`cvcrand` and `cptest`: Efficient Design and Analysis of Cluster Randomized Trials

John Gallis

in collaboration with Fan Li, Hengshi Yu and Elizabeth L. Turner

Duke University Department of Biostatistics & Bioinformatics
and Duke Global Health Institute

July 28, 2017

Presentation Outline

1. Background: Cluster Randomized Trials
2. Design: Covariate Constrained Randomization
3. Analysis: Clustered Permutation Test
4. Conclusions and Future Directions in Research

1. Background

Context: Cluster randomized trials (CRTs)

- Also known as group-randomized trials
- Randomize "clusters" of individuals
  - e.g., communities, hospitals, etc.
- Rationale
  - Cluster-level intervention
  - Risk of contamination across intervention arms
- The most common type of CRT is the **two-arm parallel**
  - Randomize clusters to two intervention arms
  - Outcome data obtained on individuals

2. Design

Problem: Baseline covariate imbalance across arms

- CRTs often recruit relatively few clusters
  - Logistical/financial reasons
  - Most randomize $\leq$24 clusters (Fiero et al., 2016)
- Covariate imbalance problems
  - High probability of severe imbalances across intervention arms
- If these variables are predictive of the outcome, this may:
  - Threaten internal validity of the trial
  - Decrease power and precision of estimates
  - Complicate statistical adjustment
    - See Ivers et al. (2012)

Balance methods: Restricted randomization

Recent review: 56% of CRTs use some form of restricted randomization (Ivers et al., 2011, 2012)

- Matching
  - Limitation: If one cluster of a pair match drops out, then neither cluster can be used in primary analysis
- Stratification
  - Limitation: Should only have as many strata as up to $\frac{1}{2}$ the total # of clusters
  - Limitation: Can only stratify on categorized variables
- Covariate constrained randomization
  - Does not require categorization of continuous variables
  - Can accommodate a large number and a variety of types of variables

DukeHealth

Duke GLOBAL HEALTH INSTITUTE

Motivating example: Dickinson et al. (2015)

- **Policy question**: Improving up-to-date immunization rates in 19- to 35-month-old children
- Location: 16 counties in Colorado
- Two interventions
  - Practice-based
  - Community-based
- Desire to balance county-level variables potentially related to being up-to-date on immunizations

Motivating example: Dickinson et al. (2015)

- These county-level covariates include:
  - Location
  - Average income ($) categorized into tertiles
  - % In Colorado Immunization Information System
  - % Hispanic
  - Estimated % up-to-date on immunizations

Covariate constrained randomization: simple example

- Start with randomizing **four** counties to the two intervention arms
- Two important county-level covariates to balance on:

County	**Location**	**% In System**
1	Rural	90
2	Urban	92
3	Urban	80
4	Rural	75

- Note: For illustration only. Four clusters is not enough for valid statistics and inference!

All potential intervention arm assignments

There are $\binom{4}{2} = 6$ possible allocations for assigning 4 counties to two interventions (practice-based and community-based).

	County 1	County 2	County 3	County 4
Allocation 1	Practice	Practice	Community	Community
Allocation 2	Practice	Community	Practice	Community
Allocation 3	Practice	Community	Community	Practice
Allocation 4	Community	Practice	Practice	Community
Allocation 5	Community	Practice	Community	Practice
Allocation 6	Community	Community	Practice	Practice

All potential intervention arm assignments

We could also display the matrix as

	County 1	County 2	County 3	County 4
Allocation 1	1	1	0	0
Allocation 2	1	0	1	0
Allocation 3	1	0	0	1
Allocation 4	0	1	1	0
Allocation 5	0	1	0	1
Allocation 6	0	0	1	1

All potential intervention arm assignments

Under simple randomization: $\frac{1}{3}$ chance of obtaining intervention arm assignments completely imbalanced on location.

