Prague is with the Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, and Inria, project team SISTM, Bordeaux, France. Fan Li is with the Department of Biostatistics and Bioinformatics, Duke University. David M. Murray is with the Office of Disease Prevention, Division of Program Coordination and Strategic Planning, and the Office of the Director, National Institutes of Health, Rockville, MD.

Correspondence should be sent to Elizabeth L. Turner, PhD, Department of Biostatistics and Bioinformatics and Duke Global Health Institute, Duke School of Medicine, Duke University, 2424 Erwin Rd, Durham, NC 27710 (e-mail: liz.turner@duke.edu). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

This article was accepted February 5, 2017.

doi: 10.2105/AJPH.2017.303707

noncompliance can be considerable in GRTs given the small number of groups randomized in many such trials.

Methods Based on Randomization Scheme

Matching or stratification in designs has been recommended for some time as a way to ensure baseline balance in terms of important potential confounders,[1] with constrained randomization more recently developed.[11] Recent reports suggest that this advice is followed in most GRTs.[6,12–15] Matching and stratification in designs can be ignored in analyses of intervention effects without harm to the type I error rate, and often the saved degrees of freedom will improve power.[16,17]

Recently, Donner et al. reported that ignoring matching can adversely affect other analyses, such as analyses examining the relationship between a risk factor and an outcome[18]; for this reason, investigators considering pair matching should consider small strata instead (e.g., strata of 4). Li et al.[19] compared model-based and permutation methods in the context of constrained randomization adjusting for group-level covariates. They found that both the adjusted F test and the permutation test maintained the nominal size and exhibited improved power under constrained randomization relative to simple randomization.

Model-Based Methods

Model-based methods can be broadly classified according to the interpretation of the model parameters. Conditional model parameters, typically estimated with mixed-effects regression via maximum likelihood estimation (MLE), are referred to as cluster-specific effects (or as subject-specific effects in the longitudinal analysis literature). Effects are conditional on the random effects used to account for clustering and on other covariates included in the analysis. Conditional models are often recommended for studies focused on within-member changes or on mediation analyses.[7]

Parameters of marginal models are usually estimated via generalized estimating equations (GEE).[20,21] They define the marginal expectation of the dependent variable as a function of the independent variables and assume that the variance is a function of the mean; they separately specify a working correlation structure for observations made on members of the same group. Marginal models are often preferred for analyses of population-level effects because the intervention effect coefficient is interpreted as a population-averaged effect. In practice, marginal models are less frequently used than conditional models.[6]

Marginal and conditional intervention effects are equal for identity and log links,[22] and the distinction between them is important only for link functions such as the logit for binary outcomes. Although some authors have advocated for the log instead of logit link for binary outcomes,[23] this approach is not widely used, possibly because of model convergence problems for certain types of data.[24,25] Alternatively, a modified Poisson approach with log link and robust standard errors could be used in the GEE framework[26] because it does not suffer from the same convergence problems as the binomial model with log link[27]; however, its use may be less common because of the familiarity of logistic regression among epidemiologists and biostatisticians.

In practice, the question about which types of effects, conditional or marginal, are desired depends on the research question. It is essential to understand the underlying assumptions of each method: conditional models rely on correct specification of untestable aspects of the data distribution, whereas marginal models rely on a correct definition of the population of interest, which can make it difficult to generalize results to other populations.[28] We address each of the 2 approaches in more detail in the sections to follow.

Conditional approaches. If the mixed-effects model used to estimate conditional effects is misspecified, the estimates are difficult to interpret and, even if regression diagnostics can help,[29] standard errors are not robust. Fortunately, Murray et al.[30] and Fu[31] have shown that mixed models are robust to substantial violations of the normality assumptions for member- and group-level errors so long as equal numbers of groups are randomized to each arm. Parameter estimation via restricted maximum likelihood estimation is preferred to MLE when few groups are available.[32–34] In the case of binary outcomes, alternative methods for specifying test degrees of freedom have been examined in small-sample GRTs, and the between–within method is recommended.[32,35]

Multiple levels of clustering in conditional models. GRTs may involve multiple levels of clustering as a result of repeated measures on individuals or groups or additional hierarchical levels in the design. Murray[1] distinguished between mixed-effects models based on the number of measurements included in the analysis and recommended mixed-effects analysis of variance (ANOVA) or covariance (ANCOVA), or mixed-effects repeated measures ANOVA or ANCOVA, for analyses involving 1 or 2 measurements per person or per group; these models can account for all sources of random variation in such data if they are properly specified.[36]

However, this is not the case in analyses involving 3 or more measurements per person or per group, wherein the sources of random variation may be different; instead, such analyses require a random coefficients model in which random trends and intercepts are calculated for each member (in cohort GRT designs) or group (in cohort and cross-sectional GRT designs), average trends and intercepts are calculated for each study arm, and the intervention effect is the net difference in the average study-arm trends.[36] Trends are often estimated as linear slopes but can take another form.

Variable group size in conditional models. Johnson et al. focused on analysis of Gaussian outcomes from GRTs with variable group sizes.[37] They compared 10 model-based approaches and found that a 1-stage mixed model with Kenward–Roger[32] degrees of freedom and unconstrained variance components performed well in GRTs with 14 or more groups per study arm. A 2-stage model weighted by the inverse of the estimated theoretical variance of the group means and with unconstrained variance components performed well in GRTs with 6 or more groups per study arm. A number of other models resulted in an inflated type I error rate when there was substantial variability in group size.

