



HAL
open science

Review of recent Methodological Developments in group-randomized trials: Part 2 - Analysis

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

► To cite this version:

Mélanie Prague, Elisabeth Turner, Gallis John, Li Fan, Murray David. Review of recent Methodological Developments in group-randomized trials: Part 2 - Analysis. *American Journal of Public Health*, 2017. hal-01579075

HAL Id: hal-01579075

<https://inria.hal.science/hal-01579075>

Submitted on 30 Aug 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



1 **REVIEW OF RECENT METHODOLOGICAL DEVELOPMENTS IN**

2 **GROUP-RANDOMIZED TRIALS: PART 2 - ANALYSIS**

3

4

5 **ABSTRACT**

6 In 2004, Murray et al. published a review of methodological developments in both the design
7 and analysis of group-randomized trials (GRTs). Over the last 13 years, there have been many
8 developments in both areas. The goal of the current paper is to review developments in analysis,
9 with a companion paper to focus on developments in design. As a pair, these papers update the
10 2004 review. This analysis paper includes developments in topics included in the earlier review,
11 such as methods for parallel-arm GRTs, inference for conditional and marginal effects, and new
12 topics including methods to account for multiple levels of clustering and alternative estimation
13 methods such as augmented GEE, targeted maximum likelihood and quadratic inference
14 functions. We also examine developments in dealing with missing outcome data, including
15 doubly robust approaches, software available for analysis, and analysis of alternative group
16 designs (including stepped wedge GRTs, network-randomized trials, pseudo-cluster randomized
17 trials and individually-randomized group treatment trials). These alternative designs, like the
18 parallel-arm GRT, require clustering to be accounted for in both their design and analysis.

19

20 **INTRODUCTION**

21 In a group-randomized trial (GRT), the unit of randomization is a group and outcome
22 measurements are obtained on members of those groups.¹ Also called a cluster-randomized trial
23 or community trial,²⁻⁵ a GRT is the best comparative design available if the intervention operates
24 at a group level, manipulates the physical or social environment, cannot be delivered to
25 individual members of the group without substantial risk of contamination, or under other
26 circumstances (e.g., a desire for herd immunity in studies of infectious disease).¹⁻⁵

27 In GRTs, outcomes on members of the same group are likely to be more similar to each other
28 than to outcomes on members from other groups.¹ Such clustering must be accounted for in the
29 design to avoid an under-powered study and in the analysis to avoid under-estimated standard
30 errors and inflated type I error for the intervention effect.¹⁻⁵ For analysis, regression modeling
31 approaches are generally preferred and most commonly used because of their ease of
32 implementation.⁶ Several textbooks now address these and other issues.¹⁻⁵
33 In 2004, Murray et al.⁷ published a review of methodological developments in both the design
34 and analysis of GRTs. In the 13 years since, there have been many developments in both areas.
35 The goal of the current paper is to focus on developments in analytic methods, including those
36 relevant to designs described in a companion paper that focuses on developments in GRT
37 design.⁸ As a pair, these papers update the 2004 review. With both papers, we seek to provide a
38 broad and comprehensive review to guide the reader to seek out appropriate materials for their
39 own circumstances.

40

41 **DEVELOPMENTS IN THE ANALYSIS OF PARALLEL GROUP-**

42 **RANDOMIZED TRIALS**

43 **Methods for Superiority, Equivalence, and Non-Inferiority**

44 In GRTs, superiority trials are more common than equivalence or non-inferiority trials: a
45 PubMed search by one of the authors (DMM) of studies published in 2015 identified 562
46 superiority GRTs but only 1 equivalence GRT and 2 non-inferiority GRTs. Similarly,
47 developments in the methods literature have focused on superiority GRTs, with developments
48 for equivalence and non-inferiority GRTs limited to small sections in two of the more recent

49 textbooks^{2,5} and a review paper on sample size methods.⁹ As a consequence, the current review
50 paper focuses on superiority GRTs.

51 **Methods for Intention-To-Treat and Alternative Intervention Effects**

52 In GRTs, protocol violations can lead to non-compliance at either the group- or member-level.⁵
53 In order to minimize bias, intention-to-treat (ITT) principles are recommended at both levels
54 rather than “on-treatment” and “per-protocol” analyses.^{2,4,5} While group-level protocol violations
55 are usually easy to identify, member-level compliance may be more difficult to ascertain in
56 practice.² Jo et al. demonstrate that analyses which ignore compliance information could be
57 underpowered to detect an ITT effect and propose a multilevel model combined with a mixture
58 model.¹⁰ Implications of group-level non-compliance can be considerable in GRTs, given the
59 small number of groups that are randomized in many GRTs.

60 **Methods Based on the Randomization Scheme**

61 Matching or stratification in the design has been recommended for some time as a way to ensure
62 baseline balance on important potential confounders,¹ with constrained randomization more
63 recently developed.¹¹ Recent reports suggest that most GRTs follow this advice.¹²⁻¹⁵ Matching
64 and stratification in the design can be ignored in the analysis of intervention effects, without
65 harm to the type I error rate, and often the saved degrees of freedom will improve power.^{16,17}
66 Recently, Donner et al. reported that ignoring matching can adversely affect other analyses, such
67 as analyses that examine the relationship between a risk factor and an outcome,¹⁸ for this reason,
68 investigators considering pair-matching should consider small strata instead (e.g., strata of 4). Li
69 et al.¹⁹ compared model-based and permutation methods in the context of constrained
70 randomization adjusting for group-level covariates. They found that both the adjusted F-test and

71 permutation test maintained the nominal size and had improved power under constrained
72 randomization compared to simple randomization.

73 **Model-Based Methods**

74 Model-based methods can be broadly classified according to the interpretation of the model
75 parameters. Conditional model parameters are typically estimated using mixed-effects regression
76 via maximum likelihood estimation (MLE) and are referred to as cluster-specific effects (or as
77 subject-specific effects in the longitudinal analysis literature). Effects are conditional on the
78 random effects used to account for clustering and on other covariates included in the analysis.

79 Conditional models are often recommended for studies focused on change within members or on
80 mediation analyses.⁷ Parameters of marginal models are usually estimated using generalized
81 estimating equations (GEE).^{20,21} They define the marginal expectation of the dependent variable
82 as a function of the independent variables and assume that the variance is a function of the mean;
83 they separately specify a working correlation structure for observations made on members of the
84 same group. Marginal models are often preferred for analyses of population-level effects because
85 the intervention effect coefficient is interpreted as a population-averaged effect. In practice,
86 marginal models are less frequently used than conditional models.⁶

87 Marginal and conditional intervention effects are equal for identity and log links²² and the
88 distinction between them is only important for link functions such as the logit for binary
89 outcomes. Although some authors have advocated for the log instead of logit link for binary
90 outcomes,²³ this approach is not widely used, possibly because of model convergence problems
91 for some data.^{24,25} Alternatively, a modified Poisson approach with log-link and robust standard
92 errors could be used in the GEE framework,²⁶ since it does not suffer from the same convergence

93 problems as the binomial model with log link,²⁷ but it may be less common because of the
94 familiarity of logistic regression among epidemiologists and biostatisticians.
95 In practice, the question about which of conditional or marginal effects are desired depends on
96 the research question. It is essential to understand the underlying assumptions of each method:
97 conditional models rely on correct specification of untestable aspects of the data distribution,
98 while marginal models rely on a correct definition of the population of interest, which can make
99 it difficult to generalize results to other populations.²⁸ We address each of the two approaches in
100 more detail below.