	County 1	County 2	County 3	County 4
Allocation 1	1	1	0	0
Allocation 2	1	0	1	0
Allocation 3	**1**	**0**	**0**	**1**
Allocation 4	**0**	**1**	**1**	**0**
Allocation 5	0	1	0	1
Allocation 6	0	0	1	1
Location	**Rural**	**Urban**	**Urban**	**Rural**
% In System	90	92	80	75

DukeHealth

Duke GLOBAL HEALTH INSTITUTE

Covariate constrained randomization: simple example

- Covariate constrained randomization method: **Define a balance score that decreases as balance improves**
  - Based on average differences in covariates between intervention arms weighted by inverse standard deviation and then summed
  - See Li et al. (2015) for technical details and theory

County 1	County 2	County 3	County 4	**Bscores**
1	1	0	0	**2.779**
1	0	1	0	**0.034**
1	0	0	1	**3.187**
0	1	1	0	**3.187**
0	1	0	1	**0.034**
0	0	1	1	**2.779**

DukeHealth

Duke GLOBAL HEALTH INSTITUTE

Covariate constrained randomization: simple example

Constraining the randomization below the 33rd percentile:

County 1	County 2	County 3	County 4	**Bscores**
1	1	0	0	**2.779**
1	0	1	0	**0.034**
1	0	0	1	**3.187**
0	1	1	0	**3.187**
0	1	0	1	**0.034**
0	0	1	1	**2.779**

Covariate constrained randomization: simple example

Constraining randomization below the 67th percentile:

County 1	County 2	County 3	County 4	**Bscores**
1	1	0	0	**2.779**
1	0	1	0	**0.034**
1	0	0	1	**3.187**
0	1	1	0	**3.187**
0	1	0	1	**0.034**
0	0	1	1	**2.779**

Introducing `cvcrand`

`cvcrand` for covariate constrained randomization

`cvcrand` *varlist* `,` `clusternum(#)` `treatmentnum(#)` `[`
    `clustername(`*varname*`)` `categorical(`*varlist*`)`
    `balancemetric(`*string*`)` `cutoff(#)` `numschemes(#)`
    `nosim` `size(#)` `weights(`*numlist*`)` `seed(#)`
    `savedata(`*string*`)` `savebscores(`*string*`)]`

This program is available to download using `ssc install cvcrand`

DukeHealth

Duke GLOBAL HEALTH INSTITUTE

Dickinson et al. (2015) Data

county	location	insystem	uptodateonimmunizations	hispanic	incomecat
1	Rural	94	37	44	0
2	Rural	85	39	23	2
3	Rural	85	42	12	0
4	Rural	93	39	18	2
5	Rural	82	31	6	2
6	Rural	80	27	15	1
7	Rural	94	49	38	0
8	Rural	100	37	39	0
9	Urban	93	51	35	1
10	Urban	89	51	17	1
11	Urban	83	54	7	2
12	Urban	70	29	13	1
13	Urban	93	50	13	2
14	Urban	85	36	10	1
15	Urban	82	38	39	0
16	Urban	84	43	28	1

Running cvcrand with the Dickinson et al. (2015) data

```
cvcrand insystem uptodate hispanic
   location incomecat,
   categorical(location incomecat)
   clusternum(16) treatmentnum(8)
   clustername(county) seed(10125)
   cutoff(0.1) balancemetric(l2)
   savedata(dickinson_constrained)
   savebscores(dickinson_bscores)
```

DukeHealth

Duke GLOBAL HEALTH INSTITUTE

```
cvcrand insystem uptodate hispanic
   location incomecat,
   categorical(location incomecat)
   clusternum(16) treatmentnum(8)
   clustername(county) seed(10125)
   cutoff(0.1) balancemetric(l2)
   savedata(dickinson_constrained)
   savebscores(dickinson_bscores)
```

DukeHealth

Duke GLOBAL HEALTH INSTITUTE

Running cvcrand with the Dickinson et al. (2015) data

```
cvcrand insystem uptodate hispanic
   location incomecat,
   categorical(location incomecat)
   clusternum(16) treatmentnum(8)
   clustername(county) seed(10125)
   cutoff(0.1) balancemetric(l2)
   savedata(dickinson_constrained)
   savebscores(dickinson_bscores)
```

