Living Textbook Grand Rounds Series

Demystifying Biostatistical Concepts for Embedded Pragmatic Clinical Trials

June 19, 2020

Elizabeth L. Turner, PhD, Duke University
Patrick J. Heagerty, PhD, University of Washington
David M. Murray, PhD, National Institutes of Health

For the NIH Collaboratory Coordinating Center
Biostatistics and Study Design Core Working Group
Overview

• Focus of this talk: demystifying design-related issues for embedded pragmatic clinical trials (ePCTs)
• Context: NIH Collaboratory–funded studies
• Three kinds of randomized trials
  • Randomized controlled trial (RCT)
  • Cluster randomized trial (CRT)
    • Parallel vs stepped-wedge
  • Individually randomized group treatment (IRGT) trial
• How to select amongst these designs?
• Other brief topics: clustering, power, and analytical issues
### EXPERIMENTAL DESIGNS AND RANDOMIZATION SCHEMES

1. Introduction  
2. Statistical Design Considerations  
3. Cluster Randomized Trials  
4. Randomization Methods  
5. Choosing Between Cluster and Individual Randomization  
6. Alternative Cluster Randomized Designs  
7. Concealment and Blinding  
8. Designing to Avoid Identification Bias  
9. Additional Resources

### ANALYSIS PLAN

1. Introduction  
2. Intraclass Correlation  
3. Unequal Cluster Sizes  
4. Accounting for Residual Confounding in the Analysis  
5. Missing Data and Intention-to-Treat Analyses  
6. EHR Data Extraction  
7. Unanticipated Changes  
8. Case Study: STOP CRC Trial
NIH Collaboratory ePCT: SPOT

• Suicide Prevention Outreach Trial (SPOT)
• Approximately 16,000 patients across 4 clinical sites
• Three-arm RCT to evaluate 2 individual-level interventions vs usual care
  • Interventions
    • Skills training program
    • Care management program
  • Intervention contact mostly through EHR
    • Low risk of “contamination”
    • Individual-level randomization appropriate
  • Unit of randomization: patient

NIH Collaboratory ePCT: STOP CRC

- Strategies and Opportunities to Stop Colorectal Cancer in Priority Populations (STOP CRC)
- 40,000+ patients across 26 clinical sites
- Intervention
  - Health system–based program to improve CRC screening rates
  - Applied to clinical site → cluster randomization
- Unit of randomization: clinical site
- Two-arm cluster randomized trial (CRT)
  - Also referred to as a group-randomized or community randomized trial

Reasons to Randomize Clusters Instead of Individuals

• Intervention targets health care units rather than individuals
  • STOP CRC: clinic-based intervention to improve screening
• Intervention targeted at individual at risk of contamination
  • Intervention adopted by members of control arm
  • For example, physicians randomized to new educational program may share knowledge with control-arm physicians in their practice
  • Contamination reduces the observed treatment effect
• Logistically easier to implement intervention by cluster
STOP CRC Cluster Randomization

**Level 2:** Randomization at the level of the clinic (ie, cluster)

- **Intervention**
- **Factors related to uptake of screening**

**Screening**

**Level 1:** Individual-level outcomes nested within clinics
STOP CRC Cluster Randomization

Level 1: Individual-level outcomes nested within clinics

- Individual-level outcomes within same clinic expected to be correlated (i.e., to cluster)
STOP CRC Cluster Randomization

**Level 1:** Individual-level outcomes nested within clinics

- Individual-level outcomes within same clinic expected to be correlated (i.e., to *cluster*)
- Reduces power to detect treatment effect if same sample size used as under individual randomization
Understanding Outcome Clustering

- Consider 10 control-arm clinics (i.e., clusters)
- Each with 5 age-eligible patients: i.e., who are not up to date with colorectal cancer (CRC) screening
- Binary outcome: refused screening (Y/N)
Understanding Outcome Clustering: Complete Clustering

- Screened
- Not screened
Understanding Outcome Clustering:
Complete Clustering

>1 participant/clinic gives no more information than a single participant/clinic since every participant in a given clinic has the same outcome
Understanding Outcome Clustering: No Clustering

- Screened
- Not screened
Understanding Outcome Clustering: **No Clustering**

- **Screened**
- **Not screened**

20% uptake of CRC screening in each clinic
No structure by clinic; more like a random sample of eligible participants
Understanding Outcome Clustering: Some Clustering