101 *Conditional Approaches*

102 If the mixed effects model used to estimate conditional effects is misspecified, the estimates are
103 difficult to interpret and, even if regression diagnostics can help,²⁹ standard errors (SEs) are not
104 robust. Fortunately, Murray et al.³⁰ and Fu³¹ have shown that mixed models are robust to
105 substantial violation of the normality assumptions for member- and group-level errors, so long as
106 balance is maintained at the group level. Parameter estimation by restricted maximum likelihood
107 estimation (REML) is preferred to MLE when few groups are available.³²⁻³⁴ For binary
108 outcomes, alternative methods for specifying the test degrees of freedom have been examined in
109 small sample GRTs and the between-within method is recommended.^{32,35}

110 *Multiple Levels of Clustering in Conditional Models.* GRTs may involve multiple levels of
111 clustering due to repeated measures on individuals or groups or additional hierarchical levels in
112 the design. Murray¹ distinguished between mixed-effects models based on the number of
113 measurements included in the analysis and recommended mixed-effects analysis of variance
114 (ANOVA) or covariance (ANCOVA), or mixed-effects repeated measures ANOVA/ANCOVA,
115 for analyses involving 1 or 2 measurements per person or per group; those models can account

116 for all sources of random variation in such data if they are properly specified.³⁶ However, that is
117 not the case in analyses involving 3 or more measurements per person or per group, where the
118 sources of random variation may be different; instead, such analyses require a random
119 coefficients model in which random trends and intercepts are calculated for each member (in
120 cohort GRT designs) and group (in cohort and cross-sectional GRT designs), average trends and
121 intercepts are calculated for each study arm, and the intervention effect is the net difference in
122 the average study-arm trends.³⁶ Trends are often estimated as linear slopes, but can take another
123 form.

124 *Variable Group Size in Conditional Models* Johnson et al. focused on the analysis of Gaussian
125 outcomes from GRTs with variable group size.³⁷ They compared ten model-based approaches
126 and found that a one-stage mixed model with Kenward-Roger³² degrees of freedom and
127 unconstrained variance components performed well for GRTs with 14 or more groups per study
128 arm. A two-stage model weighted by the inverse of the estimated theoretical variance of the
129 group means and with unconstrained variance components performed well for GRTs with 6 or
130 more groups per study arm. A number of other models resulted in an inflated type I error rate
131 when there was substantial variability in group size.

132 *Marginal Approaches*

133 When the GEE approach is used to estimate marginal effects, unbiased intervention effects can
134 be estimated even if the working correlation structure is incorrect (e.g. using robust SEs via the
135 sandwich estimator), although precision is increased if the working matrix is correct. Where
136 degrees of freedom are limited for the test of interest, as often happens in GRTs, SE estimation is
137 often biased downward and no method corrects for it in all cases, although several have been
138 proposed.³⁸⁻⁴⁴

139 *Multiple Levels of Clustering in Marginal Models.* While multilevel clustering is easy to account
140 for in mixed-effects regression, there is less literature for the GEE approach. The alternating
141 logistic regression approach⁴⁵ for binary and ordinal outcomes can be used to account for
142 correlation due to repeated measures on individuals within groups and can be implemented
143 within a GEE framework in both R (the `a1r` package) and SAS (PROC GEE).⁴⁶ The second-
144 order GEE approach which, in contrast to regular GEE, models the working correlation structure
145 as a function of covariates, can be implemented in R (`geepack` in R⁴⁷).⁴⁸ For more general
146 working correlation matrices, the user typically needs to perform additional programming in
147 order to provide the appropriate covariance matrix and convergence may not be achieved. In
148 addition, although the intervention effect is unbiased when the marginal model is not correctly
149 specified, the SEs estimated using GEE may be too small. To correct this, a robust sandwich
150 estimator of the variance can be used but such an approach leads to loss of power.⁴⁹ Because of
151 this accuracy-power trade-off, mixed-effects models may be a better option to deal with GRTs
152 involving more than two levels, although the effects estimated in such models are conditional
153 rather than marginal effects.

154 *Variable Group Size in Marginal Models.* Although GEE analysis can accommodate variable
155 group size, informative group size can negatively impact efficiency. In this case, Williamson et
156 al.⁵⁰ showed that GEE weighted by group size can correct bias in the estimated intervention
157 effect. This approach is equivalent and less computationally demanding than within-cluster
158 resampling.⁵¹

159 *Advanced GEE Approaches to Improve Efficiency.* For binary outcomes, GEE is more
160 conservative (i.e. the intervention effect will be estimated closer to the null) than mixed-effects
161 models.^{28,52} Moreover, the SE of the estimated intervention effect is also typically larger when

162 using GEE so that much recent effort has focused on efficient estimation. GEE is most efficient
163 when the true correlation structure of the data is chosen as the working correlation structure. Hin
164 et al. compared multiple selection criteria for the working correlation matrix.⁵³ An alternative
165 approach is augmented GEE (AU-GEE), a method developed for independent data using a causal
166 inference framework,⁵⁴ which has been extended to clustered data.⁵⁵ AU-GEE uses covariate
167 information to improve efficiency in a two-stage approach that specifies a model for the potential
168 outcomes under the treatment not received. AU-GEE is unbiased and robust to misspecification
169 of the potential outcome model, though correct specification improves efficiency. As for the
170 analysis of all trials, only baseline covariates should be included in AU-GEE for the analysis of
171 GRT data because adjustment for post-baseline covariates may lead to bias.⁵⁶ Alternative
172 methods are available to account for post-baseline, time-varying confounding.⁵⁷⁻⁵⁹

173 *Alternatives to GEE.* The quadratic inference function (QIF) method is an alternative to GEE for
174 the estimation of marginal effects. Song et al.⁶⁰ demonstrate that QIF has advantages over GEE:
175 it is more efficient and more robust to outliers; it has a goodness-of-fit test of the marginal mean
176 model and permits straightforward extensions to model selection. In large samples, QIF is more
177 efficient than GEE when the working correlation structure for the data is misspecified.⁶¹
178 However, the SEs may be under-estimated for small and medium sample size or for variable
179 group size.⁶² More recent work by Westgate^{63,64} provides improvements by using a bias-
180 corrected sandwich covariance estimate and by simultaneously selecting the QIF or GEE while
181 selecting the best working correlation structure.⁶⁵ Despite the many attractive properties of QIF,
182 at this time there are few applications in public health.⁶⁶⁻⁶⁸

183 A second alternative estimation method is targeted maximum likelihood estimation (tMLE).⁶⁹
184 tMLE is a maximum likelihood-based G-computation estimator that targets the fit of the data-

185 generating distribution to reduce bias in the parameter of interest. It is based on a machine
186 learning approach that fluctuates an initial estimate of the conditional mean outcome and
187 minimizes a loss function to provide an estimate of the parameter of interest.⁷⁰ The approach has
188 been used in public health^{71,72} and shows much promise for GRTs^{73,74} because it can improve
189 efficiency by simultaneously accounting for missing data and chance baseline covariate
190 imbalance without committing to a specific functional form.⁷⁵

191 **Permutation Methods**

192 Permutation analysis was introduced for GRTs by Gail et al. for the COMMIT trial.⁷⁶ They
193 found that the permutation test had nominal type I and II error rates across a variety of settings
194 common to GRTs, when the member-level errors were Gaussian or binomial, even when very
195 few heterogeneous groups were randomized to each study arm, and even when the ICC was
196 large, so long as there was balance at the level of the group. Murray et al.³⁰ extended this work,
197 showing that unadjusted permutation tests offer no more protection against confounding than
198 unadjusted model-based tests, while the adjusted versions of both tests perform similarly. The
199 permutation test was more powerful than the model-based test when the data were binomial and
200 the $ICC \geq 0.01$. Fu³¹ extended the work to heavy tailed and very skewed distributions and
201 reported similar results.

