Paper 1570-2014

Adjusting Clustering: Minimize Your Suffering!
Improving balance in baseline covariates in cluster randomized trials

Brenda Beaty, MSPH, University of Colorado School of Medicine, Aurora, CO
L. Miriam Dickinson, PhD, University of Colorado School of Medicine, Aurora, CO

ABSTRACT

The randomized controlled trial is the 'gold standard' in experimental design. In theory, assigning participants randomly to the experimental groups (intervention and control) make it more likely that any effect is due to the intervention rather than underlying characteristics of the participants. Cluster- or group- randomized trials (CRTs) are seeing increased use. There are many reasons for this, including more pragmatic trials associated with comparative effectiveness research. Examples of clusters that could be randomized for study are clinics or hospitals, counties within a state, and other geographical areas such as communities. In many of these trials, the number of clusters is relatively small. This fact complicates randomization, since a random grouping could still randomize clusters that may have very similar characteristics to one group. In order for the randomization to be most efficient, we would have baseline characteristic data about the clusters. We want to 'balance' important baseline covariates at the cluster level between the intervention and control groups. For example, if we randomize eight counties, a simple randomization could put all counties with high socioeconomic status in one group or the other, leaving us without a good way to estimate the effect size independent of socioeconomic status. However, there are strategies to prevent this. These include matching, stratification, minimization and covariate-constrained randomization. In this paper, each method is discussed, and a county-level Health Outcomes example of covariate-constrained randomization is shown for intermediate SAS® users working with SAS® Foundation for Release 9.2 and SAS/STAT® on a Windows operating system.

INTRODUCTION

The number of trials randomized at a cluster, rather than at an individual level, is increasing. Due to the current emphasis on Comparative Effectiveness Research, many pragmatic trials are being conducted among different clinics, hospital, and/or geographic locations. In many of these trials, the number of clusters is relatively small (in the 4-20 range). This can be a problem if there are important covariates at the cluster level that are not 'balanced' across the intervention and control arms of a study.

CAUSES OF UNBALANCED ALLOCATION

There are at least two reasons why unbalanced randomizations are more likely to occur in CRTs than in traditional randomized controlled trials (RCTs): the small number of clusters, and increased likelihood of geographical contiguousness (Moulton 2004). The number of clusters in a CRT is usually less than 40, in the 4-20 range. Because clusters are often practices, hospitals, communities or counties, there is also the added concern of 'contamination' between the control and intervention units.

WHY TRY TO BALANCE?

The randomized controlled trial has long been the 'gold standard' in medical research. The purpose of randomization is to insure that each subject has an equal probability of being assigned to the intervention or control arm. The same principle applies in cluster randomized trials - we want each cluster to have an equal probability of being assigned to each arm. This can be achieved with a simple randomization, but when the data are analyzed, if the clusters are unbalanced on an important covariate, the trial may produce a biased estimate of any intervention effect.

Having balanced covariates between the clusters in the intervention group and the control group increases statistical power and precision. Unbalanced cluster allocation will generally increase the amount of variance in the intervention effect, decreasing the efficiency. The amount of increased variance depends on many factors, including the number of covariates, the amount of similarity among individuals within a cluster, the number of clusters, and the number of individuals within a cluster.

EXAMPLE STUDY

Immunizations are an important part of preventive care for children. Efforts are being made to increase the up-to-date (UTD) rate in children, to improve both individual and 'herd' immunity. One strategy to increase rates is called 'Reminder/Recall'. Parents are either 'Reminded' to get their child immunized (prior to when the immunization is due) or 'Recalled' to an immunization provider (for children who have an overdue immunization).

We conducted a trial of recall for overdue immunizations for children aged 19-35 months in 16 counties (8 Urban and 8 Rural) in Colorado. We wanted to compare 'centralized' recall done at the county level (intervention group) with usual care practice-based recall done at the level of the individual medical practice (control group) for the outcomes of becoming UTD, having any new vaccine in the immunization registry, and the cost of these activities and events (Kempe, 2013). We randomized at the county level stratified by urban/rural status in hopes of avoiding too much crossover between county-based and practice-based individuals. The challenge was to randomize counties in a way that would let us draw valid conclusions about the effect of centralized recall.