Marginal approaches. When the GEE approach is used to estimate marginal effects, unbiased

intervention effects can be estimated even if the working correlation structure is incorrect (e.g., via robust standard errors with the sandwich estimator), although precision is increased if the working matrix is correct. When degrees of freedom are limited for the test of interest, as often occurs in GRTs, standard error estimation is frequently biased downward and no method corrects for this issue in all cases, although several have been proposed.[38–44]

Multiple levels of clustering in marginal models. Multilevel clustering is easy to account for in mixed-effects regression, but there is less literature for the GEE approach. The alternating logistic regression approach[45] for binary and ordinal outcomes can be used to account for correlation attributable to repeated measures on individuals within groups, and this approach can be implemented within a GEE framework in both R (the alr package) and SAS (PROC GEE).[46] The second-order GEE approach, which (by contrast with regular GEE) models the working correlation structure as a function of covariates, can be implemented in R (geepack in R[47]).[48] For more general working correlation matrices, users typically need to perform additional programming to provide the appropriate covariance matrix, and convergence may not be achieved.

In addition, although the intervention effect is unbiased when the marginal model is not correctly specified, standard errors estimated via GEE may be too small. A robust sandwich estimator of the variance can be used to correct this problem, but such an approach leads to loss of power.[49] Because of this accuracy–power trade-off, mixed-effects models may be a better option in GRTs involving more than 2 levels, although the effects estimated in such models are conditional rather than marginal effects.

Variable group size in marginal models. Although GEE analysis can accommodate variable group sizes, informative group size can negatively affect efficiency. In this case, Williamson et al.[50] showed that GEE weighted by group size can correct bias in estimated intervention effects. This approach is equivalent to and less computationally demanding than within-cluster resampling.[51]

Advanced GEE approaches to improve efficiency. For binary outcomes, GEE is more conservative (i.e., the intervention effect will be estimated closer to the null) than mixed-effects models.[28,52] Moreover, the standard error of the estimated intervention effect is typically larger when GEE is used, so much recent effort has focused on efficient estimation. GEE is most efficient when the true correlation structure of the data is selected as the working correlation structure. Hin et al. compared multiple selection criteria for the working correlation matrix.[53]

An alternative approach is augmented GEE (AU-GEE), a method developed for independent data in a causal inference framework[54] that has been extended to clustered data.[55] AU-GEE uses covariate information to improve efficiency in a 2-stage approach that specifies a model for the potential outcomes under the treatment not received. AU-GEE is unbiased and robust to misspecification of the potential outcome model, although correct specification improves efficiency. As for analysis of all trials, only baseline covariates should be included in AU-GEE for analysis of GRT data because adjustment for postbaseline covariates may lead to bias.[56] Alternative methods are available to account for postbaseline, time-varying confounding.[57–59]

Alternatives to GEE. The quadratic inference function (QIF) method is an alternative to GEE for estimation of marginal effects. Song et al.[60] demonstrated that QIF has advantages over GEE: it is more efficient and more robust to outliers, it includes a goodness-of-fit test of the marginal mean model, and it permits straightforward extensions to model selection. In large samples, QIF is more efficient than GEE when the working correlation structure for the data is misspecified.[61] However, the standard errors may be underestimated for small and medium sample sizes or for variable group sizes.[62] More recent work by Westgate[63,64] provides improvements; Westgate used a bias-corrected sandwich covariance estimate and simultaneously selected the QIF or GEE while selecting the best working correlation structure.[65] Despite the many attractive properties of QIF, at this time there are few applications in public health.[66–68]

A second alternative estimation method is targeted maximum likelihood estimation (tMLE),[69] a maximum likelihood–based G-computation estimator that targets the fit of the data-generating distribution to reduce bias in the parameter of interest. It is based on a machine learning approach that fluctuates an initial estimate of the conditional mean outcome and minimizes a loss function to provide an estimate of the parameter of interest.[70] The approach has been used in public health[71,72] and shows much promise for GRTs[73,74] because it can improve efficiency by simultaneously accounting for missing data and chance baseline covariate imbalance without committing to a specific functional form.[75]

Permutation Methods

Gail et al. introduced permutation analysis for GRTs.[76] They found that the permutation test had nominal type I error rates across a variety of settings common to GRTs when the member-level errors were Gaussian or binomial—and even when very few heterogeneous groups were randomized to each study arm and the intraclass correlation coefficient was large—so long as equal numbers of groups are randomized to each arm. Murray et al.[30] extended this work, and their results showed that unadjusted permutation tests offer no more protection against confounding than unadjusted model-based tests, whereas the adjusted versions of both tests perform similarly. The permutation test was more powerful than the model-based test when the data were binomial and the intraclass correlation coefficient was 0.01 or above. Fu[31] extended the work to heavy-tailed and very skewed distributions and reported similar results.

Li et al. compared model-based and permutation methods in the context of constrained randomization adjusting for group-level covariates. They found that the adjusted F test and the permutation test maintained the nominal size and had similar power but cautioned that the randomization distribution must be calculated within the constrained randomization space to prevent inflation of the type I error rate.[19]

DEVELOPMENTS IN THE ANALYSIS OF ALTERNATIVES

Alternative group designs can be used in place of a traditional parallel-arm GRT.[8] Four of these alternatives involve randomization and some form of clustering that must be appropriately accounted for in both their design and analysis. Thus, they share key features of the standard parallel-arm GRT, yet all have distinct and different features that are important to understand.

