An Evaluation of Constrained Randomization for the Design and Analysis of Group-randomized Trials

Fan Li[1], Yuliya Lokhnygina[1], David M. Murray[2], Patrick J. Heagerty[3] and Elizabeth R. DeLong[1]

[1]Duke University [2]National Institutes of Health, Office of Disease Prevention [3]University of Washington

Correspondence to: *frank.li@duke.edu* or *elizabeth.delong@dm.duke.edu*

Background

Covariate-based constrained randomization (CR) was proposed as an allocation technique for achieving baseline balance in group-randomized trials (GRTs) [1]. Briefly, this methodology includes:

1. specifying the important potentially confounding factors;
2. characterizing each group in terms of these factors;
3. either enumerating all or simulating a large number of potential randomization schemes (removing duplicates if any);
4. selecting a candidate subset of schemes where sufficient balance across potentially confounding covariates is achieved according to some pre-specified balance metric;
5. randomly selecting one scheme out of this smaller candidate subset and the study is implemented using that scheme.

Given the relatively fewer discussions on the analysis issues of small GRTs under CR, we aim to evaluate the interplay between CR design and analysis of small GRTs, and compare with the results under simple randomization (SR).

Simulation Strategy

A linear mixed model (LMM) was used to generate Y_{ijk}, the outcome of subject k from group j in treatment i, by

$$Y_{ijk} = \boldsymbol{z}_{ijk}^T\boldsymbol{\alpha} + \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \gamma t_i + b_{ij} + \epsilon_{ijk}$$
$$i = 0, 1,\ j = 1, \ldots, g,\ k = 1, \ldots, 300.$$

Here $\boldsymbol{z}_{ijk}$ is the 4×1 subject-level covariates; $\boldsymbol{x}_{ij}$ is the 4×1 group-level covariates. Each subject-level covariate was generated from a normal distribution with group-specific means. Each group-level covariate was independently simulated from Bern(0.3). We held $\boldsymbol{\alpha} = \boldsymbol{\beta} = \mathbf{2}_{4\times 1}$. Denote t_i as a treatment indicator depending on the realized randomization scheme. Error $\epsilon_{ijk} \sim \text{N}(0,\ \sigma_\epsilon^2 = 4)$. The random effects $b_{ij} \sim \text{N}(\mu_b = 1,\ \sigma_b^2 = \rho\sigma_\epsilon^2/(1-\rho))$, where ρ is the ICC.

Design parameter I – Balance metric

Let $\bar{x}_{0l}, \bar{x}_{1l}$ denote the average of group-level variable means from two treatment arms, ω_l be the inverse of the variance of the group means. The imbalance score (B) is defined as [2]

$$\text{B} = \sum_{l=1}^{S} \omega_l (\bar{x}_{0l} - \bar{x}_{1l})^2.$$

We proposed a $l1$ metric, total balance (TB) score, expressed as the sum of maximum absolute differences, where maximum is taken over all levels for each group-level variable as:

$$\text{TB} = \sum_{l=1}^{S} \max_\tau |n_{0l\tau} - n_{1l\tau}|.$$

Here $n_{0l\tau}$, $n_{1l\tau}$ are the number of groups assigned to the two arms that have the τth level of the lth variable.

Design parameter II - Candidate set size (R) is defined as the pre-specified number of candidate randomization schemes with the smallest B (TB). Smallest R was set as 20.

We enumerated all possible schemes for $g = 5, 7$ and simulated 20000 schemes for $g = 9, 11, 13$.

Model- versus Randomization-based Tests

Two tests for treatment effects were considered:

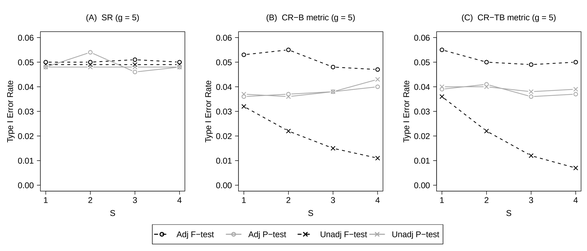
1. F-test – model-based; derived from the linear mixed model;
2. P-test – randomization based; derived from permuting the average residuals per group assuming a multiple regression model (Note: 0.05 level test does NOT exist when $R \leq 20$).

We considered both adjusted and unadjusted tests. The "adjustment" here is with respect to S $(0 \leq S \leq 4)$ of the group level potential confounders available prior to randomization. The subject-level covariates are collected as patients are recruited, often after group randomization, and they will be controlled for in all analyses. In summary, we compared the following:

		Design-based Adjustment	
		Yes	No
Analysis-based Adjustment	Yes	CR + Adjusted tests	SR + Adjusted tests
	No	CR + Unadjusted tests	SR + Unadjusted tests

The nominal significance level was chosen to be 0.05. All results were based on 10000 replicates.

Results I – Fixing R



We held $R = 100$, $\rho = 0.05$ and varied S. In terms of type I error,

1. Unadjusted F-test was NOT valid under CR;
2. P-test was slightly conservative when $R = 100$ (due to granularity of discrete p-values).

In terms of power,

1. No difference between F- and P-tests under SR;
2. Design-based adjustment improved power of the unadjusted P-test;
3. Power increased with more group-level covariates adjusted for by design, but analysis-based adjustment dominated the design-based adjustment.

Only minor differences with choice of balance metrics.

