# Power and Sample Size for Cluster Randomized Trials

Nambari Short Course

17 July 2019

# Overview

- What is a cluster RCT?
- Relevance to multilevel modeling in RCT
- Importance of multiple sources of variation in design of RCT
- Power calculations: More clusters or more individuals?

# Cluster Randomized Trial

A CRT is a randomized trial where the *cluster* is the unit of analysis.

- Carried out when individual-level randomization is not feasible
- Example: Test intervention to help individuals link to HIV care
  - Instead of randomizing individuals, randomize clinics where individuals are treated
  - Might have 10 clinics in each arm
- Example: Test intervention to improve academic achievement in students
  - Instead of randomizing individual children, randomize schools where they are enrolled
  - In a town with 30 schools, might have 15 in each arm
- Also popular with laboratory studies
  - For each experimental condition, will have certain number of replicate measures
  - How many replicates per study?

# Multilevel models and CRT

Multilevel models can be used as the main analysis vehicle for cluster RCTs.
**Notation**

$$
\begin{aligned}
j &= 1, \ldots, J \\
&= \text{cluster index} \\
i &= 1, \ldots, n \\
&= \text{individual within cluster (assume equal cluster sizes)} \\
Y_{ij} &= \text{outcome of individual } i \text{ in cluster } j \\
X_j &= \text{treatment applied to cluster } j
\end{aligned}
$$

# Multilevel model

**Level 1**

$$Y_{ij} \sim \mathcal{N}(\alpha_j + X_j\theta, \sigma^2)$$

**Level 2**

$$\alpha_j \sim \mathcal{N}(\mu, \tau^2)$$

# Consider using sample means for comparison

For illustration, consider group where $X_j = 0$ and assume $J_0$ is the number of clusters with $X_j = 0$. The sample mean is

$$\overline{Y} = (1/nJ_0)\sum_{i=1}^{n}\sum_{j=1}^{J_0} Y_{ij}$$

It can be shown that the variance of $\overline{Y}$ is

$$\text{var}(\overline{Y}) = (1/J_0)(\tau^2 + \sigma^2/n)$$

- Most of the variation comes from cluster-level variation $\tau^2$
- As clusters get larger ($n \to \infty$), variation is dominated by $\tau^2$
- If the goal is to reduce var($\overline{Y}$), might not always be a good idea to add observations to clusters

# Intraclass correlation coefficient

The ICC is the within-cluster correlation of individual measurements.

$$\text{ICC} = \frac{\tau^2}{\tau^2 + \sigma^2}$$

- As ICC gets larger, variance is dominated by $\tau^2$.
- This implies that reducing variance (increasing power) depends more on adding clusters than adding observations

# Variance inflation

Compare the variance under clustering to the variance without clustering.

$$\begin{aligned} \text{VIF} &= 1 + (n-1)\frac{\tau^2}{\tau^2 + \sigma^2} \\ &= 1 + (n-1) * ICC \end{aligned}$$

This is sometimes also called the *design effect*

# Example

The Shared Incentives Trial

Asch DA, Troxel AB, Stewart WF, Sequist TD, Hones JB, Hirsch AG, Hoffer K, Zhu J, Wang W, Hodlofski A, Frasch AB, Weiner MG, Finnerty DD, Rosenthal MB, Gangemi K, Volpp KG (2015). Effect of Financial Incentives to Physicians, Patients, or Both on Lipid Levels: A Randomized Clinical Trial. *JAMA* 314(18): 1926-35.

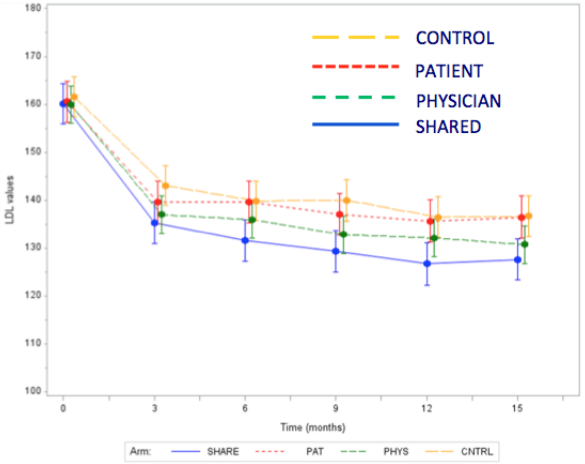
# Shared incentives participants

- 238 primary care physicians at 3 health systems
- 1,503 patients
  - age $18 - 80$
  - high cardiac risk
  - high LDL cholesterol
- approximately 6-7 patients per physician
- Four interventions
  - "control"
  - patient incentives: daily lottery for statin adherence
  - physician incentives: direct payments to physicians for quarterly goal achievement
  - shared incentives: both patient and physician, each at half value
- Primary outcome: change in LDL over 12 months

# LDL reduction at 12 months

| | Control | Patient Incentives | Physician Incentives | Shared Patient and Physician Incentives |
|---|---|---|---|---|
| $\Delta$ LDL | 26.6 | 26.4 | 30.0 | 36.8 |
| CI | $22.7 - 30.6$ | $22.5 - 30.3$ | $26.6 - 33.4$ | $32.9 - 40.6$ |

# Average LDL over time



# Sample size and power

Definition of power: the probability that we reject the *null* hypothesis when the *alternative* hypothesis is true

What affects statistical power?