Stepped-Wedge GRTs

Both between- and within-group information is available to estimate the intervention effect from a stepped wedge GRT (SW-GRT).[77,78] However, because the control condition is usually observed earlier than the intervention condition, time is a potential confounder and should be accommodated in analyses of SW-GRTs, typically by accounting for time as a predictor.[79] As with parallel GRTs, clustering by group must be taken into account, and longitudinal measures for individuals can be accommodated within either the mixed-effects or the GEE framework, although more easily with mixed-effects models (see the sections on multiple levels of clustering). Conditional approaches are more commonly used in practice and reported on in the methods literature.[79,80] Several authors have highlighted other characteristics specific to SW-GRTs, including lagged intervention effects[81] and fidelity loss over time.[79]

Network-Randomized GRTs

Because the network properties of a network-randomized GRT are primarily used at the design stage,[82] and because they differ from regular GRTs only in the novel way in which groups are defined, theories regarding analysis of parallel-arm GRTs can be applied to parallel-arm network-randomized GRTs.[83] For example, in a ring trial of an Ebola vaccine[83] in which a network was defined as all individuals who had regular physical contact with the incident (index) case of Ebola and in which all contacts received the vaccine (placebo or active), standard GRT methods were used.

For network-randomized GRTs in which the intervention is not directly administered to all individuals and it is expected that the intervention will spread over the network (e.g., snowball trials of an HIV prevention intervention for drug users[84] or a microfinance intervention[85]), methods[86,87] are available to estimate both the direct and indirect effects of the intervention. When network information is available and the outcome of interest is known to be a disseminated process, adjusting for network features such as information on the location of each individual within the network (i.e., group) can improve the efficiency as well as the power of the analysis.[88]

Pseudocluster Randomized Trials

Teerenstra et al.[89] compared analytic methods for continuous outcomes in pseudocluster randomized trials, and Campbell and Walters discussed principles in their recent textbook.[5] Clustering by the unit of randomization at the first stage (e.g., provider) must be taken into account in both the design and analysis of pseudocluster randomized trials. No explicit sample size or analytic methods are known to be available for noncontinuous outcomes.

Individually Randomized Group-Treatment Trials

Baldwin et al. compared 4 analytic models for individually randomized group treatment trials and 3 methods for calculating degrees of freedom.[90] A multilevel model adapted to reflect clustering in only 1 study arm, combined with either Satterthwaite[91] or Kenward–Roger[32] degrees of freedom, resulted in better type I error control, better efficiency, and less bias, even with heteroscedasticity at the member level. This finding is consistent with earlier reports by Pals et al.[92] and Roberts and Roberts.[93] More recently, Roberts and Walwyn[94] and Andridge et al.[95] considered circumstances in which members are associated with more than 1 small group or change agent. Both found that ignoring membership in multiple groups further inflates the type I error rate. Roberts and Walwyn reported that multiple-member multilevel models maintained the nominal type I error rate; they also provided sample size and power formulas.[94]

DEVELOPMENTS IN ADDRESSING DATA CHALLENGES

Data challenges include those related to missing outcome data, baseline imbalance of covariates, and practical implementation in software.

Missing Outcome Data

Two recent reviews[6,96] indicate that missing outcome data are common in GRTs, although investigators frequently analyze only available data without accounting for the missing data pattern. In cases in which the covariate-dependent missingness (CDM) assumption is plausible, both mixed-effects and GEE models provide unbiased estimates of the intervention effect when the CDM covariates are included in analyses of all available data.[97,98] AU-GEE also can provide unbiased effects through inclusion of all CDM covariates in the augmentation component,[55] and it has the advantage that all estimates can still be interpreted as marginal effects. Other 2-stage approaches such as multiple imputation (MI) and inverse probability weighting (IPW) can provide unbiased intervention effects under certain conditions for more general missing-at-random patterns and may provide increased precision relative to covariate-adjusted conditional or marginal models for CDM.[97,99]

Although there is less literature on how to address missing-not-at-random data,[100] sensitivity analyses are recommended.[101] A recent review showed that very few GRTs incorporated sensitivity analyses for missing data assumptions.[6]

To avoid possible type I error, MI should account for the clustered data structure.[102,103] Fixed group effects should not be used owing to reduced power.[104] For binary outcomes, Ma et al.[105] and Caille et al.[106] showed that the preferred MI method depends on the number of groups and the design effect, and they noted that bias may arise for some approaches (including CDM). Using group-specific mean imputation may be adequate for continuous outcomes.[98,102] Hossain et al.[98] showed that if the missing data mechanism includes an interaction between a covariate predictive of the outcome and the study arm, the imputation strategy must account for this interaction if it is to be unbiased.

