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

American Statistical Association (1999)

Guidelines for Statistical Practice
from the Committee on Professional Ethics

"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

- α (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.

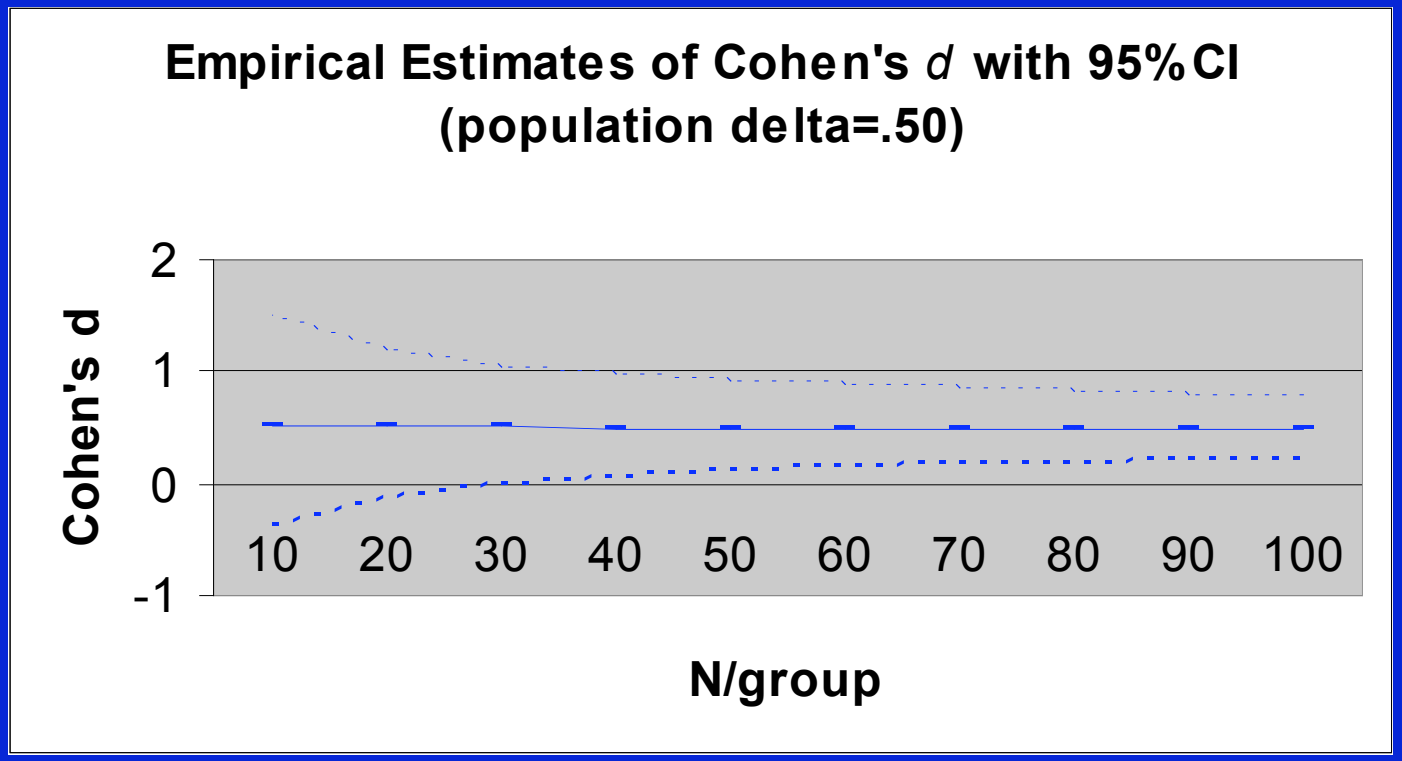
Alternatively, **consider reducing unreliability,** which will change the effect size.

Between Group Effect Size for a t-test

$$d = \frac{\overline{X}_1 - \overline{X}_2}{s}$$

Group difference in standard deviation units

RCT Design Stage:
Pilot Data to Estimate the Effect Size?