Running cvcrand with the Dickinson et al. (2015) data

```
cvcrand insystem uptodate hispanic
   location incomecat,
   categorical(location incomecat)
   clusternum(16) treatmentnum(8)
   clustername(county) seed(10125)
   cutoff(0.1) balancemetric(l2)
   savedata(dickinson_constrained)
   savebscores(dickinson_bscores)
```

Running cvcrand with the Dickinson et al. (2015) data

```
cvcrand insystem uptodate hispanic
   location incomecat,
   categorical(location incomecat)
   clusternum(16) treatmentnum(8)
   clustername(county) seed(10125)
   cutoff(0.1) balancemetric(l2)
   savedata(dickinson_constrained)
   savebscores(dickinson_bscores)
```

DukeHealth

Duke | GLOBAL HEALTH INSTITUTE

First step: Enumerate & compute balance scores

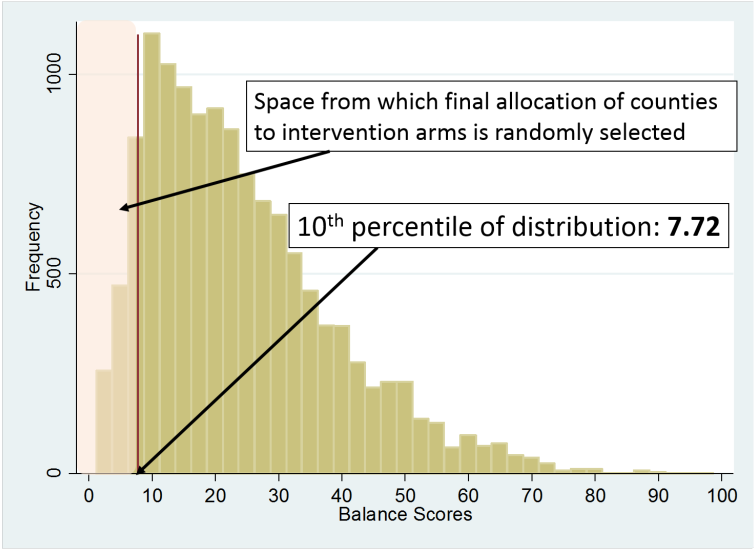
row	Cty 1	.	Cty 10	Cty 11	Cty 12	.	Cty 16	Bscores
1	1	.	0	0	0	.	0	93.56
2	1	.	0	0	0	.	0	43.57
3	1	.	1	0	0	.	0	41.62
4	1	.	0	1	0	.	0	62.06
.	.	.	.	.	.	.	.	.
12867	0	.	1	0	1	.	1	62.06
12868	0	.	0	1	1	.	1	41.62
12869	0	.	1	1	1	.	1	43.57
12870	0	.	1	1	1	.	1	93.56

First step: Enumerate & compute balance scores

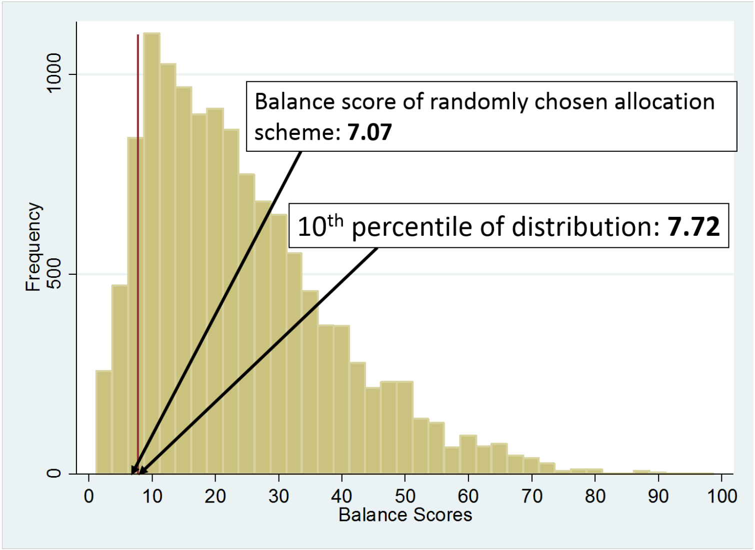
row	Cty 1	.	Cty 10	Cty 11	Cty 12	.	Cty 16	Bscores
1	1	.	0	0	0	.	0	93.56
2	1	.	0	0	0	.	0	43.57
3	1	.	1	0	0	.	0	41.62
4	1	.	0	1	0	.	0	62.06
.	.	.	.	.	.	.	.	.
12867	0	.	1	0	1	.	1	62.06
12868	0	.	0	1	1	.	1	41.62
12869	0	.	1	1	1	.	1	43.57
12870	0	.	1	1	1	.	1	93.56