Whereas MI requires specifying the distribution of the missing data conditional on covariates, IPW requires specifying the probability of missingness depending on covariates. Theoretically, both approaches can be used for any type of outcome and for CDM as well as more general missing-at-random mechanisms.[99] Although IPW requires an additional assumption of positivity (all participants have a nonzero probability of being observed), it may be viewed as easier to define, particularly in the presence of nonintermittent missingness.[107] Importantly, and as with MI, if the missing data mechanism includes an interaction between a covariate predictive of the outcome and the study arm, the weights must be generated by accounting for this interaction if the strategy is to be unbiased.[108]

Prague et al.[109,110] developed a doubly robust estimator in the context of IPW that provides an unbiased estimate of the intervention effect if either the marginal mean model or the missing data model is correctly specified. They demonstrated that a doubly robust augmented GEE approach can simultaneously account for both CDM and baseline covariate imbalance in GRTs when the parameter of interest is a marginal effect. Combining MI and IPW is a promising new approach that may be superior in performance to IPW or MI alone when there are missing covariates in addition to missing outcomes.[111]

Baseline Imbalance of Covariates

Although design strategies such as restricted randomization[8] can help to achieve baseline covariate balance, they may not be easy to implement (e.g., if group characteristics are unknown in advance), and chance imbalance may arise regardless. In this case, some form of model-based covariate adjustment could be used, such as standard multivariable regression for conditional models or AU-GEE for marginal models.[55] The advantage of AU-GEE in this case is that it is doubly robust: the consistency of the intervention effect estimate requires correct specification of either the marginal mean structure or the treatment model, and covariate adjustment is separated from intervention effect estimation, thereby reducing the risk of selecting the models that produce the most significant results. The standard multivariable regression adjustment approach does not offer either of these benefits.

Alternatively, Hansen and Bowers[112] proposed a balancing criterion and studied its randomization distribution to simultaneously test for balance of multiple covariates in both randomized controlled trials and GRTs. Leyrat et al.[113] suggested using the c-statistic of the propensity score model to measure covariate balance at the individual level. Leon et al.[114] recommended propensity score matching to correct for baseline imbalance; in a simulation study, they reported a median 90% reduction in bias. Nevertheless, the CONSORT (Consolidated Standards for Reporting of Trials) recommendation is that the adjustment covariates be specified a priori for primary analyses so that the sensitivity of the primary findings to adjustment for covariates identified post hoc can be tested in secondary analyses.[115]

Software

Table 1 presents information on 3 software programs that can be used to analyze data from GRTs. The table is organized around topics considered here. None of the 3 programs can readily implement both QIF and tMLE for

TABLE 1—Summary of Known Functions and Procedures for Analyzing Group-Randomized Trials (GRTs) via the Methods Described for Three Commonly Used Software Programs

	Software		
Method	SAS	Stata	R
Outcome analysis of all available data			
Mixed-effects models	PROC MIXED PROC NLMIXED PROC GLIMMIX	mixed melogit mepoisson	lme4 nlme
GEE	PROC GENMOD[a]	xtgee	geepack/geeM
tMLE	NA	NA	NA[b]
QIF	%qif	NA	qif[c]
Permutation tests	%ptest	NA	NA
Accounting for missing outcomes			
Multiple imputation for clustered data	%mmi_impute[d] %mmi_analyze	REALCOM-IMPUTE mi impute[d]	pan jomo[e]
Inverse probability weighting	PROC GENMOD[f]	NA[g]	CRTgeeDR
Causal inference–based methods[h]			
AU-GEE	NA	NA	CRTgeeDR
Doubly robust AU-GEE	NA	NA	CRTgeeDR

Note. AU-GEE = augmented GEE; GEE = generalized estimating equations; NA = not applicable; QIF = quadratic inference function; tMLE = targeted maximum likelihood.

[a]PROC GEE is another option, but it is in the experimental phase and has limited usefulness for GRTs over and above PROC GENMOD.

[b]In R, tmle is available for tMLE; at the time of writing, however, it did not allow for clustering.

[c]At the time of writing, we were unable to load the package, and it allows only equal cluster sizes; however, Westgate modified the code for GRTs with variable cluster sizes in the appendix of his article.[63]

[d]Only useful for continuous outcomes.

[e]In R, mice is available for multiple imputation; at the time of writing, however, it did not account for clustering.

[f]Cannot account for imprecision in weights.

[g]xtgee cannot accommodate individual-level weights but, rather, only group-specific weights.

[h]The 2 listed methods are related: AU-GEE accounts for baseline covariate imbalance, and doubly robust AU-GEE, an extension of AU-GEE, accounts for both baseline covariate imbalance and missing data.

GRTs; however, the R program offers the most ready-to-use functionality given its broad applicability to the methods we have described.

REPORTING OF RESULTS

The CONSORT guidelines for individually randomized trials were extended to GRTs in 2004,[115] and most journals now require authors to conform to these guidelines. Ivers et al. reviewed 300 GRTs published between 2000 and 2008 and reported that 60% and 70%, respectively, accounted for clustering in sample size calculation and in the analysis; 56% involved restricted randomization, and most (86%) allocated more than 4 groups per arm.[14] A more recent review of 86 trials published in 2013 and 2014 showed that 77% and 78% accounted for clustering in sample size calculation and in the analysis, respectively, and 51% involved some form of restricted randomization.[6]

Recent work on conduct and reporting has focused on the ethics of GRTs given concerns regarding this issue.[116,117] For example, Sim and Dawson discussed the challenges associated with obtaining informed consent in GRTs.[118] The Ottawa statement on ethical design and conduct of GRTs was published in 2012,[119] with a reevaluation in 2015.[120]

CONCLUSIONS

In this review, we have summarized many of the most important advances in the analysis of GRTs during the 13 years since the publication of the earlier review by Murray et al.[7] Much of our discussion has focused on marginal model parameter estimation (e.g., AU-GEE, QIF, tMLE) and missing data methods. Some topics that could not be included owing to space limitations are survival outcomes,[2,121–125] measurement bias,[126,127] validity,[128,129] Bayesian methods,[4,130–132] cost-effectiveness analyses,[4,133–136] mediation analyses seeking to uncover mechanisms of action,[137–140] and methods to analyze alternative GRT designs such as crossover GRTs.[141–144]

Our aim here has been to remind readers of the value of well-thought-out analyses of GRTs and of keeping up to date with the many recent developments in this area. We hope that this review, paired with our companion review of developments in GRT design,[8] will lead to continued improvements in the design and analysis of GRTs. AJPH

CONTRIBUTORS

E. L. Turner and D. M. Murray initiated the project and developed the outline and the topics to be covered. E. L. Turner wrote much of the first draft, to which the other authors contributed sections. All of the authors edited and reviewed the revised article.