- Screened
- Not screened
Understanding Outcome Clustering: Some Clustering

A more typical situation: proportion screened ranges from 0% - 80%

- Screened
- Not screened
Measure of Outcome Clustering: Intraclass Correlation Coefficient (ICC)

- Needed for study planning and power
- Most commonly used measure of clustering
- Ranges: 0-1; 0 = no clustering; 1 = complete clustering
- Typically < 0.2; commonly around 0.01 to 0.05
- Between-cluster outcome variance vs total outcome variance
Measure of Outcome Clustering: Intraclass Correlation Coefficient (ICC)

- Needed for study planning and power
- Most commonly used measure of clustering
- Ranges: 0-1; 0 = no clustering; 1 = complete clustering
- Typically < 0.2; commonly around 0.01 to 0.05
- Between-cluster outcome variance vs total outcome variance

\[
ICC = \frac{2^B}{2^B + 2^W} = \frac{2^B}{2^{Total}}
\]

Involves both *between-cluster* and *within-cluster* variance
In the Living Textbook: ICC Cheat Sheet

**ANALYSIS PLAN**

1. Introduction
2. Intraclass Correlation
3. Unequal Cluster Sizes
4. Accounting for Residual Confounding in the Analysis
5. Missing Data and Intention-to-Treat Analyses
6. EHR Data Extraction
7. Unanticipated Changes
8. Case Study: STOP CRC Trial

---

**Intraclass Correlation Coefficient Cheat Sheet**

**PURPOSE**

This document provides an introductory description of the intraclass correlation coefficient (ICC), a descriptive statistic that is important for the design and analysis of cluster-randomized trials. In a cluster-randomized trial, instead of being randomized by individual participant, the unit of randomization is a cluster, such as a group of participants being seen at a hospital, clinic, or primary care practice, although the outcomes may still be measured at an individual level.

**DEFINITION**

The intraclass correlation coefficient (ICC) is a descriptive statistic that describes the extent to which outcomes within each cluster are likely to be similar or not. Between different clusters, outcomes are likely to be different from each other. In ICC, the ICC is an important tool for cluster randomized pragmatic trials because this value helps determine from other clusters. The ICC is also an important tool for cluster randomized pragmatic trials because this value helps determine from other clusters. The ICC is also an important tool for cluster randomized pragmatic trials because this value helps determine from other clusters. The ICC for most pragmatic cluster-randomized trials is typically 0.1; commonly around 0.01 to 0.05.

**EXAMPLES**

In cluster-randomized trials, where groups of individuals are randomized to treatment arms, outcomes within clusters are highly correlated and when the magnitude of outcomes across clusters is quite different, then individuals within the same family are likely to have similar outcomes and the ICC will be larger. When this is the case, the data from one member of the cluster provides almost as much information as if all of the members are included. Hence, the effective sample size is closer to the number of clusters as opposed to the entire sample size of study participants.

To demonstrate why this is relevant, let's consider two examples:

1. In a dietary intake study the data from several members of the same family were likely to be very similar and would differ from those of other families. To reduce these effects, there might be little gain from sampling more than one member. On the other hand, if a cluster is an entire city and subjects within the city are randomly sampled, one might expect relatively little similarity from subject to subject relative to the rest of the sample. In this case, each individual subject would likely contribute "independent" information.

2. Suppose we have 6 providers, each with a eligible patients for a pragmatic, cluster-randomized trial. In this hypothetical case, the outcome is patient satisfaction rated on a scale from 1 to 10 with an outcome distribution as shown in Figure 1. One might expect that patients seen by a specific provider will have more similar levels of satisfaction to each other than to patients from other providers and that some providers will have consistently high patient satisfaction (e.g. provider 1) whereas others will have consistently low patient satisfaction (e.g. provider 2) whereas others will have consistently low patient satisfaction (e.g. provider 2) whereas others will have consistently low patient satisfaction (e.g. provider 2). This example is one of how outcomes within each cluster are likely to be similar. Thus, the ICC is high, and adding individuals to the cluster does not provide much additional information.
Accounting for Clustering Requires Larger Sample for Adequate Power

- Power and detectable difference is affected by...
  - Strength of the clustering effect (e.g., size of ICC)
  - Number of clusters
  - Number of patients per cluster
Impact of increasing # clusters

Example: CRT with ICC=0.1 at fixed alpha & power

Detectable difference (SD units) vs. # patients/cluster
Impact of increasing # clusters