Because of processing of large matrices, `cvcrand` uses `mata`

Second step: Sample from balance scores below the cutoff



Second step: Sample from balance scores below the cutoff



Final chosen allocation

	county	FinalScheme
1.	1	0
2.	2	1
3.	3	0
4.	4	1
5.	5	0
6.	6	0
7.	7	0
8.	8	1
9.	9	0
10.	10	1
11.	11	1
12.	12	1
13.	13	0
14.	14	0
15.	15	1
16.	16	1

Final chosen allocation

	county	FinalScheme
1.	1	Community-based
2.	2	Practice-based
3.	3	Community-based
4.	4	Practice-based
5.	5	Community-based
6.	6	Community-based
7.	7	Community-based
8.	8	Practice-based
9.	9	Community-based
10.	10	Practice-based
11.	11	Practice-based
12.	12	Practice-based
13.	13	Community-based
14.	14	Community-based
15.	15	Practice-based
16.	16	Practice-based

DukeHealth

Duke GLOBAL HEALTH INSTITUTE

Check Balance

```
. table1, by(FinalScheme) ///
> vars(inci contn \ uptod contn \ hisp contn \ loc cat \ incomecat cat) ///
> format(%2.1f)
```

Factor	Level	FinalScheme = 0	FinalScheme = 1	p-value
N		8	8	
% in CIIS, mean (SD)		88.3 (5.8)	85.8 (8.8)	0.51
% up-to-date, mean (SD)		40.4 (9.1)	41.3 (8.0)	0.84
% Hispanic, mean (SD)		21.6 (14.8)	23.0 (11.7)	0.84
Location	Rural	5 (63%)	3 (38%)	0.32
	Urban	3 (38%)	5 (63%)	
Average income	Low	3 (38%)	2 (25%)	0.82
	Med	3 (38%)	3 (38%)	
	High	2 (25%)	3 (38%)	

DukeHealth Duke GLOBAL HEALTH INSTITUTE

3. Analysis

Analysis Method: Clustered permutation test

- An appropriate analysis method **accounts for the constrained design**
  - Make inference in the constrained space
- The permutation test is ideally suited for inference when # of clusters is relatively small
  - Preserves appropriate type I error when equal # of clusters assigned to each intervention arm
- Li et al. (2015) recommend adjusting the test for the covariates used to constrain the design

Clustered permutation test: simple example

- Suppose the researchers obtain up-to-date immunization data on 20 children in each of the four counties
- This is a binary outcome variable (i.e., was the child up-to-date or not?)

Child ID	County	Up-to-date	Location	% In System
1	1	1	Rural	90
3	1	1	Rural	90
4	1	1	Rural	90
5	1	0	Rural	90
.	.	.	.	.
38	4	0	Rural	75
39	4	0	Rural	75
40	4	1	Rural	75

DukeHealth

Duke GLOBAL HEALTH INSTITUTE

Clustered permutation test: simple example

- Suppose the researchers obtain up-to-date immunization data on 20 children in each of the four counties
- This is a binary outcome variable (i.e., was the child up-to-date or not?)

```
. tab FinalScheme, summarize(outcome)
```

FinalScheme	Summary of outcome Mean	Std. Dev.	Freq.
Community	.8	.40509575	40
Practice	.875	.33493206	40
Total	.8375	.37123639	80

DukeHealth

Duke | GLOBAL HEALTH INSTITUTE

First step: Run regression

Obtain average residuals by cluster

```
. quietly logit outcome location insystem
. predict double _resid, residuals
. bys county: egen _residmn = mean(_resid)
. egen _tag = tag(county)
. quietly keep if _tag == 1
. list county location insystem _residmn
```

	county	location	insystem	_residmn
1.	1	Rural	90	.1028244
2.	2	Urban	92	-.1099574
3.	3	Urban	80	.1278469
4.	4	Rural	75	-.1301437