Results II– Varying R

Size ($S = 4$)	ICC	R	Unadj F	Unadj P	Adj F	Adj P
$g = 7$	CR	0.01	20	0.000	–	0.050	–
	CR	0.01	100	0.000	0.040	0.051	0.039
	CR	0.01	1000	0.001	0.050	0.048	0.047
	CR	0.01	2000	0.008	0.050	0.047	0.047
	CR	0.01	3000	0.032	0.052	0.049	0.051
	SR	0.01	–	0.051	0.046	0.049	0.047
	CR	0.1	20	0.001	–	0.050	–
	CR	0.1	100	0.001	0.038	0.051	0.037
	CR	0.1	1000	0.004	0.051	0.049	0.049
	CR	0.1	2000	0.012	0.048	0.053	0.049
	CR	0.1	3000	0.031	0.052	0.052	0.048
	SR	0.1	–	0.058	0.048	0.048	0.049

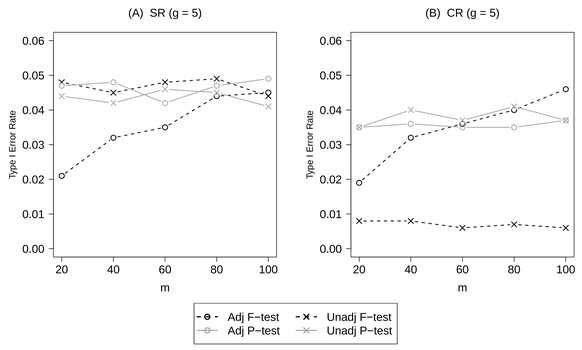
1. Unadjusted F – type I error vanished with smaller candidate set size;
2. ICC did NOT affect the test size.

Power ($S = 4$)	ICC	R	Unadj F	Unadj P	Adj F	Adj P
$g = 7$	CR	0.01	20	0.013	–	1.000	–
	CR	0.01	100	0.015	0.579	1.000	0.999
	CR	0.01	1000	0.040	0.270	1.000	0.999
	CR	0.01	2000	0.086	0.236	1.000	0.996
	CR	0.01	3000	0.123	0.170	0.999	0.980
	SR	0.01	–	0.148	0.143	0.996	0.946
	CR	0.1	20	0.022	–	0.638	–
	CR	0.1	100	0.025	0.371	0.631	0.544
	CR	0.1	1000	0.046	0.209	0.608	0.576
	CR	0.1	2000	0.084	0.196	0.574	0.536
	CR	0.1	3000	0.118	0.158	0.546	0.511
	SR	0.1	–	0.137	0.135	0.516	0.477

Except for the unadjusted F-test,

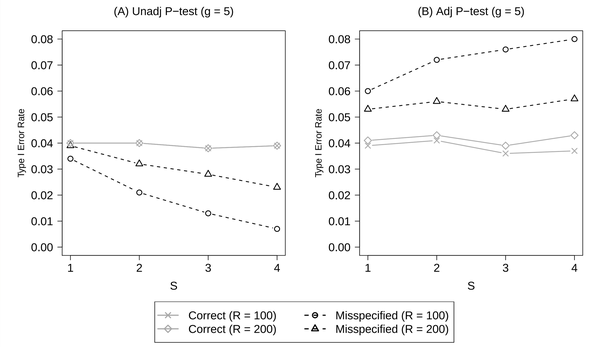
1. Smaller candidate set size corresponded to improved power;
2. Larger ICC reduced test power.

Cautionary Note I – Small Samples



Holding $R = 100$, $\rho = 0.05$ and varying the number of subjects per group (m), we observed CR (TB metric) wouldn't help much in extremely small sample scenarios since even the adjusted F-test did not carry the correct size.

Cautionary Note II – Misspecified P-test



A misspecified permutation test under CR calculates the permutational distribution with respect to the simple randomization space. Holding $\rho = 0.05$ and using the TB metric, we observed the unadjusted misspecified P-test became more conservative with increasing S and decreasing R (similar to the unadjusted F).

An important message – Both the unadjusted F- and the unadjusted misspecified P-tests did NOT account for the CR design, thus were invalid.

Conclusion

Design Aspect

- Under CR, choice of balance metric doesn't seem to make a substantial difference; smaller candidate set size could improve power with the correct analyses.
- Limitations of overly restricted randomization – nonexistence of P-test; violation of randomization ethics.

Analysis Aspect

- Under CR, the unadjusted F-test carries incorrect type I error since it fails to account for the design.
- Design-based adjustment for potential confounders could improve power (e.g., unadjusted P-test), but will be dominated by analysis-based adjustment.
- Model-based approach may suffer from convergence issues in small GRTs; randomization-based approach makes fewer modelling assumptions and is more flexible.
- The adjusted P-test is recommended under CR further due to its robustness under model misspecification; its permutational distribution should be calculated under the constrained randomization space under CR designs.

Acknowledgement – This work is supported by the NIH Common Fund through a cooperative agreement (U54 AT007748) with the NIH Health Care Systems Research Collaboratory.

References

[1] L.H. Moulton. Covariate-based constrained randomization of group-randomized trials. *Clinical Trials*, 1(3):297–305, May 2004.

[2] G M Raab and I Butcher. Balance in cluster randomized trials. *Statistics in Medicine*, 20(3):351–65, February 2001.