Implications of RCT Design on Sample Size Requirements

Andrew C. Leon, Ph.D.
Weill Medical College of Cornell University
New York City, New York
Disclosures

Data and Safety Monitoring Boards
  Pfizer, Organon, and Vanda

Consultant/Advisor
  Eli Lilly, MedAvante, FDA, and NIMH
Outline

- Reliability and Sample Size Requirements
- Multiple Endpoints and Sample Size Requirements
Goals of Randomized Controlled Clinical Trial Design

Minimize bias in estimate of treatment effect

Maintain type I error level

Sufficient statistical power

Feasible and Applicable

Leon et al., Biological Psychiatry, 2006; 59:1001-1005.
Features of RCT Design

Randomized group assignment

Double-blinded assessments

Control or comparison groups
Problems of Unreliability and Multiplicity

Unreliability introduces bias

Multiplicity inflates type I error

Unreliability reduces statistical power

Unreliability reduces RCT feasibility
RCT Design: Measurement

Choice of assessments

Feasibility of assessment

Number of primary efficacy measures

** Mode of Assessment and Intensity of Training typically overlooked -- particularly their bearing on sample size requirements**
“Avoid the use of excessive or inadequate numbers of research subjects by making informed recommendations for study size.”

www.amstat.org/profession/ethicalstatistics.html
Sample Size Determination

Informed recommendations for study size for an RCT, are guided by statistical power analyses.
Sample Size Determination

Four components of power analysis

- $\alpha$ (0.05; Except with Co-primaries)
- Power (0.80 or 0.90)
- Sample size
- Population effect size ($d$)

Given any 3, the 4th can be determined.

Typically manipulate power by changing N.

Alternatively, consider reducing unreliability, which will change the effect size.
Between Group Effect Size for a t-test

\[ d = \frac{\bar{X}_1 - \bar{X}_2}{S} \]

Group difference in standard deviation units
RCT Design Stage: Pilot Data to Estimate the Effect Size?

Empirical Estimates of Cohen's $d$ with 95% CI (population delta=.50)

Simulation Study: 10,000 simulated data sets for each combination of $d$ and $N$

95% CI: $d +/- [t * 2 / \sqrt{N}]$  
(Kraemer, AGP 2006 63:484-9)
### Sample Size Determination:
**Design to Detect a Clinically Meaningful Difference**

<table>
<thead>
<tr>
<th>$d$</th>
<th>N/group (from Cohen’s Tables)</th>
</tr>
</thead>
<tbody>
<tr>
<td>small (.20)</td>
<td>393</td>
</tr>
<tr>
<td>medium (.50)</td>
<td>64</td>
</tr>
<tr>
<td>large (.80)</td>
<td>26</td>
</tr>
</tbody>
</table>

As a benchmark:

About 200 placebo-controlled RCTs of fluoxetine for MDD: $d = .38$

Alternative approach: $N/group = 16/d^2$

- e.g., $16 / .5^2 = 64 / group$ (Lehr, Stat in Med, 1992)
Effect Size for a t-test

\[ d = \frac{\bar{X}_1 - \bar{X}_2}{S} \]

Group difference in standard deviation units
Hypothetical PANSS Ratings at Baseline

Sources of variability at baseline: true differences and measurement error
Hypothetical PANSS Ratings at Baseline: Two Assessment Methods

Equal Means, but $S_B = S_A/2$

\[ d = \frac{\bar{X}_1 - \bar{X}_2}{S} \]
As reliability of assessment increases:
(New scale, Better training, Novel modality)

The within-group variability decreases.

The between-group effect size increases.

Sample size requirements decrease.

Leon AC, Marzuk PM, Portera L. Arch Gen Psychiatry 1995;52:867-871.
Design to Evaluate New Assessment Method
(2 x 2 factorial RCT)

Randomize subjects to:

Active vs. Control

Assessment Method: A vs. B

\( H_0: \text{Active}_A - \text{Control}_A = \text{Active}_B - \text{Control}_B \)

*Treatment by Method interaction*
Placebo Responder Evaluation using Comprehensive Investigation of Symptoms and EEG (PRECISE):

- 2 academic sites
- Allocation ratio 3:1 (Placebo:Active)
- Inclusion: HAMD > 16
- 5 weeks double-blind treatment
- Site-based and Central raters
PRECISE Eligibility: HAMD > 16

Site Ratings

Central Ratings

53/62 (85%)

35/62 (56%)
## PRECISE: Reliability of Site and Centralized Raters

### Internal Consistency Reliability: Cronbach’s Coefficient alpha

<table>
<thead>
<tr>
<th></th>
<th><strong>Baseline</strong></th>
<th></th>
<th><strong>Endpoint</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>N=40</strong></td>
<td><strong>N=34</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central Raters</td>
<td>.68</td>
<td>.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Site Raters</td>
<td>.33</td>
<td>.82</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Contrast Site and Central Ratings over Time

#### Mean HAMD Score By Visit: PLACEBO ONLY

<table>
<thead>
<tr>
<th></th>
<th>Baseline (n=33)</th>
<th>Endpoint (n=27)</th>
<th>Pre-Post Change (N=27)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central Raters</td>
<td>17.2 (5.6)</td>
<td>13.4 (6.9)</td>
<td>3.7</td>
</tr>
<tr>
<td>Site Raters</td>
<td>20.4 (3.2)</td>
<td>13.1 (6.7)</td>
<td>7.6</td>
</tr>
<tr>
<td>Δ</td>
<td>-3.2 (4.1)</td>
<td>0.3 (5.2)</td>
<td>-3.9</td>
</tr>
<tr>
<td>t</td>
<td>4.54</td>
<td>-0.33</td>
<td>3.89</td>
</tr>
<tr>
<td>P value</td>
<td>&lt;.001</td>
<td>.741</td>
<td>.001</td>
</tr>
</tbody>
</table>
Response: 50% HAMD reduction