ACKNOWLEDGMENTS

This work was partly funded by the National Institutes of Health (grants R01 HD075875, R37 AI51164, R01 AI110478, and K01 MH104310).

We thank the 2 anonymous reviewers whose comments greatly helped improve the final version of this article.

Note. The content is solely the responsibility of the authors and does not necessarily represent the official views of National Institutes of Health. The study sponsors had no influence on the study design; data collection, analysis or interpretation; content of the article; nor the authors' decision to submit this article. The researchers operated independently from the funders in these matters.

HUMAN PARTICIPANT PROTECTION

No protocol approval was needed for this project because no human participants were involved.

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, FL: CRC Press; 2009.

3. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. London, England: Arnold; 2000.

4. Eldridge S, Kerry S. *A Practical Guide to Cluster Randomised Trials in Health Services Research*. Vol. 120. New York, NY: John Wiley & Sons Inc; 2012.

5. Campbell MJ, Walters SJ. *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research*. Chichester, England: John Wiley & Sons Inc; 2014.

6. 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.

7. 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.

8. Turner EL, Li F, Gallis JA, Prague M, Murray DM. Review of recent methodological developments in group-randomized trials: part 1—design. *Am J Public Health*. 2017;107(6):907–915.

9. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. *Int J Epidemiol*. 2015;44(3):1051–1067.

10. Jo B, Asparouhov T, Muthén BO. Intention-to-treat analysis in cluster randomized trials with noncompliance. *Stat Med*. 2008;27(27):5565.

11. Raab GM, Butcher I. Balance in cluster randomized trials. *Stat Med*. 2001;20(3):351–365.

12. Varnell SP, Murray DM, Janega JB, Blitstein JL. Design and analysis of group-randomized trials: a review of recent practices. *Am J Public Health*. 2004;94(3):393–399.

13. 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.

14. 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.

15. Fiero M, Huang S, Bell ML. Statistical analysis and handling of missing data in cluster randomised trials: protocol for a systematic review. *BMJ Open*. 2015;5(5):e007378.

16. Diehr P, Martin DC, Koepsell T, Cheadle A. Breaking the matches in a paired t-test for community interventions when the number of pairs is small. *Stat Med*. 1995;14(13):1491–1504.

17. Proschan MA. On the distribution of the unpaired t-statistic with paired data. *Stat Med*. 1996;15(10):1059–1063.

18. Donner A, Taljaard M, Klar N. The merits of breaking the matches: a cautionary tale. *Stat Med*. 2007;26(9):2036–2051.

19. 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.

20. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13–22.

21. Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*. 1986;42(1):121–130.

22. Ritz J, Spiegelman D. Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Stat Methods Med Res*. 2004;13(4):309–323.

23. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*. 1987;125(5):761–768.

24. Blizzard L, Hosmer W. Parameter estimation and goodness-of-fit in log binomial regression. *Biom J*. 2006;48(1):5–22.

25. Williamson T, Eliasziw M, Fick GH. Log-binomial models: exploring failed convergence. *Emerg Themes Epidemiol*. 2013;10(1):14.

26. Zou G, Donner A. Extension of the modified Poisson regression model to prospective studies with correlated binary data. *Stat Methods Med Res*. 2013;22(6):661–670.

27. 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.

28. Hubbard AE, Ahern J, Fleischer NL, et al. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*. 2010;21(4):467–474.

29. Huang X. Diagnosis of random-effect model misspecification in generalized linear mixed models for binary response. *Biometrics*. 2009;65(2):361–368.

30. Murray DM, Hannan PJ, Varnell SP, McCowen RG, Baker WL, Blitstein JL. A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial. *Stat Med*. 2006;25(3):375–388.

31. Fu D. *A Comparison Study of General Linear Mixed Model and Permutation Tests in*

Group-Randomized Trials Under Non-Normal Error Distributions [dissertation]. Memphis, TN: University of Memphis; 2006.

32. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*. 1997;53(3):983–997.

33. Localio AR, Berlin JA, Have TRT. Longitudinal and repeated cross-sectional cluster-randomization designs using mixed effects regression for binary outcomes: bias and coverage of frequentist and Bayesian methods. *Stat Med*. 2006; 25(16):2720–2736.

34. Pinheiro JC, Bates DM. *Mixed-Effects Models in S and S-PLUS*. New York, NY: Springer; 2000.

35. Li P, Redden DT. Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Med Res Methodol*. 2015;15(1):38.

36. Murray DM, Hannan PJ, Wolfinger RD, Baker WL, Dwyer JH. Analysis of data from group-randomized trials with repeat observations on the same groups. *Stat Med*. 1998;17(14): 1581–1600.

37. 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.

38. McNeish D, Stapleton LM. Modeling clustered data with very few clusters. *Multivariate Behav Res*. 2016;51(4):495–518.