The rest of this paper will focus only on the randomization of the 8 urban counties. We needed to randomize 4 counties to the intervention group and 4 to the control group. Among the counties, there was variability in the baseline covariates. For example, the percentage of the population that self-reported as Hispanic varied from 7-35% in these counties. If our randomization was unbalanced, we could have ended up with a mean of about 15% Hispanic in each of the intervention counties and a mean of 26% per county in the control group, which would complicate estimating the effect of the intervention in the Hispanic population.

POSSIBLE STRATEGIES

A recent review (Ivers, 2012) was published that discussed a number of methodologies to maximize the ability to balance baseline covariates among clusters. They included matching, stratification, minimization and covariate-constrained randomization. We will briefly discuss these in order and then apply the most appropriate one to our data set.

Pair-wise matching:
In this method, clusters are matched on all known predefined variables, such as population density and median income. Each cluster in the intervention group has a matched cluster in the control group. There are several disadvantages to this method. Loss of follow-up from one cluster also removes its match from analysis, which could be catastrophic for studies with few clusters. It also makes it difficult to properly estimate the variance within and between clusters (Donner 2004).

Stratification:
In this technique, clusters are stratified by one or more baseline covariates, and each level of the baseline covariate (sometimes called 'blocks') is treated as a separate experiment. Randomization is then performed within each stratum. Similar to an individually randomized trial, when using stratification in a cluster randomized trial, the number of strata must be few. Simulations suggest that as the total number of strata approach half the total number of clusters, stratification becomes ineffective.

Minimization:
This method randomly assigns the first cluster. For each following cluster, the covariates of the cluster(s) previously enrolled are considered, and the newest cluster is assigned to the study group that provides better overall balance. This method is only available for sequentially enrolled clusters, and there is concern that selection bias could occur due to the fact that the next assignment can be predicted in some situations (Scott 2002). Also, continuous variables must be split into categories, and there must be a rapid way of determining randomization status at the time of study entry.

Minimization represents an option for when clusters are recruited and allocated sequentially.

Covariate-constrained randomization:
If covariate data is available for all clusters *prior* to allocation, we can use this technique. Here, all possible allocations of participating clusters are calculated, and then the list is limited to those that met the pre-specified criteria for balance. The actual allocation is then chosen randomly from this acceptable list. The main drawback to this method is that the recruitment of clusters must be completed prior to cluster allocation.

In our case, the counties had all agreed to participate before being randomized and we were able to obtain county-level data, so we chose to use covariate-constrained randomization.

OUTLINE OF COVARIATE-CONSTRAINED RANDOMIZATION PROCESS

1. *Standardize all important covariates to create z scores (PROC STANDARD)*
2. *Create all combinations of possible allocations (PROC PLAN)*
3. *Match all possible allocations to county data*
4. *Compare sum of squared differences in z scores over covariates for all possible allocations*
5. *Take a pre-determined percentage or number of combinations with the lowest total sum of squared differences between groups*
6. *Select final allocation randomly from the subset of most balanced allocations*

SAMPLE OF DATA SET

We obtained baseline information on the counties from the Colorado Department of Public Health and Environment (CDPHE), the Colorado Immunization Information System (CIIS), and 2010 Census data.

The county-level baseline covariates we used are listed below:
1. Percent of children between 0 and 4 who had ≥2 immunization records in CIIS
2. Number of 19-35 month olds by county
3. Percent up-to-date (UTD) rate at baseline
4. Population by race (white, black or other), ethnicity (Hispanic/Latino or not), and median income
5. Number of Pediatric (Peds), Family Medicine (FM) practices, and Community Health Centers (CHC)

County	% in CIIS	19-35 mos	% UTD	% White	% Black	% Hispanic / Latino	Median Income	N Peds	N FM	N CHC
1	93%	3,779	51%	80%	4%	35%	$52,923	6	40	11
2	89%	11,807	51%	80%	10%	17%	$58,302	21	47	6
3	83%	9,453	54%	92%	2%	7%	$93,819	14	23	1
4	70%	12,354	29%	84%	8%	13%	$54,839	14	53	10
5	93%	10,008	50%	90%	2%	13%	$63,857	18	53	3
6	85%	5,343	36%	93%	2%	10%	$53,502	7	38	7
7	82%	3,143	38%	85%	3%	39%	$39,570	6	22	7
8	84%	6,056	43%	87%	1%	28%	$52,457	2	20	8