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

95% CI: $d \pm [t * 2/\sqrt{N}]$ (Kraemer, AGP 2006 63:484-9)

Sample Size Determination:

Design to Detect a Clinically Meaningful Difference

d	**N/group** (from Cohen's Tables)
small (.20)	393
medium (.50)	64
large (.80)	26

As a benchmark:

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

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

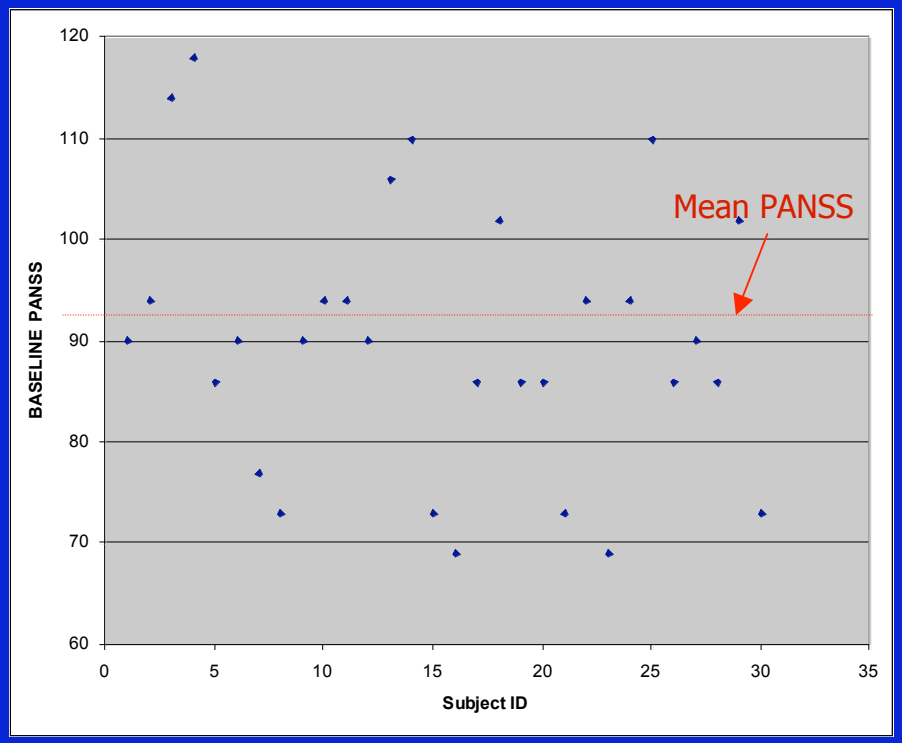
e.g., $16 / .5^2 = 64$ /group (Lehr, Stat in Med, 1992)

Effect Size for a t-test

$$d = \frac{\overline{X}_1 - \overline{X}_2}{s}$$

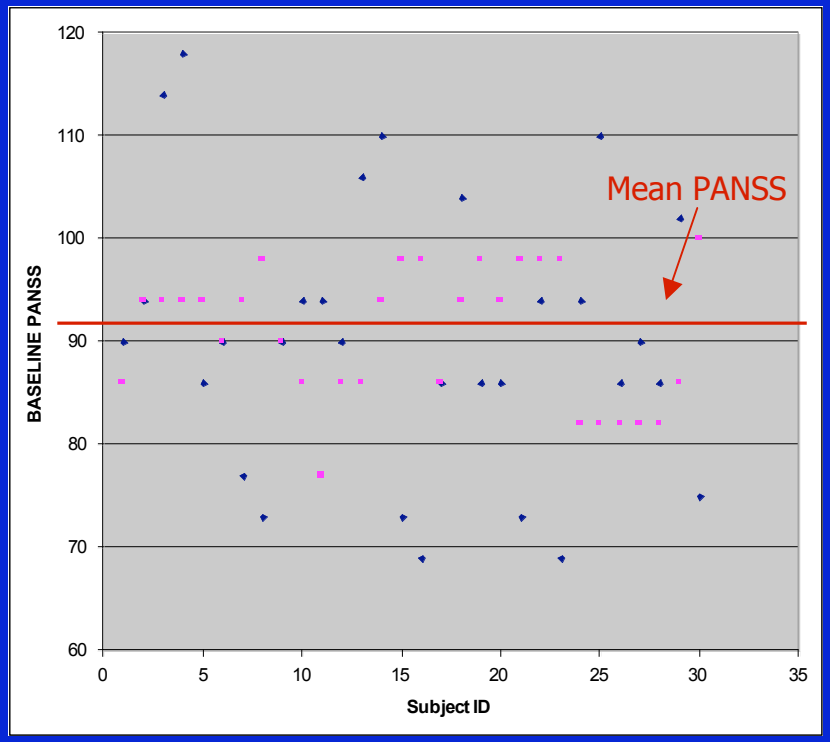
Group difference in standard deviation units