39. Li P, Redden DT. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Stat Med*. 2015; 34(2):281–296.

40. Fay MP, Graubard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*. 2001; 57(4):1198–1206.

41. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics*. 2001;57(1):126–134.

42. Morel J, Bokossa M, Neerchal N. Small sample correction for the variance of GEE estimators. *Biom J*. 2003;45(4):395–409.

43. Preisser JS, Lu B, Qaqish BF. Finite sample adjustments in estimating equations and covariance estimators for intracluster correlations. *Stat Med*. 2008; 27(27):5764–5785.

44. Pan W, Wall MM. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Stat Med*. 2002;21(10): 1429–1441.

45. Carey V, Zeger SL, Diggle P. Modelling multivariate binary data with alternating logistic regressions. *Biometrika*. 1993;80(3):517–526.

46. By K, Qaqish BF, Preisser JS, Perin J, Zink RC. ORTH: R and SAS software for regression models of correlated binary data based on orthogonalized residuals and alternating logistic regressions. *Comput Methods Programs Biomed*. 2014;113(2): 557–568.

47. Halekoh U, Højsgaard S, Yan J. The R package geepack for generalized estimating equations. *J Stat Softw*. 2006;15(2):1–11.

48. Crespi CM, Wong WK, Mishra SI. Using second-order generalized estimating equations to model heterogeneous intraclass correlation in cluster-randomized trials. *Stat Med*. 2009;28(5): 814–827.

49. 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.

50. Williamson JM, Datta S, Satten GA. Marginal analyses of clustered data when cluster size is informative. *Biometrics*. 2003; 59(1):36–42.

51. Hoffman EB, Sen PK, Weinberg CR. Within-cluster resampling. *Biometrika*. 2001;88(4):1121–1134.

52. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int Stat Rev*. 1991;59(1):25–35.

53. Hin L-Y, Carey VJ, Wang Y-G. Criteria for working-correlation-structure selection in GEE: assessment via simulation. *Am Stat*. 2007;61(4):360–364.

54. Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Stat Med*. 2008;27(23): 4658–4677.

55. Stephens AJ, Tchetgen Tchetgen EJ, Gruttola VD. 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(10):915–930.

56. Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. *Int J Epidemiol*. 2013;42(5):1511–1519.

57. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc*. 1995;90(429):106–121.

58. Robins JM, Greenland S, Hu F-C. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *J Am Stat Assoc*. 1999;94(447):687–700.

59. Miglioretti DL, Heagerty PJ. Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics*. 2004; 5(3):381–398.

60. Song PXK, Jiang Z, Park E, Qu A. Quadratic inference functions in marginal models for longitudinal data. *Stat Med*. 2009;28(29):3683–3696.

61. Khajeh-Kazemi R, Golestan B, Mohammad K, Mahmoudi M, Nedjat S, Pakravan M. Comparison of generalized estimating equations and quadratic inference functions in superior versus inferior Ahmed glaucoma valve implantation. *J Res Med Sci*. 2011;16(3): 235–244.

62. Westgate PM, Braun TM. The effect of cluster size imbalance and covariates on the estimation performance of quadratic inference functions. *Stat Med*. 2012; 31(20):2209–2222.

63. Westgate PM. A bias-corrected covariance estimate for improved inference with quadratic inference functions. *Stat Med*. 2012;31(29):4003–4022.

64. Westgate PM. A covariance correction that accounts for correlation estimation to improve finite-sample inference with generalized estimating equations: a study on its applicability with structured correlation matrices. *J Stat Comput Simul*. 2016;86(10):1891–1900.

65. Westgate PM. Criterion for the simultaneous selection of a working correlation structure and either generalized estimating equations or the quadratic inference function approach. *Biom J*. 2014; 56(3):461–476.

66. Asgari F, Biglarian A, Seifi B, Bakhshi A, Miri HH, Bakhshi E. Using quadratic inference functions to determine the factors associated with obesity: findings from the STEPS Survey in Iran. *Ann Epidemiol*. 2013;23(9):534–538.

67. Bakhshi E, Etemad K, Seifi B, Mohammad K, Biglarian A, Koohpayehzadeh J. Changes in obesity odds ratio among Iranian adults since 2000: quadratic inference functions method. *Comput Math Methods Med*. 2016;5: 1–7.

68. Yang K, Tao L, Mahara G, et al. An association of platelet indices with blood pressure in Beijing adults: applying quadratic inference function for a longitudinal study. *Medicine (Baltimore)*. 2016;95(39):e4964.

69. van der Laan MJ, Robins JM. *Unified Methods for Censored Longitudinal Data and Causality*. New York, NY: Springer; 2003.

70. Gruber S, van der Laan MJ. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *Int J Biostat*. 2010;6(1):1–18.

71. Kotwani P, Balzer L, Kwarisiima D, et al. Evaluating linkage to care for hypertension after community-based screening in rural Uganda. *Trop Med Int Health*. 2014;19(4):459–468.

72. Ahern J, Karasek D, Luedtke AR, Bruckner TA, van der Laan MJ. Racial/ethnic differences in the role of childhood adversities for mental disorders among a nationally representative sample of adolescents. *Epidemiology*. 2016; 27(5):697–704.

73. Balzer LB, Petersen ML, van der Laan MJ. Targeted estimation and inference for the sample average treatment effect in trials with and without pair-matching. *Stat Med*. 2016;35(21):3717–3732.