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

207 **DEVELOPMENTS IN THE ANALYSIS OF ALTERNATIVES TO THE**
208 **PARALLEL GRT**

209 **Stepped Wedge GRT**

210 Both between- and within-group information is available to estimate the intervention effect from
211 a stepped wedge group randomized trial (SW-GRT).^{77,78} However, because the control condition
212 is typically observed earlier than the intervention condition, time is a potential confounder and
213 should be accommodated in the analysis of SW-GRTs, typically by accounting for time as a
214 predictor.⁷⁹ As for parallel GRTs, clustering by group must be accounted for, and longitudinal
215 measures on individuals can be accommodated within either the mixed-effects or GEE
216 framework, though more easily using mixed-effects models (see both *Multiple Levels of*
217 *Clustering* sections). Conditional approaches are more commonly used in practice and reported
218 on in the methods literature.^{79,80} Several authors have highlighted other characteristics specific to
219 SW-GRT including lagged intervention effects⁸¹ and fidelity loss over time.⁷⁹

220 **Network-Randomized GRT**

221 Because the network properties of a network-randomized GRT are primarily used at the design
222 stage,⁸² and because they differ from regular GRTs only in the novel way in which groups are
223 defined, the theory on the analysis of parallel-arm GRTs can be applied to parallel-arm network-
224 randomized GRTs.⁸³ For example, in a ring trial of an Ebola vaccine,⁸³ in which a network was
225 defined as all individuals who had regular physical contact with the incident (index) case of
226 Ebola and in which all contacts received the vaccine (placebo or active), standard GRT methods
227 were used. For network-randomized GRTs in which the intervention is not directly administered
228 to all individuals and in which it is expected that the intervention spreads over the network (e.g.
229 the snowball trials of a HIV prevention intervention for drug users⁸⁴ or a microfinance

230 intervention⁸⁵), methods^{86,87} are available to estimate both the direct and indirect effects of the
231 intervention. When network information is available and the outcome of interest is known to be a
232 disseminated process, adjusting for network features such as information on the location of each
233 individual within the network (i.e. group) can improve both the efficiency and power of the
234 analysis.⁸⁸

235 **Pseudo-Cluster Randomized Trial**

236 Teerenstra et al.⁸⁹ compared analytic methods for continuous outcomes in pseudo-cluster
237 randomized trials (PCRT) and Campbell and Walters discussed principles in their recent
238 textbook.⁵ Clustering by the unit of randomization at the first stage (e.g. provider) must be
239 accounted for in both the design and analysis of PCRT. No explicit sample size or analytic
240 methods are known to be available for non-continuous outcomes.

241 **Individually Randomized Group Treatment Trial**

242 Baldwin et al. compared four analytic models for IRGTs and three methods for calculating
243 degrees of freedom.⁹⁰ A multilevel model adapted to reflect clustering in only one study arm,
244 combined with either Satterthwaite⁹¹ or Kenward-Roger³² degrees of freedom, provided better
245 type I error control, better efficiency, and less bias, even with heteroscedasticity at the member
246 level. This finding is consistent with earlier reports by Pals et al.⁹² and Roberts et al.⁹³ More
247 recently, Roberts & Walwyn⁹⁴ and Andridge et al.⁹⁵ considered the circumstance in which
248 members are associated with more than one small group or change agent. Both found that
249 ignoring membership in multiple groups further inflates the type I error rate. Roberts & Walwyn
250 reported that multiple member multilevel models maintained the nominal type I error rate; they
251 also provide sample size and power formulae.⁹⁴

252 **DEVELOPMENTS TO ADDRESS DATA CHALLENGES**

253 **Missing Outcome Data**

254 Two recent reviews^{6,96} indicate that missing outcome data is common in GRTs, though
255 investigators frequently analyze only available data without accounting for the missing data
256 pattern. When the covariate-dependent missingness (CDM) assumption is plausible, both mixed
257 effects and GEE models provide unbiased estimates of the intervention effect when the CDM
258 covariates are included in an analysis of all available data.^{97,98} AU-GEE also can provide
259 unbiased effects by including all CDM covariates in the augmentation component⁵⁵ and has the
260 advantage that all estimates can still be interpreted as marginal effects. Other two-stage
261 approaches such as multiple imputation (MI) or inverse probability weighting (IPW) can provide
262 unbiased intervention effects under certain conditions for more general missing at random
263 (MAR) patterns and may provide increased precision compared to covariate-adjusted conditional
264 or marginal models for CDM.^{97,99} Although there is less literature on how to deal with missing
265 not-at-random (MNAR) data,¹⁰⁰ sensitivity analyses are recommended.¹⁰¹ A recent review
266 showed that very few GRTs performed any sensitivity analyses for their missing data
267 assumptions.⁶

268 To avoid possible type I error, MI should account for the clustered data structure.^{102,103} Fixed
269 group effects should not be used due to reduced power.¹⁰⁴ For binary outcomes, Ma et al.¹⁰⁵ and
270 Caille et al.¹⁰⁶ show that the preferred MI method depends on the number of groups and the
271 design effect, and note that bias may arise for some approaches even for CDM missingness.
272 Using group-specific mean imputation may be adequate for continuous outcomes.^{98,102} Hossain
273 et al.⁹⁸ show that if the missing data mechanism has an interaction between a covariate predictive
274 of the outcome and study arm, the imputation strategy must account for this interaction to be
275 unbiased.

276 Whereas MI requires specifying the distribution of the missing data conditional on covariates,
277 IPW requires specifying the probability of being missing depending on covariates. Theoretically,
278 both approaches can be used for any type of outcome and for both CDM and more general forms
279 of MAR mechanisms.⁹⁹ While IPW requires an additional assumption of positivity (all
280 participants have a non-zero probability of being observed), it may be viewed as easier to define,
281 particularly in the presence of non-intermittent missingness.¹⁰⁷ Importantly, and as for MI, if the
282 missing data mechanism has an interaction between a covariate predictive of the outcome and
283 study arm, the weights must be generated by accounting for this interaction in order to be
284 unbiased.¹⁰⁸ Prague et al.^{109,110} developed a doubly robust estimator in the context of IPW, which
285 provides an unbiased estimate if either the marginal mean model or the missing data model is
286 correctly specified. They demonstrated that a doubly-robust augmented GEE approach can
287 simultaneously account for both CDM and baseline covariate imbalance in GRTs when the
288 parameter of interest is a marginal effect. Combining MI and IPW is a promising new approach
289 which may have superior performance to IPW or MI alone when there are missing covariates in
290 addition to missing outcomes.¹¹¹