1. Standardize all important covariates to create z scores (PROC STANDARD)

Abbreviated code:

```
proc standard data=countydata mean=0 std=1 out=zscores print;
  var ciis--chc;
run;
```

Partial output:

```
county    ciis   nkids   utd   white    black     hispanic   income

1         1.10  -1.10    0.78   -1.26   -0.00     1.22      -0.36
2         0.56   1.13    0.78   -1.26    1.85    -0.27      -0.02
3        -0.25   0.47    1.12    1.12   -0.62    -1.09       2.23
4        -2.01   1.28   -1.68   -0.47    1.23    -0.60      -0.24
```

2. Create all combinations of possible allocations (PROC PLAN)

Quick review of combinations:

Combinations are selected from a group where the order of the elements does not matter. In a situation where n things are taken k at a time without repetition, the number of combinations is:

$$\frac{n!}{k!\,(n-k)!}$$

In our case, we want to choose 4 elements from a group of 8. The number of possible combinations is:

$$\frac{8!}{4!(8-4)!} = \frac{8!}{4!4!} = \frac{8*7*6*5*4!}{4*3*2*1*4!} = \frac{1680}{24} = 70$$

One way to create all combinations is using PROC PLAN, as shown below:

```
proc plan seed=60359; ❶
  factors block=70 ordered
    county=4 of 8 comb; ❷
  ods output plan=comb; ❸
run;
```

❶ Here we select a seed, so that we can reproduce these results. If we leave this out, SAS will take the seed number from the clock.
❷ The factors statement defines the design of the plan.
❸ Creates an ODS data set of the plan shown in part below.

Partial output data:

block	county1	county2	county3	county4
1	1	2	3	4
2	1	2	3	5
3	1	2	3	6
4	1	2	3	7
5	1	2	3	8
6	1	2	4	5
7	1	2	4	6
8	1	2	4	7
9	1	2	4	8
10	1	2	5	6

We will use 'block' to match the design to our data.

Because we program in SAS, there are at least several other ways to generate all combinations, as shown below:

- M. Ashraf Chaudhary and Lawrence H. Moulton wrote a paper in 2006 entitled "A SAS Macro for constrained randomization of group-randomized designs", which is referenced at the end of this paper if you'd rather use a macro.

- There is also the ALLCOMB function that could be used for this purpose as well as shown below:

```
data comb1 (keep=j c1-c8);
array c(*) c1-c8 (1 1 1 1 0 0 0 0);
n=dim(c);
k=4;
ncomb=comb(n,k);
do j=1 to ncomb;
a=allcomb(ncomb,k,of c1-c8);
   output;
end;
run;
```

- SAS macro TS 498 Generating Combinations and Permutations

3. Match all possible allocations to county data

We need to create a matrix of combinations with a 0/1 indicator variable to for each county to match with intervention (or control) group.

In the 'comb' data set, we have a list of selected counties for the intervention group. We then need to assign the counties *not* in that list to the control group, so we have a group variable for each of the 8 counties.

```
data comb2 (drop=county1-county4);
  set comb;
  array c (*) c1-c8; ❶
  array county (*) county1-county4; ❷
  do i=1 to 8; ❸
    do j=1 to 4; ❹
       if county(j)=i then c(i)=1;
    end;
  if c(i)=. then c(i)=0;
  end;
  drop i j;
run;
```

❶ The array c will contain 8 elements, c1-c8 to assign a county to either intervention (1) or control (0) group.
❷ These are the counties selected in the PROC PLAN step.
❸ The outer DO loop creates the 8 variables c1-c8.
❹ The inner DO loop assigns the county value to be 1 only if the county number was selected in the PROC PLAN.

The data set 'comb2' will have 70 records, one for each block.