Example: CRT with ICC=0.1 at fixed alpha & power

![Graph showing the impact of increasing number of clusters on detectable difference in SD units. The x-axis represents the number of patients per cluster, ranging from 0 to 350. The y-axis represents the detectable difference in SD units, ranging from 0 to 2.5. The graph includes lines for different numbers of clusters per arm: 2, 4, 8, 16, and 32.]
Impact of increasing # clusters

Example: CRT with ICC=0.1 at fixed alpha & power

Detectable difference (SD units)

Total # clusters = 4

# patients/cluster

Number of clusters per arm:
- 2
- 4
- 8
- 16
- 32
Impact of increasing # clusters

Example: CRT with ICC=0.1 at fixed alpha & power

- Detectable difference (SD units)
- Total # clusters = 4
- Total # clusters = 8
- # clusters per arm:
  - 2
  - 4
  - 8
  - 16
  - 32

# patients/cluster
Impact of increasing # clusters

Example: CRT with **ICC=0.1** at fixed alpha & power
Impact of increasing # clusters

Example: CRT with smaller ICC=0.01 at fixed alpha & power

![Graph showing the impact of increasing number of clusters on detectable difference (SD units) for different numbers of patients per cluster. The x-axis represents the number of patients per cluster, and the y-axis represents the detectable difference in SD units. The graph includes lines for different numbers of clusters per arm: 2, 4, 8, 16, and 32.](image)
Impact of increasing # clusters/groups

Example: CRT with even smaller ICC=0.001 at fixed alpha & power

Detectable difference (SD units) vs. # patients/cluster for different # clusters per arm:
- Red: 2 clusters
- Blue: 4 clusters
- Green: 8 clusters
- Pink: 16 clusters
- Black: 32 clusters
Accounting for Clustering in Design

- Power and sample size for CRT
  - Account for anticipated clustering
  - Inflate RCT sample size
  - Work with statistician to do correctly
- Use ICC for outcome
  - ICC often 0.01-0.05
  - STOP CRC: ICC = 0.03 for primary outcome
  - Depends on outcome and study characteristics
  - Different outcome = different ICC, even in same CRT
Estimating ICC to Plan Study

• How to get good estimate of ICC for a particular outcome?
  • Depends on outcome and study characteristics
  • CONSORT statement recommends ICC reported
  • Look at other articles with similar settings
  • Use available EHR data
• Be cautious when using pilot data from small study
  • ICC might have a wide confidence interval
NIH Collaboratory ePCT: LIRE

- Lumbar Imaging with Reporting of Epidemiology (LIRE)
- Goal: reduce unnecessary spine interventions by providing info on prevalence of normal findings
- Patients of 1700 PCPs across 100 clinics
- Clinic-level intervention → cluster randomization
- Unit of randomization: clinic
- Pragmatic trial
  - All clinics will eventually receive intervention
  - Stepped-wedge CRT

NIH Collaboratory ePCT: LIRE

Types of CRT Designs

Parallel

Stepped-wedge
Types of CRT Designs

Parallel  Stepped-wedge

Complete  Incomplete

In complete designs, measurements are taken from every cluster at every time point. In incomplete designs, some clusters do not provide measurements at all time points.
Types of CRT Designs

Examples with 8 clusters: 1-year intervention

- Control period
- Intervention period

Parallel design

Cluster 1

Cluster 8

Time since baseline

Types of CRT Designs

Examples with 8 clusters: 1-year intervention

Types of CRT Designs

Examples with 8 clusters: 1-year intervention

Types of CRT Designs

Examples with 8 clusters: 1-year intervention

Types of CRT Designs

Examples with 8 clusters: 1-year intervention

- **Parallel design**
  - Cluster 1
  - Cluster 8
  - Time since baseline: 0, 1

- **Complete stepped-wedge design**
  - Time since baseline: 0, 1, 2, 3, 4

- **Incomplete stepped-wedge design**
  - Time since baseline: 0, 1, 2, 3, 4
CRT Analysis: Treatment Effects

Estimated (primarily) using between-cluster information, i.e., \textit{vertical} information.