291 **Baseline Imbalance of Covariates**

292 While design strategies such as restricted randomization⁸ can help to achieve baseline covariate
293 balance, they may not be easy to implement (e.g. if group characteristics are unknown in
294 advance) and chance imbalance may arise regardless. In this case, some form of model-based
295 covariate adjustment could be used such as standard multivariate regression for conditional
296 models or AU-GEE for marginal models.⁵⁵ The advantage of AU-GEE in this case is that it is
297 doubly robust in that the consistency of intervention effect estimate requires correct specification
298 of either the marginal mean structure or the treatment model, and it separates covariate

299 adjustment from intervention effect estimation thereby reducing the risk of choosing the
300 adjustment models to obtain the most significant results. The standard multivariate regression
301 adjustment approach does not enjoy either of these benefits.
302 Alternatively, Hansen and Bowers¹¹² proposed a balancing criterion and studied its
303 randomization distribution in order to simultaneously test for balance of multiple covariates in
304 both RCTs and GRTs. Leyrat et al.¹¹³ suggested to use the c-statistic of the propensity score
305 model to measure covariate balance at the individual level. Leon et al.¹¹⁴ recommended
306 propensity score matching to correct for baseline imbalance; in a simulation study, they report a
307 median 90% reduction in bias. Nevertheless, the Consolidated Standards for Reporting of Trials
308 (CONSORT)¹¹⁵ recommends that the adjustment covariates be specified a priori for primary
309 analyses so that secondary analyses could test sensitivity of the primary findings to adjustment
310 for covariates identified post hoc.

311 **Software**

312 Table 1 identifies three software programs that can be used to analyze data from GRTs. The
313 table is organized around topics considered in the current paper. While none of the three software
314 programs can readily implement both QIF and tMLE for GRTs, the R program offers the most
315 ready-to-use functionality given its broad applicability to the methods cited in the current paper.

316 [TABLE 1 ABOUT HERE.]

317 **REPORTING OF RESULTS**

318 The CONSORT guidelines for individually randomized trials were extended to GRTs in 2004¹¹⁵
319 and most journals now require authors to conform to these guidelines. Based on a review of 300
320 GRTs published between 2000-2008, Ivers et al. reported that 60% and 70% accounted for
321 clustering in the sample size calculation and in the analysis, respectively, 56% used restricted

322 randomization, and most (86%) allocated more than 4 groups per arm.¹⁴ A more recent review
323 of 86 trials published in 2013-2014 showed that 77% and 78% accounted for clustering in the
324 sample size calculation and in the analysis, respectively, and that 51% used some form of
325 restricted randomization.¹⁵

326 Given concerns about the ethical conduct of GRTs,^{116,117} recent reports on conduct and reporting
327 have focused on the ethics of GRTs. For example, Sim and Dawson discuss the challenges
328 associated with obtaining informed consent in GRTs.¹¹⁸ The Ottawa Statement on the ethical
329 design and conduct of GRTs was published in 2012¹¹⁹ with a reevaluation in 2015.¹²⁰

330 **DISCUSSION**

331 In this review, we have summarized many of the most important advances in the analysis of
332 GRTs during the 13 years since the publication of the earlier review by Murray et al.⁷ Many of
333 these developments have focused on developments in marginal model parameter estimation (e.g.
334 augmented GEE, QIF and tMLE) and missing data methods. Some topics that space limitations
335 have prevented include review of recent developments in survival outcomes,^{2,121-125} measurement
336 bias,^{126,127} validity,^{128,129} Bayesian methods,^{4,130-132} cost-effectiveness analyses^{4,133-136} and
337 mediation analyses to uncover mechanisms of action.¹³⁷⁻¹⁴⁰

338 Through this review, we have sought to ensure that the reader is reminded of the value of well-
339 thought out analysis of GRTs and of keeping up to date with the many recent developments in
340 this area. Pairing this knowledge with our companion review of developments in the design of
341 GRTs,⁸ we hope that our review leads to continued improvements in the design and analysis of
342 GRTs.

343 **APPENDIX: GLOSSARY**

344 **Augmented GEE:** “Augmenting the standard GEE with a function of baseline covariates.”⁵⁵

345 These methods adapt semiparametric theory developed by Robins¹⁴¹ and Robins, Rotnitzky, and
346 Zhao¹⁴² for observational studies with time-varying exposures and missing data problems,
347 respectively. They consist of leveraging the estimating equation by a predictor function for
348 counterfactual outcomes under the intervention not received by the group/cluster considered
349 missing.⁵⁵

350 **Baseline covariate balance:** The group-level and individual-level covariate distributions are
351 similar in all study arms.¹¹

352 **Choice of balancing criterion:** Li et al. describe several balancing criteria to assess how well a
353 GRT is balanced across covariates. These include the “best balance” (BB) metric of de Hoop et
354 al.,¹⁴³ the balance criterion (B) of Raab and Butcher,¹¹ and the total balance score introduced by
355 Li et al.¹⁹

356 **Coefficient of variation:** A measure of between-group variation, defined in Table 1 of our
357 companion paper.⁸

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

360 **Constrained randomization:** Refers “to those designs that go beyond the basic design
361 constraints to specify classes of randomization outcomes that satisfy certain balancing criteria,
362 while retaining validity of the design.”¹⁴⁴

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

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

366 **Intraclass correlation:** A measure of between-group variation, defined in Table 1 of our
367 companion paper.⁸

368 **Covariate-dependent missingness (CDM) assumption:** The assumption that “missingness in
369 outcomes depends on covariates measured at baseline, but not on the outcome itself.”⁹⁸

370 **Doubly-robust augmented GEE approach:** Combining augmented GEE and IPW, a doubly-
371 robust estimator is obtained, which provides an unbiased estimate if either the marginal mean
372 model or the missing data model is correctly specified.^{109,110}

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

375 **G-computation estimator:** A computational method to estimate causal effect in structural
376 nested models. These models are designed to deal with confounding by variables affected by
377 intervention.¹⁴⁵

378 **Individually Randomized Group Treatment Trials:** “Studies that randomize individuals to
379 study arms but deliver treatments in small groups or through a common change agent.”^{8,92}

380 **Informative cluster size:** When the outcome measured is related to the size of the cluster.⁵⁰

381 **Missing at Random (MAR) assumption:** Rubin’s (1976) definition is that “data are missing at
382 random if for each possible value of the parameter ϕ [the parameter of the conditional
383 distribution of the missing data indicator given the data], the conditional probability of the
384 observed pattern of missing data, given the missing data and the value of the observed data, is
385 the same for all possible values of the missing data.”¹⁴⁶

386 **Network-Randomized GRT:** “The network-randomized GRT is a novel design that uses
387 network information to address the challenge of potential contamination in GRTs of infectious
388 diseases.”^{8,82,84,147}

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

391 **On treatment analyses:** When groups are analyzed “according to the intervention they actually
392 received.”²

393 **Per protocol analyses:** When groups “not receiving the correct intervention are excluded.”²

394 **Pseudo-cluster randomized trial:** Intervention is allocated to individuals in a two-stage
395 process. “In the first stage, providers are randomized to a patient allocation-mix.... In the
396 second stage, patients recruited to the PCRT are individually randomized to intervention or
397 control according to the allocation probability of their provider.”⁸

398 **Stepped Wedge GRT:** “A one-directional crossover GRT in which time is divided into intervals
399 and in which all groups eventually receive the intervention.”^{8,78}

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

404 **Within-cluster resampling:** Randomly sample one observation from each cluster, with
405 replacement. Then analyze this resampled dataset. Repeat this process a large number of times.
406 “The within-cluster resampling estimator is constructed as the average” of all of the resample-
407 based estimates (see Hoffman et al.⁵¹ pp. 1122-3).

