Design and Analytic Issues

David M. Murray, Ph.D.
Associate Director for Prevention
Director, Office of Disease Prevention
National Institutes of Health

NCCIH Workshop
August 8, 2016

National Institutes of Health
Office of Disease Prevention

Choosing the Design and Analytic Plan

- The delivery plan for the intervention shapes the design, which in turn shapes the analytic plan.
  - How, where, with whom, and from whom do participants receive their treatment or control intervention?
  - The answers often guide the choice of the unit of assignment and other features of the design, with implications for the analysis.

Three Kinds of Randomized Trials

- Individually Randomized Clinical Trials (RCTs)
  - Individuals randomized to study conditions with no connection among participants after randomization.
    - Most surgical and drug trials, some behavioral trials
- Individually Randomized Group Treatment Trials (IRGTs)
  - Individuals randomized to study conditions with some connection among participants after randomization.
    - Many behavioral trials with interventions delivered in small groups or through a common change agent.
- Group-Randomized Trials (GRTs)
  - Groups randomized to study conditions with some connection among participants before and after randomization.
    - Many trials with interventions delivered in communities, worksites, schools, etc.

Individually Randomized Group Treatment Trials

- K01AT005270
  - 6 groups of 10-12 women each, N=60 recruited
  - Groups and leaders are nested within conditions
- R34MH083866
  - 16 groups of 10 women each, stratified by site, N=160 recruited
  - Groups and leaders are nested within conditions; group format in intervention only
- R21AT007708
  - 4 groups of 15 parent/child dyads each, N=60 recruited
  - Groups and leaders are nested within conditions
- R34DA035946
  - 4 schools with 16 groups of six parent/child dyads each, N=64 recruited
  - Groups are nested within conditions; schools and leaders are crossed with conditions

Group Randomized Trials

- R34DA029237
  - Six schools randomized, 3 to I and 3 to C
  - Each school has 2 groups of 20-25 students, N=266 recruited
  - Schools, groups and leaders are nested within conditions
- R34DA032756
  - 12 classes randomized, 6 to I and 6 to C, all from a single school
  - Each class has 20-25 students, N=283 recruited
  - Classes and leaders are nested within conditions
- R01DA0224764
  - 4 dormitories randomized 2 to I and 2 to C
  - Each dormitory has 5 groups of 12 youth, N=240 recruited
  - Dormitories, groups, and leaders are nested within conditions.

Impact on the Design

- Randomized clinical trials
  - There is usually good opportunity for randomization to distribute potential confounders evenly, as most RCTs have N>100.
  - If well executed, confounding is not usually a concern.
- Individually randomized group treatment trials
  - There may be less opportunity for randomization to distribute potential confounders evenly, as most IRGTs have N<100. Confounding can be more of a concern in IRGTs than in RCTs.
- Group-randomized trials
  - GRTs often involve a limited number of groups, often <50.
  - There may be limited opportunity for randomization to distribute potential confounders evenly.
  - Confounding is usually a concern in GRTs if G is <50.

Impact on the Analysis

- Observations on randomized individuals who do not interact are independent and are analyzed with standard methods.
- The members of the same group in a GRT will share some physical, geographic, social or other connection.
- The members of groups created for an IRGT will develop similar connections.
- Those connections will create a positive intraclass correlation that reflects extra variation attributable to the group.

$$\mathrm{ICC}_{m:g:c} = \mathrm{corr}\left(y_{i:k:l}, y_{i':k:l}\right)$$

Impact on the Analysis in a GRT or IRGT

- Given m members in each of g groups...
  - When group membership is established by random assignment,

$$\sigma^2_{\bar{y}_g} = \frac{\sigma^2_y}{m}$$

  - When group membership is not established by random assignment, or if members develop connections post-randomization.

$$\sigma^2_{\bar{y}_g} = \frac{\sigma^2_y}{m}\left(1 + (m-1)\,ICC\right)$$

Impact on the Analysis and Power

- The variance of any group-level statistic will be larger.
- The df to estimate the ICC will be based on the number of groups, and so is often limited.
  - This is almost always true in a GRT, can be true in an IRGT.
- Any analysis that ignores the extra variation or the limited df will have a Type I error rate that is inflated, often badly.
  - Type I error rate may be 30-50% in a GRT, even with small ICC
  - Type I error rate may be 15-25% in an IRGT, even with small ICC
- Extra variation and limited df always reduce power.
- Nested factors must be modeled as random effects (Zucker, 1990), including groups and facilitators, if nested.