DukeHealth

Duke GLOBAL HEALTH INSTITUTE

Second step: Input the constrained matrix

County 1	County 2	County 3	County 4	**Bscores**
1	1	0	0	**2.779**
1	0	1	0	**0.034**
1	0	0	1	**3.187**
0	1	1	0	**3.187**
0	1	0	1	**0.034**
0	0	1	1	**2.779**

Second step: Input the constrained matrix

For computational reasons, replace 0 with -1

County 1	County 2	County 3	County 4	**Bscores**
1	1	-1	-1	**2.779**
1	-1	1	-1	**0.034**
1	-1	-1	1	**3.187**
-1	1	1	-1	**3.187**
-1	1	-1	1	**0.034**
-1	-1	1	1	**2.779**

DukeHealth

Duke GLOBAL HEALTH INSTITUTE

Second step: Input the constrained matrix

County 1	County 2	County 3	County 4
1	1	-1	-1
1	-1	1	-1
-1	1	-1	1
-1	-1	1	1

Third step: Multiply the constrained and residual matrix

$$
\underset{\text{Permutation Matrix}}{\begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \end{pmatrix}}
\underset{\substack{\text{Average}\\\text{Residuals}}}{\begin{pmatrix} 0.1028 \\ -0.1099 \\ 0.1278 \\ -0.1301 \end{pmatrix}}
= \left| \begin{pmatrix} -0.0048 \\ 0.4708 \\ -0.4708 \\ 0.0048 \end{pmatrix} \right|
= \underset{\substack{\text{Test}\\\text{Statistics}}}{\begin{pmatrix} 0.0048 \\ 0.4708 \\ \boxed{0.4708} \\ 0.0048 \end{pmatrix}}
$$

- Intervention effect p-value: Percentage of times other test statistics are greater than the observed test statistic (0.4708)
- In this case: p = 0.00
- In larger data examples, these matrices can get large, requiring `mata` to process

DukeHealth Duke GLOBAL HEALTH INSTITUTE

Introducing cptest

`cptest` **for clustered permutation test**

cptest *varlist*, clustername(*varname*) directory(*string*)
    cspacedatname(*string*) outcometype(#) [
    categorical(*varlist*)]

This program is available to download using `ssc install cvcrand`

DukeHealth   Duke GLOBAL HEALTH INSTITUTE

Analysis of Dickinson et al. (2015) data

- Researchers have collected up-to-date immunization status on 300 children in each county (simulated data)
  - Binary outcome (1 = up-to-date on immunizations; 0 = not up-to-date)
- Is there a significant difference in up-to-date immunization rate between the two interventions?

Simulated outcome data

```
. tab FinalScheme, summarize(outcome)
```

FinalScheme	Summary of outcome Mean	Std. Dev.	Freq.
0	.78916667	.40798529	2,400
1	.85958333	.34749121	2,400
Total	.824375	.38054044	4,800

```
. tab FinalScheme, summarize(outcome)
```

FinalScheme	Summary of outcome Mean	Std. Dev.	Freq.
Community	.78916667	.40798529	2,400
Practice	.85958333	.34749121	2,400
Total	.824375	.38054044	4,800

Run cptest on Dickinson et al. (2015) simulated data

```
cptest outcome insystem uptodate
   hispanic location incomecat,
   clustername(county)
   directory(P:\Program\Stata Conf)
   cspacedatname(dickinson_constrained)
   outcometype(Binary)
   categorical(location incomecat)
```

Run cptest on Dickinson et al. (2015) simulated data

```
cptest outcome insystem uptodate
   hispanic location incomecat,
   clustername(county)
   directory(P:\Program\Stata Conf)
   cspacedatname(dickinson_constrained)
   outcometype(Binary)
   categorical(location incomecat)
```