408 **ACKNOWLEDGEMENTS**

409 Removed to avoid unblinding during the review process.

410 **CONTRIBUTORS**

411 Removed to avoid unblinding during the review process.

412 **HUMAN PARTICIPANT PROTECTION**

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

REFERENCES References

- 414
415
416 1. Murray DM. *Design and Analysis of Group-Randomized Trials*. New York, NY: Oxford University
417 Press; 1998.
418 2. Hayes RJ, Moulton LH. *Cluster Randomised Trials*. Boca Raton: CRC Press; 2009.
419 3. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*.
420 London: Arnold; 2000.
421 4. Eldridge S, Kerry S. *A Practical Guide to Cluster Randomised Trials in Health Services Research*.
422 Vol 120: John Wiley & Sons; 2012.
423 5. Campbell MJ, Walters SJ. *How to Design, Analyse and Report Cluster Randomised Trials in*
424 *Medicine and Health Related Research*. Chichester, West Sussex: John Wiley & Sons; 2014.
425 6. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster
426 randomized trials: a systematic review. *Trials*. 2016;17(1):72.
427 7. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of
428 recent methodological developments. *Am J Public Health*. 2004;94(3):423-432.
429 8. Turner EL, Li F, Gallis JA, Prague M, Murray DM. Review of Recent Methodological
430 Developments in Group-Randomized Trials: Part 1 - Design. *Am J Public Health*. Submitted.
431 9. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized
432 trials. *Int J Epidemiol*. 2015;44(3):1051-1067.
433 10. Jo B, Asparouhov T, Muthén BO. Intention-to-treat analysis in cluster randomized trials with
434 noncompliance. *Stat Med*. 2008;27(27):5565.
435 11. Raab GM, Butcher I. Balance in cluster randomized trials. *Stat Med*. 2001;20(3):351-365.
436 12. Varnell SP, Murray DM, Janega JB, Blitstein JL. Design and analysis of group-randomized trials: a
437 review of recent practices. *Am J Public Health*. 2004;94(3):393-399.
438 13. Murray DM, Pals SP, Blitstein JL, Alfano CM, Lehman J. Design and analysis of group-randomized
439 trials in cancer: a review of current practices. *J Natl Cancer Inst*. 2008;100(7):483-491.
440 14. Ivers NM, Halperin IJ, Barnsley J, et al. Allocation techniques for balance at baseline in cluster
441 randomized trials: a methodological review. *Trials*. 2012;13:120.
442 15. Fiero M, Huang S, Bell ML. Statistical analysis and handling of missing data in cluster randomised
443 trials: protocol for a systematic review. *BMJ Open*. 2015;5(5):e007378.
444 16. Diehr P, Martin DC, Koepsell T, Cheadle A. Breaking the matches in a paired t-test for community
445 interventions when the number of pairs is small. *Stat Med*. 1995;14(13):1491-1504.
446 17. Proschan MA. On the distribution of the unpaired t-statistic with paired data. *Stat Med*.
447 1996;15(10):1059-1063.
448 18. Donner A, Taljaard M, Klar N. The merits of breaking the matches: a cautionary tale. *Stat Med*.
449 2007;26(9):2036-2051.
450 19. Li F, Lokhnygina Y, Murray DM, Heagerty PJ, DeLong ER. An evaluation of constrained
451 randomization for the design and analysis of group-randomized trials. *Stat Med*.
452 2015;35(10):1565-1579.
453 20. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*.
454 1986;73(1):13-22.
455 21. Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*.
456 1986;42(1):121-130.
457 22. Ritz J, Spiegelman D. Equivalence of conditional and marginal regression models for clustered
458 and longitudinal data. *Stat Methods Med Res*. 2004;13(4):309-323.
459 23. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J*
460 *Epidemiol*. 1987;125(5):761-768.

- 461 24. Blizzard L, Hosmer W. Parameter Estimation and Goodness-of-Fit in Log Binomial Regression.
462 *Biom J.* 2006;48(1):5-22.
- 463 25. Williamson T, Eliasziw M, Fick GH. Log-binomial models: exploring failed convergence. *Emerging*
464 *themes in epidemiology.* 2013;10(1):1-10.
- 465 26. Zou G, Donner A. Extension of the modified Poisson regression model to prospective studies
466 with correlated binary data. *Stat Methods Med Res.* 2013;22(6):661-670.
- 467 27. Yelland LN, Salter AB, Ryan P. Performance of the modified Poisson regression approach for
468 estimating relative risks from clustered prospective data. *Am J Epidemiol.* 2011;174(8):984-992.
- 469 28. Hubbard AE, Ahern J, Fleischer NL, et al. To GEE or not to GEE: comparing population average
470 and mixed models for estimating the associations between neighborhood risk factors and
471 health. *Epidemiology.* 2010;21(4):467-474.
- 472 29. Huang X. Diagnosis of Random-Effect Model Misspecification in Generalized Linear Mixed
473 Models for Binary Response. *Biometrics.* 2009;65(2):361-368.
- 474 30. Murray DM, Hannan PJ, Varnell SP, McCowen RG, Baker WL, Blitstein JL. A comparison of
475 permutation and mixed-model regression methods for the analysis of simulated data in the
476 context of a group-randomized trial. *Stat Med.* 2006;25(3):375-388.
- 477 31. Fu D. *A comparison study of general linear mixed model and permutation tests in group-*
478 *randomized trials under non-normal error distributions* [Dissertation]. Memphis: Statistics,
479 University of Memphis; 2006.
- 480 32. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum
481 likelihood. *Biometrics.* 1997;53(3):983-997.
- 482 33. Localio AR, Berlin JA, Have TRT. Longitudinal and repeated cross-sectional cluster-randomization
483 designs using mixed effects regression for binary outcomes: bias and coverage of frequentist
484 and Bayesian methods. *Stat Med.* 2006;25(16):2720-2736.
- 485 34. Pinheiro JC, Bates DM. *Mixed-effects models in S and S-PLUS.* New York: Springer; 2000.
- 486 35. Li P, Redden DT. Comparing denominator degrees of freedom approximations for the
487 generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized
488 trials. *BMC Med Res Methodol.* 2015;15(1):38.
- 489 36. Murray DM, Hannan PJ, Wolfinger RD, Baker WL, Dwyer JH. Analysis of data from group-
490 randomized trials with repeat observations on the same groups. *Stat Med.* 1998;17(14):1581-
491 1600.
- 492 37. Johnson JL, Kreidler SM, Catellier DJ, Murray DM, Muller KE, Glueck DH. Recommendations for
493 choosing an analysis method that controls Type I error for unbalanced cluster sample designs
494 with Gaussian outcomes. *Stat Med.* 2015;34(27):3531-3545.
- 495 38. McNeish D, Stapleton LM. Modeling clustered data with very few clusters. *Multivariate Behav*
496 *Res.* 2016;51(4):495-518.
- 497 39. Li P, Redden DT. Small sample performance of bias-corrected sandwich estimators for cluster-
498 randomized trials with binary outcomes. *Stat Med.* 2015;34(2):281-296.
- 499 40. Fay MP, Graubard BI. Small-Sample Adjustments for Wald-Type Tests Using Sandwich
500 Estimators. *Biometrics.* 2001;57(4):1198-1206.
- 501 41. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties.
502 *Biometrics.* 2001;57(1):126-134.
- 503 42. Morel J, Bokossa M, Neerchal N. Small sample correction for the variance of GEE estimators.
504 *Biom J.* 2003;45(4):395-409.
- 505 43. Preisser JS, Lu B, Qaqish BF. Finite sample adjustments in estimating equations and covariance
506 estimators for intracluster correlations. *Stat Med.* 2008;27(27):5764-5785.
- 507 44. Pan W, Wall MM. Small-sample adjustments in using the sandwich variance estimator in
508 generalized estimating equations. *Stat Med.* 2002;21(10):1429-1441.