Partial output data:

block	c1	c2	c3	c4	c5	c6	c7	c8
1	1	1	1	1	0	0	0	0
2	1	1	1	0	1	0	0	0
3	1	1	1	0	0	1	0	0
4	1	1	1	0	0	0	1	0
5	1	1	1	0	0	0	0	1
6	1	1	0	1	1	0	0	0
7	1	1	0	1	0	1	0	0
8	1	1	0	1	0	0	1	0
9	1	1	0	1	0	0	0	1
10	1	1	0	0	1	1	0	0

Next we need to copy the county data 70 times and merge with the allocation data. There are 8 records in data set 'zscores' (one for each county), so data set 'z' will have 8 X 70 records = 560 records.

```
data z;
  set zscores;
  do block=1 to 70;
    output;
  end;
run;
```

Merge county data with all possible allocations:

```
proc sort data=z; by block county; run;

data z2;
  merge comb2 z;
  by block;
run;
```

Partial output data:

block	c1	c2	c3	c4	c5	c6	c7	c8	county
1	1	1	1	1	0	0	0	0	1
1	1	1	1	1	0	0	0	0	2
1	1	1	1	1	0	0	0	0	3
1	1	1	1	1	0	0	0	0	4
1	1	1	1	1	0	0	0	0	5
1	1	1	1	1	0	0	0	0	6
1	1	1	1	1	0	0	0	0	7
1	1	1	1	1	0	0	0	0	8

We're almost there! Now, we just need to assign each county to either control or intervention according to the possible combinations. To do this, we use the variables c1-c8.

```
data z3;
  set z2;
  array c(*) c1-c8;
  do i=1 to dim(c);
    group=(c(i)=c(county));
  end;
run;
```

Partial output data:

block	county	group	ciis	nkids	utd
1	1	0	1.10	-1.10	0.78
1	2	0	0.56	1.13	0.78
1	3	0	-0.25	0.47	1.12
1	4	0	-2.01	1.28	-1.68
1	5	1	1.10	0.63	0.67
1	6	1	0.02	-0.66	-0.89
1	7	1	-0.39	-1.27	-0.67
1	8	1	-0.12	-0.47	-0.11
2	1	0	1.10	-1.10	0.78
2	2	0	0.56	1.13	0.78

4. Compare sum of squared differences in z scores over covariates for all possible allocations

```
proc means data=z3 mean;
  class block group;
  var ciis--chc;
  ods output summary=s1 (drop=nobs f: v:);
run;
```

Data set 's1' contains the mean values by group for all possible combinations of counties. Partial output data:

block	group	ciis_Mean	nkids_Mean	utd_Mean
1	0	-0.15	0.44	0.25
1	1	0.15	-0.44	-0.25
2	0	0.62	0.28	0.84
2	1	-0.62	-0.28	-0.84

Now, we summarize each possible combination by calculating the sum of squared differences between the control and intervention groups of counties.

```
proc sort data=s1; by block group; run;
data s2 (drop=group i);
  set s1;
  by block;
  retain fc fk futd fw fb fh fi fp ffm fchc;
  array f(*) fc fk futd fw fb fh fi fp ffm fchc;
  array z(*) ciis nkids utd white black hisp income peds fm chc;
  array sqdiff (*) dciis dkids dutd dwhite dblack dhisp dincome dpeds dfm dchc;

  if first.block then do i=1 to dim(f);
    f(i)=z(i);
  end;

  else if last.block then do;
    do i=1 to dim(f);
    sqdiff (i)=(f(i)-z(i))**2;
    end;
  totalssq=sum(dciis, dkids, dutd, dwhite, dblack, dhisp, dincome, dpeds, dfm, dchc);
  output;
  end;
run;
```

Partial output data:

```
block     dciis      dkids       dutd      dwhite     dblack   totalssq
1         0.09256    0.79029     0.25313   0.88791    1.51351  5.33719
2         1.56432    0.31833     2.81250   0.12052    0.09459  8.45858
3         0.50392    0.00666     0.80000   0.00246    0.09459  2.36804
```

5. Take a pre-determined percentage or number of combinations with the lowest total sum of squared differences between groups

```
* N=70;
proc univariate data=s2;
  var totalssq;
  id block;
  ods output quantiles=q (drop=varname);
run;

proc print data=q noobs; run;
```