DukeHealth

Duke GLOBAL HEALTH INSTITUTE

Run cptest on Dickinson et al. (2015) simulated data

```
cptest outcome insystem uptodate
   hispanic location incomecat,
   clustername(county)
   directory(P:\Program\Stata Conf)
   cspacedatname(dickinson_constrained)
   outcometype(Binary)
   categorical(location incomecat)
```

Run cptest on Dickinson et al. (2015) simulated data

```
cptest outcome insystem uptodate
   hispanic location incomecat,
   clustername(county)
   directory(P:\Program\Stata Conf)
   cspacedatname(dickinson_constrained)
   outcometype(Binary)
   categorical(location incomecat)
```

```
Logistic regression was performed
```

(*output omitted*)

Clustered permutation test p-value = $\boxed{0.0047}$

4. Conclusions and Future Research

Conclusion

- CRTs in general should use some form of restricted randomization
- Constrained randomization is a good option
  - especially when the number of clusters to randomize is small
  - and when there are several covariates to balance across intervention arms
- `cvcrand` is an easy-to-implement program to perform constrained randomization
- Constrained randomization may be followed up by a clustered permutation test, implemented using the program `cptest`

Future Research

- Covariate constrained randomization methods for CRTs with more than two intervention arms
- Evaluating the performance of covariate constrained randomization when cluster sizes are expected to be unequal

Acknowledgments

- Coauthors
  - Elizabeth Turner
  - Fan Li
  - Hengshi Yu
- Duke Global Health Institute Research Design & Analysis Core
- Joy Noel Baumgartner
  - The `cvcrand` program was used in the design of the study *Evaluation of an Early Childhood Development Intervention for HIV-Exposed Children in Cameroon* sponsored by Catholic Relief Services
- Helpful resources
  - Statalist forums
  - Resources on `mata` and Stata programming by Dr. Christopher Baum

DukeHealth Duke GLOBAL HEALTH INSTITUTE

References

Carter, B. R., and K. Hood. 2008. Balance algorithm for cluster randomized trials. *BMC Medical Research Methodology* 8: 65.

Dickinson, L. M., B. Beaty, C. Fox, W. Pace, W. P. Dickinson, C. Emsermann, and A. Kempe. 2015. Pragmatic cluster randomized trials using covariate constrained randomization: A method for practice-based research networks (PBRNs). *The Journal of the American Board of Family Medicine* 28(5): 663–672.

Fiero, M. H., S. Huang, E. Oren, and M. L. Bell. 2016. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials* 17(1): 72.

Gallis, J. A., F. Li, H. Yu, and E. L. Turner. Submitted. `cvcrand` and `cptest`: Efficient Design and Analysis of Cluster Randomized Trials. *Stata Journal* .

Ivers, N., M. Taljaard, S. Dixon, C. Bennett, A. McRae, J. Taleban, Z. Skea, J. Brehaut, R. Boruch, and M. Eccles. 2011. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. *BMJ* 343: d5886.

Ivers, N. M., I. J. Halperin, J. Barnsley, J. M. Grimshaw, B. R. Shah, K. Tu, R. Upshur, and M. Zwarenstein. 2012. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials* 13: 120.

Li, F., Y. Lokhnygina, D. M. Murray, P. J. Heagerty, and E. R. DeLong. 2015. An evaluation of constrained randomization for the design and analysis of group-randomized trials. *Statistics in Medicine* 35(10): 1565–79.

Li, F., E. L. Turner, P. J. Heagerty, D. M. Murray, W. M. Vollmer, and E. R. DeLong. 2017. An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes. *Statistics in Medicine* .

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

Raab, G. M., and I. Butcher. 2001. Balance in cluster randomized trials. *Statistics in medicine* 20(3): 351–365.

Turner, E. L., F. Li, J. A. Gallis, M. Prague, and D. Murray. 2017a. Review of Recent Methodological Developments in Group-Randomized Trials: Part 1 - Design. *American journal of public health* 107(6): 907–15.

Turner, E. L., M. Prague, J. A. Gallis, F. Li, and D. Murray. 2017b. Review of Recent Methodological Developments in Group-Randomized Trials: Part 2 - Analysis. *American Journal of Public Health* 107(7): 1078–1086.