- 509 45. Carey V, Zeger SL, Diggle P. Modelling multivariate binary data with alternating logistic
510 regressions. *Biometrika*. 1993;80(3):517-526.
- 511 46. By K, Qaqish BF, Preisser JS, Perin J, Zink RC. ORTH: R and SAS software for regression models of
512 correlated binary data based on orthogonalized residuals and alternating logistic regressions.
513 *Comput Methods Programs Biomed*. 2014;113(2):557-568.
- 514 47. Halekoh U, Højsgaard S, Yan J. The R package geepack for generalized estimating equations.
515 *Journal of Statistical Software*. 2006;15(2):1-11.
- 516 48. Crespi CM, Wong WK, Mishra SI. Using second-order generalized estimating equations to model
517 heterogeneous intraclass correlation in cluster-randomized trials. *Stat Med*. 2009;28(5):814-827.
- 518 49. Teerenstra S, Lu B, Preisser JS, van Achterberg T, Borm GF. Sample size considerations for GEE
519 analyses of three-level cluster randomized trials. *Biometrics*. 2010;66(4):1230-1237.
- 520 50. Williamson JM, Datta S, Satten GA. Marginal analyses of clustered data when cluster size is
521 informative. *Biometrics*. 2003;59(1):36-42.
- 522 51. Hoffman EB, Sen PK, Weinberg CR. Within-cluster resampling. *Biometrika*. 2001;88(4):1121-
523 1134.
- 524 52. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-
525 averaged approaches for analyzing correlated binary data. *Int Stat Rev*. 1991;59(1):25-35.
- 526 53. Hin L-Y, Carey VJ, Wang Y-G. Criteria for working–correlation–structure selection in GEE:
527 Assessment via simulation. *Am Stat*. 2007;61(4):360-364.
- 528 54. Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment
529 comparisons in randomized clinical trials: A principled yet flexible approach. *Stat Med*.
530 2008;27(23):4658-4677.
- 531 55. Stephens AJ, Tchetgen Tchetgen EJ, Gruttola VD. Augmented generalized estimating equations
532 for improving efficiency and validity of estimation in cluster randomized trials by leveraging
533 cluster-level and individual-level covariates. *Stat Med*. 2012;31(10):915-930.
- 534 56. Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation
535 and bias. *Int J Epidemiol*. 2013;42(5):1511-1519.
- 536 57. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated
537 outcomes in the presence of missing data. *J Am Stat Assoc*. 1995;90(429):106-121.
- 538 58. Robins JM, Greenland S, Hu F-C. Estimation of the causal effect of a time-varying exposure on
539 the marginal mean of a repeated binary outcome. *J Am Stat Assoc*. 1999;94(447):687-700.
- 540 59. Miglioretti DL, Heagerty PJ. Marginal modeling of multilevel binary data with time-varying
541 covariates. *Biostatistics*. 2004;5(3):381-398.
- 542 60. Song PJK, Jiang Z, Park E, Qu A. Quadratic inference functions in marginal models for
543 longitudinal data. *Stat Med*. 2009;28(29):3683-3696.
- 544 61. Khajeh-Kazemi R, Golestan B, Mohammad K, Mahmoudi M, Nedjat S, Pakravan M. Comparison
545 of Generalized Estimating Equations and Quadratic Inference Functions in superior versus
546 inferior Ahmed Glaucoma Valve implantation. *J Res Med Sci*. 2011;16(3):235-244.
- 547 62. Westgate PM, Braun TM. The effect of cluster size imbalance and covariates on the estimation
548 performance of quadratic inference functions. *Stat Med*. 2012;31(20):2209-2222.
- 549 63. Westgate PM. A bias-corrected covariance estimate for improved inference with quadratic
550 inference functions. *Stat Med*. 2012;31(29):4003-4022.
- 551 64. Westgate PM. A covariance correction that accounts for correlation estimation to improve
552 finite-sample inference with generalized estimating equations: a study on its applicability with
553 structured correlation matrices. *J Stat Comput Simul*. 2016;86(10):1891-1900.
- 554 65. Westgate PM. Criterion for the simultaneous selection of a working correlation structure and
555 either generalized estimating equations or the quadratic inference function approach. *Biom J*.
556 2014;56(3):461-476.