This gives:

```
Quantile       Estimate

…
25% Q1          2.85355
10%             1.71596
5%              1.66583
1%              1.65852
0% Min          1.65852
```

Here, we took the most balanced (lowest sum-of-squares) approximately 10% of all possible randomizations (N=8) as a 'good' match for control and intervention groupings.

```
data all;
  set s2;
  if totalssq<1.72 then Top10Percent=1;
    else top10percent=0;
  format totalssq 8.2;
run;
```

Comparison of the most balanced 10% percent of allocations in terms of baseline covariates versus the less balanced 90%:

Squared differences:	Top 10% of possible randomizations (n=8)	Lower 90% of possible randomizations (n=62)	P value (Wilcoxon)
Median total sum of squared differences	1.68	5.21	<.0001
Median sum of squared differences in:			
Percent of children with records in CIIS	0.37	0.41	0.56
Number of 19-35 month old children	0.02	0.32	0.003
Percent UTD	0.04	0.25	0.0009
Percent White race	0.16	0.30	0.30
Percent Black race	0.02	0.21	0.002
Percent Hispanic ethnicity	0.20	0.25	0.26
Median Income	0.25	0.53	0.08
Number of Pediatric practices	0.09	0.36	0.38
Number of Family Medicine practices	0.16	0.26	0.24
Number of Community Health Centers	0.21	0.28	0.92

6. Select final allocation randomly from the subset of most balanced allocations

```
data subset;
  set all;
  where top10percent=1;
  rand1=ranuni(69872);
run;

proc sort data=subset; by rand1; run;

data FinalAllocationOfGlory;
  set subset;
  if _n_=1;
run;

proc print data=FinalAllocationOfGlory noobs; var block totalssq; run;

block    totalssq
18        1.67
```

CONCLUSION

There are evolving ways to avoid unbalanced baseline covariates in cluster-randomized trials. The choice of optimum method depends on the source and timing of available data. In our case, covariate-constrained randomization proved a very useful tool.

REFERENCES

Chaudhary MA, Moulton LH. A SAS macro for constrained randomization of group-randomized designs. *Comput Methods Programs Biomed.* 2006;83:205-210. PM:16870302

Donner A, Klar N. Pitfalls and Controversies in Cluster Randomization Trials. Amer J Pub Health 2004;94(3).

Ivers NM, Halperin IJ, Barnsley J, Grimshaw JM, Shah BR, Tu K,Upshur R, Zwarenstein M. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. Trials 2012, 13:120.

Kempe A, Saville A, Dickinson LM, Eisert S, Reynolds J, Herrero D, Beaty B, Albright K, Dibert E, Koehler V, Lockhart S, Calonge N. Population-based versus practice-based recall for childhood immunizations: a randomized controlled comparative effectiveness trial. Am J Public Health 2013 Jun;103(6):1116-23.

Moulton LH. Covariate-based constrained randomization of group-randomized trials. Clinical trials 2004;1:297-305.

Scott NA, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials: a review. Controlled Clinical Trials 23 (2002) 662-67.

http://support.sas.com/techsup/technote/ts498.html accessed 10/5/13.

ACKNOWLEDGEMENTS

The author gratefully acknowledges Drs. Miriam Dickinson and Michelle Torok, as well as Anna Furniss and Patrick Thornton for their kind help and thoughtful review.

RECOMMENDED READING

Glynn RJ, Brookhart MA, Stedman M, Avorn J, Solomon DH. Design of cluster-randomized trials of quality improvement interventions aimed at medical care providers. *Med Care.* 2007;45:S38-S43. PM:17909381

Samuel-Hodge CD, Kraschnewski JL, Keyserling TC, Bangdiwala SI et al. Optimized probability sampling of study sites to improve generalizability in a multisite intervention trial. *Prev Chronic Dis.* 2010;7:A10. PM:20040225

Raab GM, Butcher I. Balance in cluster randomized trials. *Stat Med.* 2001;20:351-365. PM:11180306

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Brenda Beaty, MSPH
Colorado Health Outcomes Program
University of Colorado Denver
Mail Stop F443
13199 E. Montview Ave., Suite 300
Aurora, CO 80045-0508
Work Phone: (303) 724-1076
E-mail: Brenda.Beaty@ucdenver.edu