74. Schnitzer ME, van der Laan MJ, Moodie EE, Platt RW. Effect of breastfeeding on gastrointestinal infection in infants: a targeted maximum likelihood approach for clustered longitudinal data. *Ann Appl Stat*. 2014;8(2): 703–725.

75. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6:1.

76. Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomization-based inference for community intervention trials. *Stat Med*. 1996;15(11): 1069–1092.

77. 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.

78. 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.

79. Davey C, Hargreaves J, Thompson JA, et al. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials*. 2015;16(1): 358.

80. 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.

81. 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.

82. Harling G, Wang R, Onnela J, De Gruttola V. Leveraging contact network structure in the design of cluster randomized trials. *Clin Trials*. 2017;14(1): 37–47.

83. Ebola ça Suffit Ring Vaccination Trial Consortium. 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: h3740.

84. 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.

85. Banerjee A, Chandrasekhar AG, Duflo E, Jackson MO. The diffusion of microfinance. *Science*. 2013;341(6144): 1236498.

86. Ogburn EL, VanderWeele TJ. Causal diagrams for interference. *Stat Sci*. 2014; 29(4):559–578.

87. Vanderweele TJ, Tchetgen EJT, Halloran ME. Components of the indirect effect in vaccine trials: identification of contagion and infectiousness effects. *Epidemiology*. 2012;23(5):751.

88. Staples P, Prague M, Victor DG, Onnela J-P. Leveraging contact network information in clustered randomized trials of infectious processes. Available at: https://arxiv.org/pdf/1610.00039.pdf. Accessed February 26, 2017.

89. Teerenstra S, Moerbeek M, Melis RJ, Borm GF. A comparison of methods to analyse continuous data from pseudo cluster randomized trials. *Stat Med*. 2007; 26(22):4100–4115.

90. Baldwin SA, Bauer DJ, Stice E, Rohde P. Evaluating models for partially clustered designs. *Psychol Methods*. 2011;16(2): 149–165.

91. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics*. 1946;2(6): 110–114.

92. Pals S, 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.

93. Roberts C, Roberts SA. Design and analysis of clinical trials with clustering effects due to treatment. *Clin Trials*. 2005; 2(2):152–162.

94. Roberts C, Walwyn R. Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. *Stat Med*. 2013;32(1): 81–98.

95. 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.

96. Díaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clin Trials*. 2014;11(5): 590–600.

97. DeSouza CM, Legedza AT, Sankoh AJ. An overview of practical approaches for handling missing data in clinical trials. *J Biopharm Stat*. 2009;19(6): 1055–1073.

98. Hossain A, Diaz-Ordaz K, Bartlett JW. Missing continuous outcomes under covariate dependent missingness in cluster randomised trials. *Stat Methods Med Res*. 2016;Epub ahead of print.

99. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res*. 2013;22(3):278–295.

100. Vansteelandt S, Rotnitzky A, Robins J. Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika*. 2007;94(4):841–860.

101. Thabane L, Mbuagbaw L, Zhang S, et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Med Res Methodol*. 2013; 13(1):92.

102. Taljaard M, Donner A, Klar N. Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biom J*. 2008;50(3):329–345.

103. Ma J, Akhtar-Danesh N, Dolovich L, Thabane L. Imputation strategies for missing binary outcomes in cluster randomized trials. *BMC Med Res Methodol*. 2011;11(1):18.

104. Andridge RR. Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biom J*. 2011;53(1): 57–74.

105. Ma J, Raina P, Beyene J, Thabane L. Comparing the performance of different multiple imputation strategies for missing binary outcomes in cluster randomized trials: a simulation study. *J Open Access Med Stat*. 2012;2:93–103.

106. Caille A, Leyrat C, Giraudeau B. A comparison of imputation strategies in cluster randomized trials with missing binary outcomes. *Stat Methods Med Res*. 2016;25(6):2650–2669.

107. Seaman S, Galati J, Jackson D, Carlin J. What is meant by "missing at random"? *Stat Sci*. 2013;28(2):257–268.

108. Belitser SV, Martens EP, Pestman WR, Groenwold RH, Boer A, Klungel OH. Measuring balance and model selection in propensity score methods. *Pharmacoepidemiol Drug Saf*. 2011;20(11): 1115–1129.

109. Prague M, Wang R, De Gruttola V. *CRTgeeDR: An R Package for Doubly Robust Generalized Estimating Equations Estimations in Cluster Randomized Trials with Missing Data*. Cambridge, MA: Harvard University; 2016.

110. Prague M, Wang R, Stephens A, Tchetgen Tchetgen E, DeGruttola V. Accounting for interactions and complex inter-subject dependency in estimating treatment effect in cluster-randomized trials with missing outcomes. *Biometrics*. 2016;72(4):1066–1077.

111. Seaman SR, White IR, Copas AJ, Li L. Combining multiple imputation and inverse-probability weighting. *Biometrics*. 2012;68(1):129–137.

112. Hansen BB, Bowers J. Covariate balance in simple, stratified and clustered comparative studies. *Stat Sci*. 2008;23(2): 219–236.

113. Leyrat C, Caille A, Foucher Y, Giraudeau B. Propensity score to detect baseline imbalance in cluster randomized trials: the role of the c-statistic. *BMC Med Res Methodol*. 2016;16(1):9.

114. Leon AC, Demirtas H, Li C, Hedeker D. Subject-level matching for imbalance in cluster randomized trials with a small number of clusters. *Pharm Stat*. 2013;12(5):268–274.

115. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. *BMJ*. 2004;328(7441):702–708.