- 557 66. Asgari F, Biglarian A, Seifi B, Bakhshi A, Miri HH, Bakhshi E. Using quadratic inference functions
558 to determine the factors associated with obesity: findings from the STEPS Survey in Iran. *Ann*
559 *Epidemiol.* 2013;23(9):534-538.
- 560 67. Bakhshi E, Etemad K, Seifi B, Mohammad K, Biglarian A, Koohpayehzadeh J. Changes in Obesity
561 Odds Ratio among Iranian Adults, since 2000: Quadratic Inference Functions Method. *Comput*
562 *Math Methods Med.* 2016;2016:1-7.
- 563 68. Yang K, Tao L, Mahara G, et al. An association of platelet indices with blood pressure in Beijing
564 adults: Applying quadratic inference function for a longitudinal study. *Medicine (Baltimore).*
565 2016;95(39):e4964.
- 566 69. Van der Laan MJ, Robins JM. *Unified methods for censored longitudinal data and causality.*
567 Springer Science & Business Media; 2003.
- 568 70. Gruber S, van der Laan MJ. A targeted maximum likelihood estimator of a causal effect on a
569 bounded continuous outcome. *Int J Biostat.* 2010;6(1):1-18.
- 570 71. Kotwani P, Balzer L, Kwarisiima D, et al. Evaluating linkage to care for hypertension after
571 community-based screening in rural Uganda. *Trop Med Int Health.* 2014;19(4):459-468.
- 572 72. Ahern J, Karasek D, Luedtke AR, Bruckner TA, van der Laan MJ. Racial/ethnic differences in the
573 role of childhood adversities for mental disorders among a nationally representative sample of
574 adolescents. *Epidemiology.* 2016;27(5):697-704.
- 575 73. Balzer LB, Petersen ML, van der Laan MJ. Targeted estimation and inference for the sample
576 average treatment effect in trials with and without pair-matching. *Stat Med.* 2016;35(21):3717-
577 3732.
- 578 74. Schnitzer ME, van der Laan MJ, Moodie EE, Platt RW. Effect of breastfeeding on gastrointestinal
579 infection in infants: a targeted maximum likelihood approach for clustered longitudinal data.
580 *Ann Appl Stat.* 2014;8(2):703-725.
- 581 75. Van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol.* 2007;6(1).
- 582 76. Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomization-
583 based inference for community intervention trials. *Stat Med.* 1996;15(11):1069-1092.
- 584 77. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised
585 trial: rationale, design, analysis, and reporting. *BMJ.* 2015;350:h391.
- 586 78. Spiegelman D. Evaluating public health interventions: 2. Stepping up to routine public health
587 evaluation with the stepped wedge design. *Am J Public Health.* 2016;106(3):453-457.
- 588 79. Davey C, Hargreaves J, Thompson JA, et al. Analysis and reporting of stepped wedge randomised
589 controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials.*
590 2015;16(1):358.
- 591 80. Mdege ND, Man M-S, Taylor CA, Torgerson DJ. Systematic review of stepped wedge cluster
592 randomized trials shows that design is particularly used to evaluate interventions during routine
593 implementation. *J Clin Epidemiol.* 2011;64(9):936-948.
- 594 81. Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G, Hargreaves JR. Designing a stepped wedge
595 trial: three main designs, carry-over effects and randomisation approaches. *Trials.*
596 2015;16(1):352.
- 597 82. Harling G, Wang R, Onnela J, De Gruttola V. Leveraging contact network structure in the design
598 of cluster randomized trials. *Clin Trials.* 2016 [Epub ahead of print].
- 599 83. Ebola ça Suffit Ring Vaccination Trial Consortium. The ring vaccination trial: a novel cluster
600 randomised controlled trial design to evaluate vaccine efficacy and effectiveness during
601 outbreaks, with special reference to Ebola. *BMJ.* 2015;351:h3740.
- 602 84. Latkin C, Donnell D, Liu TY, Davey-Rothwell M, Celentano D, Metzger D. The dynamic
603 relationship between social norms and behaviors: the results of an HIV prevention network
604 intervention for injection drug users. *Addiction.* 2013;108(5):934-943.

- 605 85. Banerjee A, Chandrasekhar AG, Duflo E, Jackson MO. The diffusion of microfinance. *Science*.
606 2013;341(6144).
- 607 86. Ogburn EL, VanderWeele TJ. Causal diagrams for interference. *Stat Sci*. 2014;29(4):559-578.
- 608 87. VanderWeele TJ, Tchetgen EJT, Halloran ME. Components of the indirect effect in vaccine trials:
609 identification of contagion and infectiousness effects. *Epidemiology*. 2012;23(5):751.
- 610 88. Staples P, Prague M, Victor DG, Onnela J-P. Leveraging Contact Network Information in
611 Clustered Randomized Trials of Infectious Processes. *arXiv preprint arXiv:1610.00039*. 2016.
- 612 89. Teerenstra S, Moerbeek M, Melis RJ, Borm GF. A comparison of methods to analyse continuous
613 data from pseudo cluster randomized trials. *Stat Med*. 2007;26(22):4100-4115.
- 614 90. Baldwin SA, Bauer DJ, Stice E, Rohde P. Evaluating models for partially clustered designs.
615 *Psychological Methods*. 2011;16(2):149-165.
- 616 91. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics*.
617 1946;2(6):110-114.
- 618 92. Pals SP, Murray DM, Alfano CM, Shadish WR, Hannan PJ, Baker WL. Individually randomized
619 group treatment trials: a critical appraisal of frequently used design and analytic approaches.
620 *Am J Public Health*. 2008;98(8):1418-1424.
- 621 93. Roberts C, Roberts SA. Design and analysis of clinical trials with clustering effects due to
622 treatment. *Clin Trials*. 2005;2(2):152-162.
- 623 94. Roberts C, Walwyn R. Design and analysis of non-pharmacological treatment trials with multiple
624 therapists per patient. *Stat Med*. 2013;32(1):81-98.
- 625 95. Andridge RR, Shoben AB, Muller KE, Murray DM. Analytic methods for individually randomized
626 group treatment trials and group-randomized trials when subjects belong to multiple groups.
627 *Stat Med*. 2014;33(13):2178-2190.
- 628 96. Díaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately
629 handled in cluster randomised trials? A systematic review and guidelines. *Clin Trials*.
630 2014;11(5):590-600.
- 631 97. DeSouza CM, Legedza AT, Sankoh AJ. An overview of practical approaches for handling missing
632 data in clinical trials. *J Biopharm Stat*. 2009;19(6):1055-1073.
- 633 98. Hossain A, Diaz-Ordaz K, Bartlett JW. Missing continuous outcomes under covariate dependent
634 missingness in cluster randomised trials. *Stat Methods Med Res*. 2016.
- 635 99. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat*
636 *Methods Med Res*. 2013;22(3):278-295.
- 637 100. Vansteelandt S, Rotnitzky A, Robins J. Estimation of regression models for the mean of repeated
638 outcomes under nonignorable nonmonotone nonresponse. *Biometrika*. 2007;94(4):841-860.
- 639 101. Thabane L, Mbuagbaw L, Zhang S, et al. A tutorial on sensitivity analyses in clinical trials: the
640 what, why, when and how. *BMC Med Res Methodol*. 2013;13(1):92.
- 641 102. Taljaard M, Donner A, Klar N. Imputation strategies for missing continuous outcomes in cluster
642 randomized trials. *Biom J*. 2008;50(3):329-345.
- 643 103. Ma J, Akhtar-Danesh N, Dolovich L, Thabane L. Imputation strategies for missing binary
644 outcomes in cluster randomized trials. *BMC Med Res Methodol*. 2011;11(1):18.
- 645 104. Andridge RR. Quantifying the impact of fixed effects modeling of clusters in multiple imputation
646 for cluster randomized trials. *Biom J*. 2011;53(1):57-74.
- 647 105. Ma J, Raina P, Beyene J, Thabane L. Comparing the performance of different multiple imputation
648 strategies for missing binary outcomes in cluster randomized trials: a simulation study. *J Open*
649 *Access Med Stat*. 2012;2:93-103.
- 650 106. Caille A, Leyrat C, Giraudeau B. A comparison of imputation strategies in cluster randomized
651 trials with missing binary outcomes. *Stat Methods Med Res*. 2016;25(6):2650-2669.

