U.S. FOOD & DRUG ADMINISTRATION

Statistical Considerations

Edward Bein

Outline

Key concepts for cluster randomized trials (CRTs)

- Sampling/randomization for CRTs, between-cluster variation, and intracluster correlation
- Cons and pros of CRTs
- Using cluster-level summaries vs. subject-level endpoint values in efficacy analyses
- Matched-pair vs. stratified vs. complete randomization

Tip of the cap: Hayes & Moulton (2017) & Donner & Klar (2000).

A Simple CRT Sampling/Randomization Scheme

- There is a large **population of clusters**, typically of different sizes (i.e., different numbers of individuals belong to them).
- The **population of individuals** includes all individuals who belong to a cluster.
- Select a random sample of *N* clusters from the population of clusters.
- Each of the selected clusters is randomly assigned to treatment 1 or treatment 0.
- Observe the endpoint values (e.g., binary *infected: Y or N*) of all individuals who belong to a selected cluster.

Between-cluster Variance

- Consider the N_1 clusters randomly assigned to treatment 1. Typically these clusters have variable true clinical success rates π_{1i}.
- This variability is due to differences in cluster-level characteristics.
- We conceptualize the π_{1i} as belonging to a population of cluster-specific treatment 1 true success rates.
- The variance of this population is termed **between-cluster variance**.
- Ditto treatment 0…

Intracluster (aka Intraclass) Correlation

- Positive treatment-1 between-cluster variance implies that endpoint values from pairs of individuals belonging to the same treatment 1 cluster are positively correlated. This correlation is termed the **intracluster correlation** (ICC), denoted ρ_1.
- Endpoint values from pairs of individuals from different treatment 1 clusters are independent.
- Ditto treatment 0 and ρ_0.

CRT Cons

- Standard statistical methods (e.g., t-tests, chi square tests) assume independent observations.
- When ρ_1 & ρ_0 are positive, this assumption is violated. This implies that applying standard methods to CRT individual-level data will yield overly optimistic p-values and overly-narrow CIs.
- Further, the effective sample size when valid non-standard analysis methods are used is smaller than the nominal sample size…

Example: Estimating/Testing the Risk Difference (RD)

- Let $\widehat{RD}$ be the usual estimator of the RD. $\widehat{RD}$ is unbiased (under certain conditions) in all three cases considered below. All three cases have 100 subjects/arm.
- Case 1: Individual-randomized trial, 100 subjects/arm.
- Case 2: CRT, both $\rho = .02$, 50 clusters/arm, 2 subjects/cluster. Then this case has the same statistical power to test $H_0: RD = 0$ that would be obtained from an individual-randomized trial with 98 subjects/arm. That is, 98 subjects/arm is the **effective sample size** for Case 2.
- Case 3: CRT, both $\rho = .02$, 10 clusters/arm, 10 subjects/cluster. Then this case has the same statistical power to test $H_0: RD = 0$ that would be obtained from an individual-randomized trial with 85 subjects/arm. Its **effective sample size** is 85 subjects/arm.
- Bottom line: statistical power for testing $H_0: RD = 0$ is greatest in Case 1, smallest in Case 3.

CRT Pros

- Only appropriate trial design when evaluating treatments intended to be administered cluster-wide.
- Intended to handle within-cluster contamination/interference between treatments. There is **contamination/interference between treatments** when patients' clinical outcomes are influenced by both the treatments they themselves receive and the treatments others receive.
  - ➢ treatments for diabetes: no interference.
  - ➢ vaccines/treatments for infectious diseases: interference.

Analyzing Cluster-level Summaries vs. Subject-level Endpoint Values

- Example of cluster-level summary: for each cluster in the trial, compute its infection rate, and then compare treatment 1 and treatment 0 clusters' rates using a t-test or nonparametric test.
- Example of subject-level endpoint: analyze all the individual binary infection outcomes using logistic regression GEE (generalized estimating equations) to compare treatments. This is a version of logistic regression appropriate for hierarchical data.

Cluster-level vs. individual-level treatment effects

- Each cluster has its own specific risk difference. The **cluster-level RD** is the mean cluster-specific RD over the population of clusters.
- Imagine that all individuals in the population of individuals receive treatment 1; call the resulting infection rate $rate_1$. Ditto for treatment 0 and $rate_0$. The **individual-level RD** equals the difference between $rate_1$ and $rate_0$.

Caveat emptor…

- In general,

  individual-level RD ≠ cluster-level RD

- In the slide 9 example of t-test of cluster-level infection rates, H_0 is: *cluster-level RD = 0*.
- In the slide 9 example of logistic regression GEE, H_0 is: *individual-level RD = 0*.
- **Bottom line**: the method of analysis should target the treatment effect at the level of clinical interest.
- Note 1: there are analysis methods for cluster-level summaries that target the individual-level RD & methods for individual outcomes that target the cluster-level RD.
- Note 2: ***individual-level RD = cluster-level RD*** when cluster-specific RD is uncorrelated with cluster size .

Types of CRT Designs

- Matched pairs: clusters are paired based on similarity on baseline characteristic(s) predictive of outcome and one member of the pair is randomized to treatment 1.
- Stratified: clusters are grouped into strata defined in terms of baseline characteristic(s) predictive of outcome and at least 2 of the clusters in each stratum are randomized to each arm.
- Completely randomized: no matching or stratification.
- Etc.: crossover design, stepped wedge design.

Some Considerations for Choosing Type of CRT Design

- Level of concern about between-arm imbalance on important cluster-level baseline covariates: concern is minimal when many clusters are included in the trial; otherwise, pair matching or stratification can improve balance.
- Hayes & Moulton (2017) generally recommend stratification over pair matching, both for CRTs with small and with large numbers of clusters.
- Covariate-adjusted analyses can adjust for between-arm imbalance and increase statistical power (but such analyses require care…).

References

Books

- Donner, A., & Klar, N. (2000). Design and Analysis of Cluster Randomization Trials in Health Research. Wiley.
- Hayes, R. J., & Moulton, L. H. (2017). Cluster Randomized Trials (2nd ed.). CRC Press.

Literature Reviews

- Turner et al. (2017). Review of recent methodological developments in group-randomized trials: Part 1 – design. American Journal of Public Health, 107, 907-915.
- Turner et al. (2017). Review of recent methodological developments in group-randomized trials: Part 2 – analysis. American Journal of Public Health, 107, 1078-1086.

References, continued

Recently developed statistical methods

- Benitez et al. (January 20, 2022). Defining and estimating effects in cluster randomized trials: A methods comparison. https://arxiv.org/abs/2110.09633

Extra: Toy Example of *individual-level RD ≠ cluster-level RD*

- Population of clusters: 10 “large” clusters (100 individuals each) & 10 “small” clusters (10 individuals each).
- Population of individuals: 1100 individuals, pooled across the 20 clusters.
- Treatment 0 infection rate is 50% in all clusters.
- Treatment 1 infection rate is 72% in small clusters, 30% in large clusters.
- Cluster-level RD (treatments 1 vs. 0): **1%**.
- Individual-level RD (treatments 1 vs. 0): **-16.2**%.