- Zucker DM. An analysis of variance pitfall: The fixed effects analysis in a nested design. Educational and Psychological Measurement. 1990;50(4):731-8.

The Warning

Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception, however, and should be discouraged.

Cornfield (1978)

- Though Cornfield's remarks were addressed to GRTs, his comments also apply to IRGTs.

- Cornfield J. Randomization by group: a formal analysis. American Journal of Epidemiology. 1978;108(2):100-2.

Why Have Connections Among Participants?

- The intervention may operate at a group level, or rely on group processes.
- The intervention may manipulate the social or physical environment.
- It may be less expensive, or logistically simpler, to deliver components of the intervention to groups rather than individuals, or through a common change agent.
- It may not be possible to deliver the interventions to selected individuals without substantial risk of contamination to others in the same group.

- Avoid connections if possible, and plan for them if not.

Recommendations for Design

- A GRT remains the best comparative design available when the investigator wants to evaluate an intervention that…
  - operates at a group level
  - manipulates the social or physical environment
  - cannot be delivered to individuals without contamination
- An IRGT is the best comparative design when...
  - Individual randomization is possible without contamination
  - There are good reasons to deliver the intervention in groups
- Alternatives discussed in a few minutes…

Recommendations for Analysis

- GRTs and IRGTs require analyses that reflect the nested designs inherent in these studies.
- Used alone, the usual methods based on the General or Generalized Linear Model are invalid, and investigators risk overstating the significance of any effects.
- Methods based on the General Linear Mixed Model and on the Generalized Linear Mixed Model are widely applicable.
  - For designs having one or two time intervals, mixed-model ANOVA/ANCOVA is recommended.
  - For designs having three or more time intervals, random coefficients models are recommended.
- Other methods can be used effectively, with proper care, including randomization tests, GEE, and two-stage methods.

Recommendations for Analysis

- Other approaches may not be appropriate, including analysis at a subgroup level and ignoring the unit of assignment if the ICC is not significant.
- Even with an otherwise strong design and analytic plan, unbalanced designs can create analytic problems and an inflated Type I error rate.
  - Balance at the group level is most important.
  - Imbalance at the member level beyond 2:1 requires special analytic methods (Johnson et al., 2015).
- For pilot studies, the requirements can be relaxed, but investigators risk overstating the significance of any effects.

- Johnson JL et al. Recommendations for choosing an analysis method that controls Type I error for unbalanced cluster sample designs with Gaussian outcomes. Statistics in Medicine. 2015;34(27):3531-45.

Power in GRTs and IRGTs

- Power in a GRT is driven by the number of groups and ICC.
  - For ES of 0.25, 10-12 groups per condition are typically required for ICCs found in public health and medicine.
- Power in an IRGT is driven by several factors.
  - If groups exist in all conditions, the issues are the same as for a GRT.
  - If a participant belongs to more than one small group, or if the groups change over time, the impact is greater.
  - If there are small groups in only one condition, that should be accounted for in the power calculations and the analysis, but the impact is less.
  - If the facilitators are crossed with conditions, the impact is less.
  - If the group interaction is limited, the impact is less.

Power in GRTs and IRGTs

- Investigators are encouraged to work with a methodologist familiar with these issues.
- For efficacy and effectiveness trials, it is important that these issues be considered fully in planning the trial.
- For pilot studies, the requirements can be relaxed, but the studies risk being underpowered for a valid analysis.

What About Alternative Designs?

- Many alternatives to GRTs have been proposed.
  - Multiple baseline designs
  - Time series designs
  - Quasi-experimental designs
  - Stepped wedge designs
  - Regression discontinuity designs
- Murray et al. (2010) compared these alternatives to GRTs for power and cost in terms of sample size and time.