- 652 107. Seaman S, Galati J, Jackson D, Carlin J. What is meant by “missing at random”? *Stat Sci.*
653 2013;28(2):257-268.
- 654 108. Belitser SV, Martens EP, Pestman WR, Groenwold RH, Boer A, Klungel OH. Measuring balance
655 and model selection in propensity score methods. *Pharmacoepidemiol Drug Saf.*
656 2011;20(11):1115-1129.
- 657 109. Prague M, Wang R, De Gruttola V. CRTgeeDR: An R Package for Doubly Robust Generalized
658 Estimating Equations Estimations in Cluster Randomized Trials with Missing Data. *Harvard*
659 *University Biostatistics Working Paper Series*: Harvard University; 2016.
- 660 110. Prague M, Wang R, Stephens A, Tchetgen Tchetgen E, DeGruttola V. Accounting for interactions
661 and complex inter-subject dependency in estimating treatment effect in cluster-randomized
662 trials with missing outcomes. *Biometrics.* 2016;72(4):1066-1077.
- 663 111. Seaman SR, White IR, Copas AJ, Li L. Combining multiple imputation and inverse-probability
664 weighting. *Biometrics.* 2012;68(1):129-137.
- 665 112. Hansen BB, Bowers J. Covariate Balance in Simple, Stratified and Clustered Comparative Studies.
666 *Stat Sci.* 2008;23(2):219-236.
- 667 113. Leyrat C, Caille A, Foucher Y, Giraudeau B. Propensity score to detect baseline imbalance in
668 cluster randomized trials: the role of the c-statistic. *BMC Med Res Methodol.* 2016;16(1):9.
- 669 114. Leon AC, Demirtas H, Li C, Hedeker D. Subject-level matching for imbalance in cluster
670 randomized trials with a small number of clusters. *Pharm Stat.* 2013;12(5):268-274.
- 671 115. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised
672 trials. *Br Med J.* 2004;328(7441):702-708.
- 673 116. Hutton JL. Are distinctive ethical principles required for cluster randomized controlled trials?
674 *Stat Med.* 2001;20(3):473-488.
- 675 117. Taljaard M, Chaudhry SH, Brehaut JC, et al. Survey of consent practices in cluster randomized
676 trials: improvements are needed in ethical conduct and reporting. *Clin Trials.* 2014;11(1):60-69.
- 677 118. Sim J, Dawson A. Informed consent and cluster-randomized trials. *Am J Public Health.*
678 2012;102(3):480-485.
- 679 119. Weijer C, Grimshaw JM, Eccles MP, et al. The Ottawa statement on the ethical design and
680 conduct of cluster randomized trials. *PLoS Med.* 2012;9(11).
- 681 120. van der Graaf R, Koffijberg H, Grobbee DE, et al. The ethics of cluster-randomized trials requires
682 further evaluation: a refinement of the Ottawa Statement. *J Clin Epidemiol.* 2015;68(9):1108-
683 1114.
- 684 121. Zeng D, Lin D, Lin X. Semiparametric transformation models with random effects for clustered
685 failure time data. *Stat Sin.* 2008;18(1):355-377.
- 686 122. Cai T, Cheng S, Wei L. Semiparametric mixed-effects models for clustered failure time data. *J Am*
687 *Stat Assoc.* 2002;97(458):514-522.
- 688 123. Zhong Y, Cook RJ. Sample size and robust marginal methods for cluster-randomized trials with
689 censored event times. *Stat Med.* 2015;34(6):901-923.
- 690 124. Zhan Z, de Bock GH, Wiggers T, Heuvel E. The analysis of terminal endpoint events in stepped
691 wedge designs. *Stat Med.* 2016;35(24):4413-4426.
- 692 125. Xu Z. *Statistical Design and Survival Analysis in Cluster Randomized Trials* [Dissertation], The
693 University of Michigan; 2011.
- 694 126. Kramer MS, Martin RM, Sterne JA, Shapiro S, Dahhou M, Platt RW. The double jeopardy of
695 clustered measurement and cluster randomisation. *BMJ.* 2009;339.
- 696 127. Cho S-J, Preacher KJ. Measurement Error Correction Formula for Cluster-Level Group Differences
697 in Cluster Randomized and Observational Studies. *Educ Psychol Meas.* 2016;76(5):771-786.
- 698 128. Eldridge S, Ashby D, Bennett C, Wakelin M, Feder G. Internal and external validity of cluster
699 randomised trials: systematic review of recent trials. *BMJ.* 2008;336(7649):876-880.

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

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

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

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

709 133. Gomes M, Ng ES-W, Grieve R, Nixon R, Carpenter J, Thompson SG. Developing appropriate
710 methods for cost-effectiveness analysis of cluster randomized trials. *Med Decis Making*.
711 2012;32(2):350-361.

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

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

717 136. Díaz-Ordaz K, Kenward MG, Grieve R. Handling missing values in cost effectiveness analyses that
718 use data from cluster randomized trials. *J R Stat Soc Ser A Stat Soc*. 2014;177(2):457-474.

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

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

723 139. Vanderweele TJ, Hong G, Jones SM, Brown JL. Mediation and spillover effects in group-
724 randomized trials: a case study of the 4Rs educational intervention. *J Am Stat Assoc*.
725 2013;108(502):469-482.

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

728 141. Robins JM. Marginal structural models versus structural nested models as tools for causal
729 inference. In: Halloran ME, Berry DA, eds. *Statistical models in epidemiology, the environment
730 and clinical trials*. New York: Springer; 1999:pp. 95-134.

731 142. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are
732 not always observed. *J Am Stat Assoc*. 1994;89(427):846-866.

733 143. de Hoop E, Teerenstra S, van Gaal BG, Moerbeek M, Borm GF. The "best balance" allocation led
734 to optimal balance in cluster-controlled trials. *J Clin Epidemiol*. 2012;65(2):132-137.

735 144. Moulton LH. Covariate-based constrained randomization of group-randomized trials. *Clin Trials*.
736 2004;1(3):297-305.

737 145. Vansteelandt S, Joffe M. Structural nested models and g-estimation: The partially realized
738 promise. *Stat Sci*. 2014;29(4):707-731.

739 146. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592.

740 147. Staples PC, Ogburn EL, Onnela J-P. Incorporating Contact Network Structure in Cluster
741 Randomized Trials. *Sci Rep*. 2015;5:17581.

742

743

744 **Table 1. Summary of known functions and procedures to analyze GRTs using methods**
 745 **described in the current review.**

Method	SAS	Software Stata	R
Outcomes analysis of all available data			
Mixed-effects models	PROC MIXED PROC NL MIXED PROC GLIMMIX PROC GENMOD ¹	mixed melogit mepoisson xtgee	lme4 nlme geeglm/geeM
Generalized estimating equations (GEE)			
Targeted maximum likelihood (tMLE)	N/A	N/A	N/A ²
Quadratic inference function (QIF)	%qif	N/A	qif ³
Permutation tests	%ptest	N/A	N/A
Accounting for missing outcomes			
Multiple imputation for clustered data	%mmi_impute ⁴ %mmi_analyze	REALCOM Impute mi impute ⁴	pan jomo ⁵
Inverse probability weighting (IPW)	PROC GENMOD ⁶	N/A ⁷	CRTgeeDR
Causal-inference based methods⁸			
Augmented GEE (AU-GEE)	N/A	N/A	CRTgeeDR
Doubly robust AU-GEE	N/A	N/A	CRTgeeDR

Footnotes: 1. PROC GEE is another option, but is in experimental phase and has limited usefulness for GRTs over and above PROC GENMOD. 2. In R, tmlme is available for tMLE, but at the time of writing, does not allow for clustering. 3. As of the writing, the authors have been unable to load the package and it only allows equal cluster size, but Westgate has modified the code for GRTs with variable cluster size in the appendix of his paper⁶³ 4. Only useful for continuous outcomes. 5. In R, mice is available for multiple imputation but at the time of writing, does not account for clustering. 6. Cannot account for imprecision in the weights. 7. xtgee cannot accommodate individual-level weights but only group-specific weights. 8. Both of the 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. N/A: not available at the time of writing.