<table>
<thead>
<tr>
<th>Placebo Response (N=27)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central Raters</td>
</tr>
<tr>
<td>22%</td>
</tr>
<tr>
<td>Site Raters</td>
</tr>
<tr>
<td>30%</td>
</tr>
</tbody>
</table>
Placebo Response Rates: Implications for Sample Size Requirements (per group)*

*For power of .80 using $\chi^2$ test with 2-tailed alpha=.05
Central Raters - RCT for Schizophrenia

Study Design

- 289 acutely psychotic, hospitalized patients
- Moderate to severely ill (70 ≤ PANSS ≤ 120)
- 35 sites
- 6 weeks of treatment
- Active comparator vs. 2 inv. doses vs. placebo
- Central Ratings were the primary outcome measure
- Sponsor only allowed publication of Central Ratings for comparator and placebo cells
Centralized Raters’ Score Distribution: Screen

Screening Visit: All Subjects (PANSS)
Central Raters in Schizophrenia: Results

PANSS Means

Mixed Model p = .022

<table>
<thead>
<tr>
<th></th>
<th>PL 68</th>
<th>58</th>
<th>46</th>
<th>38</th>
<th>30</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLZ</td>
<td>68</td>
<td>58</td>
<td>48</td>
<td>43</td>
<td>39</td>
<td>36</td>
</tr>
</tbody>
</table>
Recommendations

Improve the assessment process with *More Reliable Methods of Assessment*

A Reduction in Unreliability translates into:

- Reduced sample size requirements
- Reduced risks to human subjects
- Reduced RCT study time
- Reduced RCT costs
Multiple Endpoints and Sample Size Requirements

Multiple endpoints increase:

- research costs
- study duration
- N exposed to risks
Randomized Clinical Trial Design

Tension between
   Falsely concluding that an ineffective agent is efficacious
     *Type I error*

   Failing to conclude that an effective agent works
     *Type II error*
Simulation Study: Type I Error

N=100/group per response rate

10,000 $\chi^2$ tests/ response rate
“It may sometimes be desirable to use more than one primary variable

... the method of controlling type I error should be given in the protocol.”

**Multiple outcomes**: MATRICS battery
Bonferroni Adjustment

* Partitions the $\alpha=0.05$ among the $k$ tests

$$\alpha/k, \text{ for } k=1,2,3 \text{ endpoints: } \alpha^* = .05, .025, .0167...$$

* Sets an upper limit on *Experimentwise* Type I error ($\alpha_{EW}$)
Concerns about Bonferroni Adjustment

Does not account for correlations between endpoints.

Reduced statistical power – can lead to false negative findings.
Multiplicity-Adjusted Sample Sizes*

- Maintain statistical power if sample size estimates are based on adjusted alpha level (at design stage)

- Sample Size Requirements Increase with the Number of Tests

- Must increase N by about 20% for 2 tests; 30% for 3 tests.

<table>
<thead>
<tr>
<th># tests</th>
<th>adjusted _</th>
<th>d=0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.050</td>
<td>64</td>
</tr>
<tr>
<td>2</td>
<td>0.025</td>
<td>78</td>
</tr>
<tr>
<td>3</td>
<td>0.017</td>
<td>86</td>
</tr>
<tr>
<td>4</td>
<td>0.013</td>
<td>91</td>
</tr>
<tr>
<td>5</td>
<td>0.010</td>
<td>96</td>
</tr>
</tbody>
</table>

Assume: 2-tailed t-test, *power=0.80
(Leon, JCP, 2004)
Alternatives to Bonferroni Adjustment

Hochberg’s Sequentially-Rejective Tests

Each successively smaller p-value has a more rigorous alpha threshold.

<table>
<thead>
<tr>
<th>test #</th>
<th>Hochberg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0500</td>
</tr>
<tr>
<td>2</td>
<td>0.0250</td>
</tr>
<tr>
<td>3</td>
<td>0.0167</td>
</tr>
<tr>
<td>4</td>
<td>0.0125</td>
</tr>
<tr>
<td>5</td>
<td>0.0100</td>
</tr>
</tbody>
</table>

James adjustment  (Stat Med, 1991)

Incorporates correlations among endpoints
Simulation Studies
Adjustment Strategies for Multiple $\chi^2$ Tests: Type I Error

Endpoint rates = 30% vs. 30%; and $k = 3$

10,000 Simulated data sets per correlation.

Leon & Heo, J Biopharm Stat, 2006
Power Relative to Bonferroni

Power of 1 or more significant result.  
N/group=152  
10,000 simulated data sets/correlation. 

Endpoint rates of 25% vs. 40%; $k = 3$; 

Leon & Heo, Stat in Med, 2007
Pre-specify one primary efficacy measure.

• If multiple measures are absolutely necessary, pre-specify alpha adjustment strategy
  • Hochberg (r< 0.50) or James (r>0.50)

• Estimate sample size using adjusted alpha

• Multiple endpoints increase required sample size:
  • research costs
  • study duration
  • N exposed to risks