Parallel design

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time since baseline</td>
<td></td>
</tr>
</tbody>
</table>

Complete SW design

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time since baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Control period
- Intervention period
CRT Analysis: Treatment Effects

Estimated (primarily) using between-cluster information, i.e., \textit{vertical} information

Parallel design

Complete SW design

- Control period
- Intervention period
CRT Analysis: Treatment Effects

Estimated (primarily) using between-cluster information

Estimated using both vertical & horizontal (i.e., within-cluster) information

Parallel design

Complete SW design

<table>
<thead>
<tr>
<th>Control period</th>
<th>Intervention period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control period</td>
<td>Intervention period</td>
</tr>
</tbody>
</table>
CRT Analysis: Treatment Effects

Estimated (primarily) using between-cluster information, i.e., **vertical** information

Estimated using both **vertical** & **horizontal** (i.e., within-cluster) information

Parallel design

Complete SW design
CRT Analysis: Treatment Effects

Estimated (primarily) using between-cluster information, i.e., **vertical** information

Estimated using both **vertical** & **horizontal** (i.e., within-cluster) information

Parallel design

Control period

Intervention period
**CRT Analysis: Treatment Effects**

Estimated (primarily) using between-cluster information

Estimated using both **vertical** & **horizontal** (ie, within-cluster) information

**Parallel design**

**Complete SW design**

- **Control** period
- **Intervention** period
Choosing the Right Type of CRT

- Arguments **for** stepped-wedge CRT:
  - Cannot immediately implement intervention in 1/2 clusters
  - Pragmatic research: eventually implement in all clusters
  - Have few clusters and might gain power
Choosing the Right Type of CRT

• Arguments **for** stepped-wedge CRT:
  • Cannot immediately implement intervention in 1/2 clusters
  • Pragmatic research: eventually implement in all clusters
  • Have few clusters and might gain power

• Arguments **against** stepped-wedge CRT:
  • Risk confounding treatment effect with time effect
  • Risk of interruption or external events that could affect the outcome (eg, a pandemic!)
Recommendations for CRT Design

- Use a parallel CRT design if you can
- If stepped-wedge, plan for time effects in design & analysis
- Work with statistician to account for clustering in design and analysis of both designs
Choosing Study Design

Is there a strong rationale for randomizing groups/clusters rather than individuals to study conditions?
Choosing Study Design

Is there a strong rationale for randomizing groups/clusters rather than individuals to study conditions?

Yes
Choosing Study Design

Is there a strong rationale for randomizing groups/clusters rather than individuals to study conditions?

Examples with clinic/health-system-level interventions:

- STOP CRC colorectal cancer screening CRT
- LIRE lumbar imaging trial SW-CRT
Choosing Study Design

Is there a strong rationale for randomizing groups/clusters rather than individuals to study conditions?

Yes

Is there a strong rationale for rolling out the intervention to all clusters before the end of the trial?
Choosing Study Design

Is there a strong rationale for randomizing groups/clusters rather than individuals to study conditions?

- Yes
  - Is there a strong rationale for rolling out the intervention to all clusters before the end of the trial?
    - Yes
    - No
      - STOP CRC colorectal cancer screening CRT
      - CRT
Choosing Study Design

Is there a strong rationale for randomizing groups/clusters rather than individuals to study conditions?

Yes

Is there a strong rationale for rolling out the intervention to all clusters before the end of the trial?

Yes

LIRE lumbar imaging SW-CRT

SW-CRT
Choosing Study Design

Is there a strong rationale for randomizing groups/clusters rather than individuals to study conditions?

No
Choosing Study Design

Is there a strong rationale for randomizing groups/clusters rather than individuals to study conditions?

No

Examples with individual-level randomization:
- SPOT suicide prevention RCT
- OPTIMUM mindfulness for back-pain RCT
Choosing Study Design

Is there a strong rationale for randomizing groups/clusters rather than individuals to study conditions?

No

Do participants receive their treatment in a group format or from a shared interventionist?
Choosing Study Design

- Is there a strong rationale for randomizing groups/clusters rather than individuals to study conditions? **No**
- Do participants receive their treatment in a group format or from a shared interventionist? **No**
- **SPOT suicide prevention RCT**
  - Intervention is targeted at the individual
Choosing Study Design

Is there a strong rationale for randomizing groups/clusters rather than individuals to study conditions?

No

Do participants receive their treatment in a group format or from a shared interventionist?

Yes

Individually-randomized group treatment (IRGT) trial

Clustering must be accounted for in both design and analysis
Choosing Study Design

Is there a strong rationale for randomizing groups/clusters rather than individuals to study conditions?

No

Do participants receive their treatment in a group format or from a shared interventionist?