116. Hutton JL. Are distinctive ethical principles required for cluster randomized controlled trials? *Stat Med*. 2001;20(3): 473–488.

117. Taljaard M, Chaudhry SH, Brehaut JC, et al. Survey of consent practices in cluster randomized trials: improvements are needed in ethical conduct and reporting. *Clin Trials*. 2014;11(1): 60–69.

118. Sim J, Dawson A. Informed consent and cluster-randomized trials. *Am J Public Health*. 2012;102(3):480–485.

119. Weijer C, Grimshaw JM, Eccles MP, et al. The Ottawa statement on the ethical design and conduct of cluster randomized trials. *PLoS Med*. 2012;9(11): e1001346.

120. van der Graaf R, Koffijberg H, Grobbee DE, et al. The ethics of cluster-randomized trials requires further evaluation: a refinement of the Ottawa statement. *J Clin Epidemiol*. 2015;68(9): 1108–1114.

121. Zeng D, Lin D, Lin X. Semiparametric transformation models with random effects for clustered failure time data. *Stat Sin*. 2008;18(1):355–377.

122. Cai T, Cheng S, Wei L. Semiparametric mixed-effects models for clustered failure time data. *J Am Stat Assoc*. 2002;97(458):514–522.

123. Zhong Y, Cook RJ. Sample size and robust marginal methods for cluster-randomized trials with censored event times. *Stat Med*. 2015; 34(6):901–923.

124. Zhan Z, de Bock GH, Wiggers T, Heuvel E. The analysis of terminal endpoint events in stepped wedge designs. *Stat Med*. 2016;35(24): 4413–4426.

125. Xu Z. *Statistical Design and Survival Analysis in Cluster Randomized Trials* [dissertation]. Ann Arbor, MI: University of Michigan; 2011.

126. Kramer MS, Martin RM, Sterne JA, Shapiro S, Dahhou M, Platt RW. The double jeopardy of clustered measurement and cluster randomisation. *BMJ*. 2009;339:b2900.

127. Cho S-J, Preacher KJ. Measurement error correction formula for cluster-level group differences in cluster randomized and observational studies. *Educ Psychol Meas*. 2016;76(5):771–786.

128. Eldridge S, Ashby D, Bennett C, Wakelin M, Feder G. Internal and external validity of cluster randomised trials: systematic review of recent trials. *BMJ*. 2008;336(7649):876–880.

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

130. Ma J, Thabane L, Kaczorowski J, et al. Comparison of Bayesian and classical methods in the analysis of cluster randomized controlled trials with a binary outcome: the Community Hypertension Assessment Trial (CHAT). *BMC Med Res Methodol*. 2009; 9(1):37.

131. Grieve R, Nixon R, Thompson SG. Bayesian hierarchical models for cost-effectiveness analyses that use data from cluster randomized trials. *Med Decis Making*. 2010;30(2):163–175.

132. Clark AB, Bachmann MO. Bayesian methods of analysis for cluster randomized trials with count outcome data. *Stat Med*. 2010;29(2):199–209.

133. Gomes M, Ng ES, Grieve R, Nixon R, Carpenter J, Thompson SG. Developing appropriate methods for cost-effectiveness analysis of cluster randomized trials. *Med Decis Making*. 2012; 32(2):350–361.

134. Diaz-Ordaz K, Kenward M, Gomes M, Grieve R. Multiple imputation methods for bivariate outcomes in cluster randomised trials. *Stat Med*. 2016;35(20): 3482–3496.

135. Ng ES, Diaz-Ordaz K, Grieve R, Nixon RM, Thompson SG, Carpenter JR. Multilevel models for cost-effectiveness analyses that use cluster randomised trial data: an approach to model choice. *Stat Methods Med Res*. 2013; 25(5):2036–2052.

136. Diaz-Ordaz K, Kenward MG, Grieve R. Handling missing values in cost effectiveness analyses that use data from

cluster randomized trials. *J R Stat Soc Ser A Stat Soc*. 2014;177(2):457–474.

137. Hox JJ, Moerbeek M, Kluytmans A, van de Schoot R. Analyzing indirect effects in cluster randomized trials. The effect of estimation method, number of groups and group sizes on accuracy and power. *Front Psychol*. 2014;5:78.

138. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. *Annu Rev Psychol*. 2007;58:593–614.

139. Vanderweele TJ, Hong G, Jones SM, Brown JL. Mediation and spillover effects in group-randomized trials: a case study of the 4Rs educational intervention. *J Am Stat Assoc*. 2013;108(502):469–482.

140. VanderWeele TJ. A unification of mediation and interaction: a 4-way decomposition. *Epidemiology*. 2014;25(5):749–761.

141. Turner RM, White IR, Croudace T. Analysis of cluster randomized cross-over trial data: a comparison of methods. *Stat Med*. 2007;26(2):274–289.

142. Forbes AB, Akram M, Picher D, Cooper J, Bellomo R. Cluster randomised crossover trials with binary data and unbalanced cluster sizes: application to studies of near-universal interventions in intensive care. *Clin Trials*. 2015;12(1):34–44.

143. Morgan KE, Forbes AB, Keogh RH, Jairath V, Kahan BC. Choosing appropriate analysis methods for cluster randomised cross-over trials with a binary outcome. *Stat Med*. 2017;36(2):318–333.

144. Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McKenzie JE. Appropriate statistical methods were infrequently used in cluster-randomized crossover trials. *J Clin Epidemiol*. 2016;74:40–50.