- True alternative effect size
- Variation
- Sample size

# Review: sample size per group for a two-group test

Suppose

- Effect size: $\mu_1 - \mu_2$
- Significance level: $\alpha$
- Power: $1 - \beta$
- Standard deviation: $\sigma$

$$N = \frac{(z_{\alpha/2} + z_{\beta})^2 2(\sigma^2)}{(\mu_1 - \mu_2)^2}$$

# Sample size per arm for cluster-randomized two-group test

- Sample size must be inflated by the *design effect*
  - $m =$ number of individuals within each cluster
  - $\rho = \text{ICC} = \frac{\tau^2}{\tau^2+\sigma^2}$

$$
\begin{aligned}
\text{Deff} &= 1 + (m-1)\rho \\
N^* &= \frac{(z_{\alpha/2} + z_{\beta})^2 2\sigma^2 (1 + (m-1)\rho)}{(\mu_1 - \mu_2)^2} \\
N^* &= N \times \text{Deff}
\end{aligned}
$$

# Example

- Study of intervention to improve hypertension care
  - $\mu_1 = 140$, $\mu_2 = 120$
  - $\sigma = 30$
  - $\alpha = 0.05$, $\beta = 0.8$
- Ignoring clustering

$$\begin{aligned} N &= \frac{(z_{\alpha/2} + z_{\beta})^2 2\sigma^2}{(\mu_1 - \mu_2)^2} \\ &= \frac{(1.96 + 0.84)^2 2(30)^2}{(120 - 140)^2} = 36 \end{aligned}$$

# Example

What if the intervention is delivered at the clinic level?

- Two-arm CRT in clinics
- $m = 10$ individuals enrolled within each clinic
- $\rho = \text{ICC} = 0.05$

$$\begin{aligned} \text{Deff} &= 1 + (m-1)\rho = 1 + 9(0.05) = 1.45 \\ N^* &= N \times \text{Deff} = 36 * 1.45 = 52 \end{aligned}$$

# Example, cont

$$\begin{aligned} \text{Deff} &= 1 + (m-1)\rho = 1 + 9(0.05) = 1.45 \\ N^* &= N \times \text{Deff} = 36 * 1.45 = 52 \end{aligned}$$

How many clusters do we need?

$$\begin{aligned} N^* &= 52 \\ \text{number of clusters} &= 52/10 = 5.2 \end{aligned}$$

Round up to 6 clusters per group, or 120 participants

# Example: R code

```
> # calculate the sample size per group if there were no clustering
> ss.simple = power.t.test(n=NULL, delta=20, sd=30, power=0.8, sig.level=0.05)
> ss.simple

     Two-sample t test power calculation

              n = 36.3058
          delta = 20
             sd = 30
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

# Example: R code

```
> # Assume cluster size of 10 and ICC of 0.05
> # Calculate the design effect
> J = 10
> ICC = 0.05
> deff = 1 + (J-1)*ICC
> deff
[1] 1.45

> # calculate the inflated sample size
> ss.star = ceiling(ss.simple$n)*deff
> ss.star
[1] 53.65

> # calculate the minimum number of clusters
> clusters.simple = ceiling(ss.star/J)
> clusters.simple
[1] 6
```

# Example: R code

```
> # calculate minimum number of clusters uinsg clusterPower routine
> install.packages("clusterPower")
> library("clusterPower")
> cluster.R1 = crtpwr.2mean(alpha=0.05, power=0.8,
+                           m=NA, n=10,
+                           cv=0,
+                           d=20, varw=900
+                           icc=0.05)
> cluster.R1
       m
6.244887
>
> # calculate total needed sample size
> nstar.R1 = ceiling(cluster.R1)*J
> nstar.R1
       m
70
```

# Sample sizes for cluster-randomized trial

What if the cluster sizes are not all the same?

- The *design effect* must include an additional factor
- calculate coefficient of variation of cluster sizes
- *cv* is the ratio of the variance to the mean
  - $K$ is the number of clusters
  - cluster $k$ has size $j_k$
  - the average cluster size is $\bar{j}$

$$
\begin{aligned}
cv &= \frac{\frac{1}{K-1}\sum_{k=1}^{K}(j_k - \bar{j})^2}{\bar{j}} \\
\text{Deff} &= 1 + \left[(cv^2 + 1)\bar{j} - 1\right] ICC \\
N^* &= N \times \text{Deff}
\end{aligned}
$$

# Summary

- CRTs are useful/necessary for certain kinds of interventions
- BUT they are less efficient that individual RCTs
  - Avoid large cluster sizes
  - Maximize the number of clusters
  - Avoid settings with large ICCs
- Account for clustering properly in the analysis