Hypothetical PANSS Ratings at Baseline



Sources of variability at baseline: true differences and measurement error

Hypothetical PANSS Ratings at Baseline: Two Assessment Methods



$$d = \frac{\overline{X}_1 - \overline{X}_2}{s}$$

Equal Means, but $S_B = S_A/2$

More Reliable Assessment Procedures Reduce Sample Size Requirements

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

Method	Treatment	
	Active	Control
A		
B		

H_0: $Active_A$ - $Control_A$ = $Active_B$ - $Control_B$

Treatment by Method interaction

Study Compared Site and Central Raters

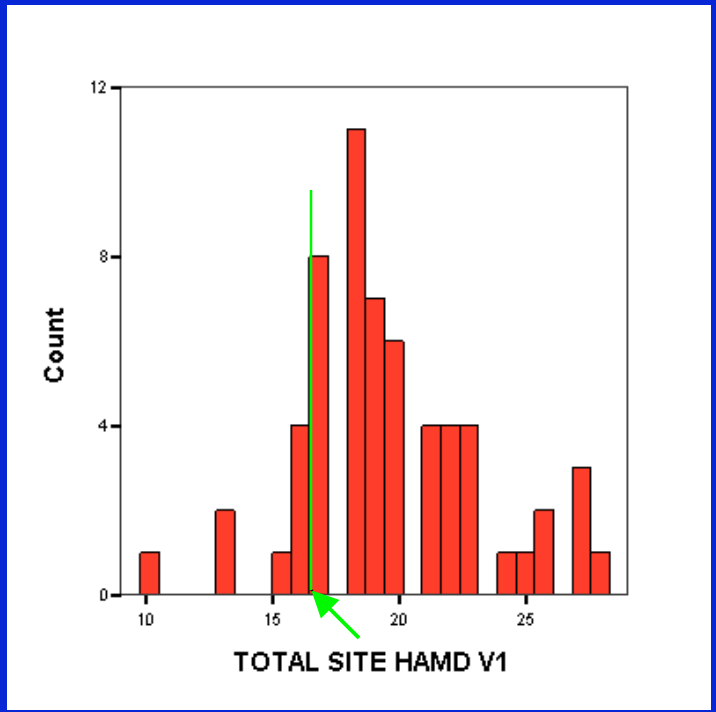
Placebo Responder Evaluation using Comprehensive Investigation of Symptoms and EEG (PRECISE):

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

Kobak, DeBrota, Engelhardt, Williams. Site vs. Centralized Raters in MDD RCT. NCDEU, 2006

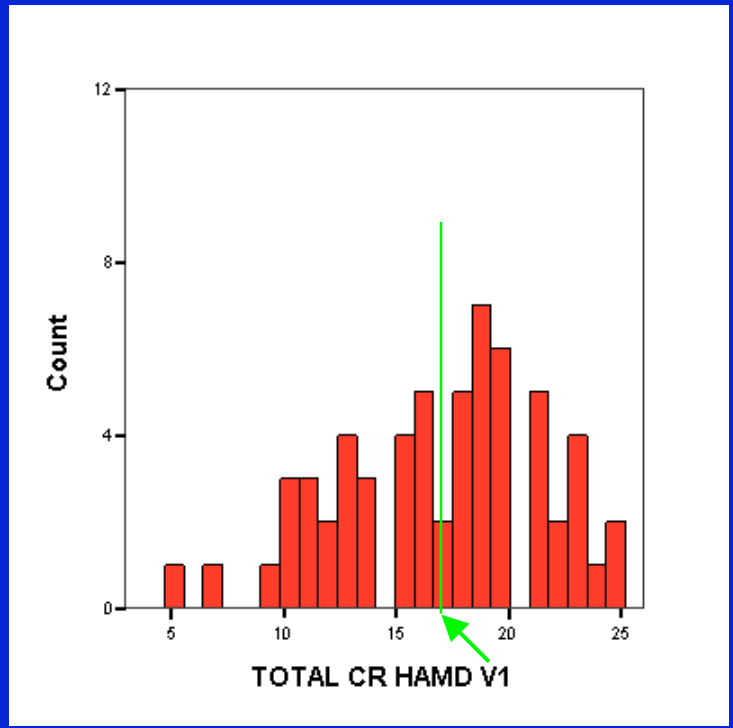
PRECISE Eligibility: HAMD > 16

Site Ratings



53/62 (85%)

Central Ratings



35/62 (56%)