No

Optimum mindfulness for back-pain RCT

Intervention is group-based

Yes

Clustering must be accounted for in both design and analysis

Individually-randomized group treatment (IRGT) trial
NIH Collaboratory ePCT: OPTIMUM

- OPTIMUM: optimizing pain treatment in medical settings using group-based mindfulness
- ~450 patients across 3 clinical sites
- Two-arm RCT
  - Intervention vs usual care
- Unit of randomization: individual
- Group-based intervention
  - Clustering of outcomes in intervention arm
  - Must be accounted for in both design and analysis
  - “Individually randomized group treatment (IRGT) trial”
Choosing Study Design

Is there a strong rationale for randomizing groups/clusters rather than individuals to study conditions?

No

Do participants receive their treatment in a group format or from a shared interventionist?

No

RCT

Yes

IRGT trial

Is there a strong rationale for rolling out the intervention to all clusters before the end of the trial?

Yes

SW-GRT

No

GRT

See Figure: Murray DM, Taljaard M, Turner EL, George SM, Ann Rev Pub Health 2020. 41:1-19
Choosing Study Design

- Is there a strong rationale for randomizing groups/clusters rather than individuals to study conditions?
  - No
  - Do participants receive their treatment in a group format or from a shared interventionist?
    - No
      - RCT
    - Yes
      - Is there a strong rationale for rolling out the intervention to all clusters before the end of the trial?
        - Yes
          - SW-GRT
        - No
          - GRT

Clustering must be accounted for in both design and analysis.

See Figure: Murray DM, Taljaard M, Turner EL, George SM, Ann Rev Pub Health 2020. 41:1-19
Important Things to Know

• Question drives design, design drives analysis
• Randomization
  • Individual-level preferred for statistical reasons
  • But cluster randomization often needed
• Account for clustering in design and analysis of:
  • CRT
  • IRGT trial
• Good design is difficult but critical
  • Need input from diverse team, including statistician
  • Analysis may not be able to overcome design flaws
Important Things to Do

- Focus on the research question
- Select design features with analysis in mind
- Collaborate early with a statistician
- Choose individual randomization, but only if possible
- Weigh statistical choices vs implementation challenges
- Write and publish a protocol paper
In the Living Textbook

**DESIGN**

**EXPERIMENTAL DESIGNS AND RANDOMIZATION SCHEMES**

1. Introduction
2. Statistical Design Considerations
3. Cluster Randomized Trials
4. Randomization Methods
5. Choosing Between Cluster and Individual Randomization
6. Alternative Cluster Randomized Designs
7. Concealment and Blinding
8. Designing to Avoid Identification Bias
9. Additional Resources

**ANALYSIS PLAN**

1. Introduction
2. Intraclass Correlation
3. Unequal Cluster Sizes
4. Accounting for Residual Confounding in the Analysis
5. Missing Data and Intention-to-Treat Analyses
6. EHR Data Extraction
7. Unanticipated Changes
8. Case Study: STOP CRC Trial
Summary

• Focus of this talk: demystifying design-related issues for embedded pragmatic clinical trials (ePCTs)
• Context: NIH Collaboratory–funded studies
• Three kinds of randomized trials
  • Randomized controlled trial (RCT)
  • Cluster randomized trial (CRT)
    • Parallel vs stepped-wedge
  • Individually randomized group treatment (IRGT) trial
• How to select amongst these designs?
• Other brief topics: clustering, power, and analytical issues
Design and Analysis Methods

NIH Resources

• Pragmatic and Group-Randomized Trials in Public Health and Medicine
  • [https://prevention.nih.gov/grt](https://prevention.nih.gov/grt)
  • 7-part online course on GRTs and IRGTs

• Mind the Gap Webinars
    • Analytic methods for SW-GRTs (Fan Li, July 14, 2020)
    • SW-GRTs for Disease Prevention Research (Monica Taljaard, July 11, 2018)
    • Design and Analysis of IRGTs in Public Health (Sherri Pals, April 24, 2017)
    • Research Methods Resources for Clinical Trials Involving Groups or Clusters (David Murray, December 13, 2017)

• Research Methods Resources Website
  • [https://researchmethodsresources.nih.gov/](https://researchmethodsresources.nih.gov/)
  • Material on GRTs and IRGTs and a sample size calculator for GRTs.
Demystifying Biostatistical Concepts for Embedded Pragmatic Clinical Trials

June 19, 2020

Elizabeth L. Turner, PhD, Duke University
Patrick J. Heagerty, PhD, University of Washington
David M. Murray, PhD, National Institutes of Health

Thank you
Any questions or comments?