- Murray DM, Pennell M, Rhoda D, Hade EM, Paskett ED. Designing studies that would address the multilayered nature of health care. Journal of the National Cancer Institute Monographs. 2010(40):90-6. PMC3482955.
- See also Shadish WR, Cook TD, Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston, MA: Houghton Mifflin Company; 2002.

Multiple Baseline Designs

- Evaluation relies on logic rather than statistical evidence.
  - Replication of the pattern in each group, coupled with the absence of such changes otherwise, is taken as evidence of an intervention effect.
  - With just a few groups, there is little power for a valid analysis.
- Good choice if effects are expected to be large and rapid.
- Poor choice if effects are expected to be small or gradual.
- Very poor choice if the intervention effect is expected to be inconsistent across groups.

- Rhoda DA, Murray DM, Andridge RR, Pennell ML, Hade EM. Studies with staggered starts: multiple baseline designs and group-randomized trials. American Journal of Public Health. 2011;101(11):2164-9. PMC3222403.

Quasi-Experimental Designs

- QEs have all the features of experiments except randomization.
  - Causal inference requires elimination of plausible alternatives.
  - If groups are assigned and members are observed, analysis and power issues are the same as in GRTs.
  - If participants receive at least part of the intervention in small groups, analysis and power issues are the same as in IRGTs.
- Useful when randomization is not possible.
  - Can provide experience with recruitment, measurement, intervention.
  - Can provide evidence of treatment effects if executed properly.
- Well-designed and analyzed QEs are usually more difficult and more expensive than well-designed and analyzed GRTs and IRGTs.

Regression Discontinuity Design

- Individuals assigned to conditions based on a score, often reflecting need (Shadish et al., 2002).
- The analysis models the relationship between the assignment variable and the outcome.
- Because assignment is fully explained by the assignment variable, proper modeling supports causal inference.
- RDs avoid randomization, but are as valid as randomized trials.
- RDs are less efficient than the standard RCT or GRT, often requiring twice as many participants.
- RDs can be with groups or members (Pennell, et al., 2011).

- Pennell ML, Hade EM, Murray DM, Rhoda DA. Cutoff designs for community-based intervention studies. Statistics in Medicine. 2011;30(15):1865-82. PMC3127461.

Summary

- A GRT remains the best design available whenever the investigator wants to evaluate an intervention that…
  - operates at a group level
  - manipulates the social or physical environment
  - cannot be delivered to individuals
- GRTs provide better or equal quality evidence and are either more efficient or take less time than the alternatives.
- Even so, GRTs are more challenging than the usual RCT.
- IRGTs present many of the same issues found in GRTs.
- Investigators new to GRTs and IRGTs should collaborate with more experienced colleagues, especially experienced methodologists.

Summary

- Many alternatives to GRTs have been proposed.
  - Multiple baseline designs
  - Time series designs
  - Quasi-experimental designs
  - Stepped wedge designs
  - Regression discontinuity designs
- Under the right conditions, these alternatives can provide good evidence for causal inference.
  - Some rely on logic more than statistical evidence.
    - Multiple baseline designs, time-series designs
  - Others require studies as large or larger than GRTs and may take longer to complete
    - Quasi-experimental designs, regression discontinuity, stepped wedge

Important References on GRTs, IRGTs

- Murray DM. Design and Analysis of Group-Randomized Trials. New York, NY: Oxford University Press; 1998.
- Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. London: Arnold; 2000.
- Hayes RJ, Moulton LH. Cluster Randomised Trials. Boca Raton, FL: Taylor & Francis Group, LLC; 2009.
- Campbell MJ, Walters SJ. How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research. Chichester: John Wiley & Sons Ltd.; 2014.
- Pals SP, Murray DM, Alfano CM, Shadish WR, Hannan PJ, Baker WL. Individually randomized group treatment trials: a critical appraisal of frequently used design and analytic approaches. American Journal of Public Health. 2008;98(8):1418-24.
- Baldwin SA, Bauer DJ, Stice E, Rohde P. Evaluating models for partially clustered designs. Psychological Methods. 2011;16(2):149-65.