PRECISE: Reliability of Site and Centralized Raters

Internal Consistency Reliability:
Cronbach's Coefficient alpha

	Baseline N=40		**Endpoint** N=34
Central Raters	.68		.81
Site Raters	.33		.82

Contrast Site and Central Ratings over Time

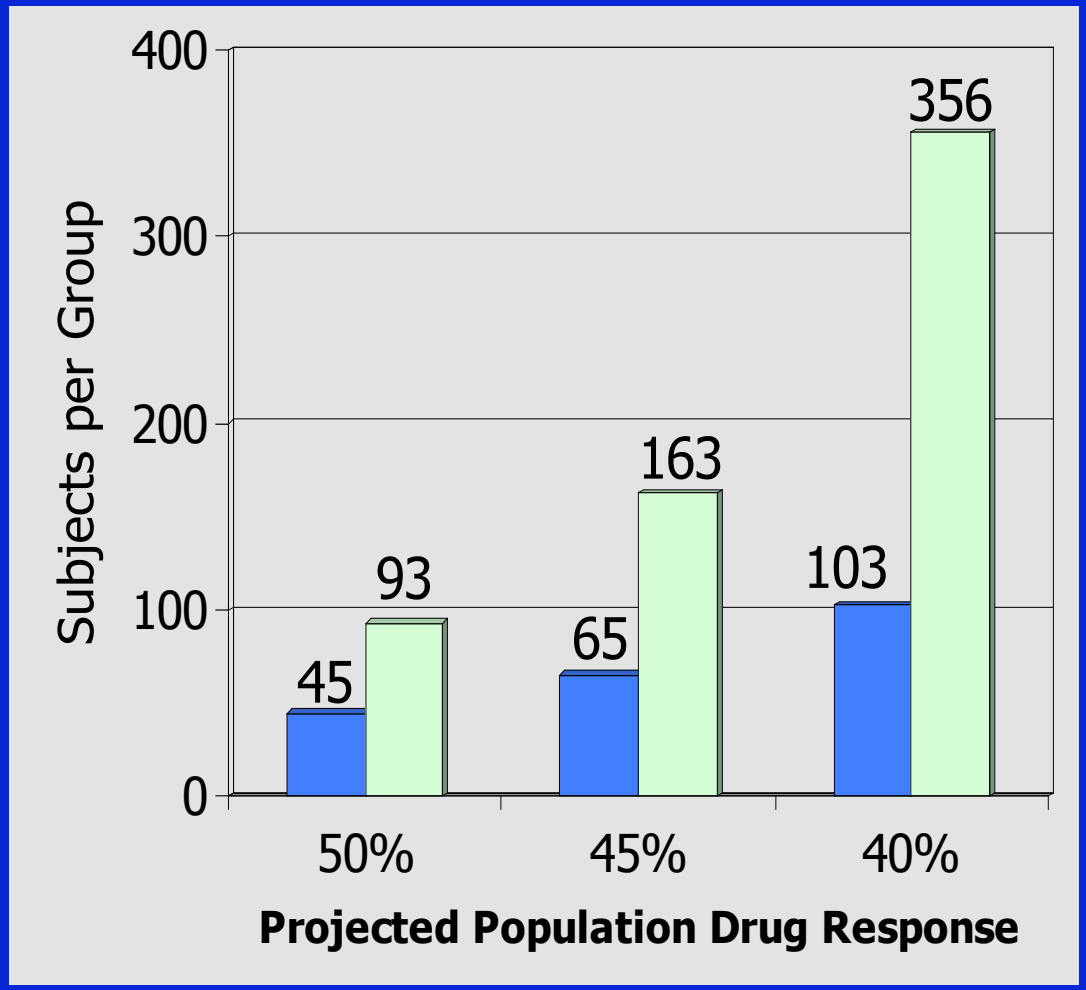
Mean HAMD Score By Visit: PLACEBO ONLY				
		Baseline (n=33)	**Endpoint (n=27)**	**Pre-Post Change (N=27)**
Central Raters		17.2 (5.6)	13.4 (6.9)	3.7
Site Raters		20.4 (3.2)	13.1 (6.7)	7.6
Δ		-3.2 (4.1)	0.3 (5.2)	-3.9
t		4.54	-0.33	3.89
P value		<.001	.741	.001

PRECISE Study: Placebo Responder Rates

Response: 50% HAMD reduction

	Placebo Response (N=27)
Central Raters	22%
Site Raters	30%

Placebo Response Rates: Implications for Sample Size Requirements (per group)*



*For power of .80 using χ^2 test with 2-tailed alpha=.05

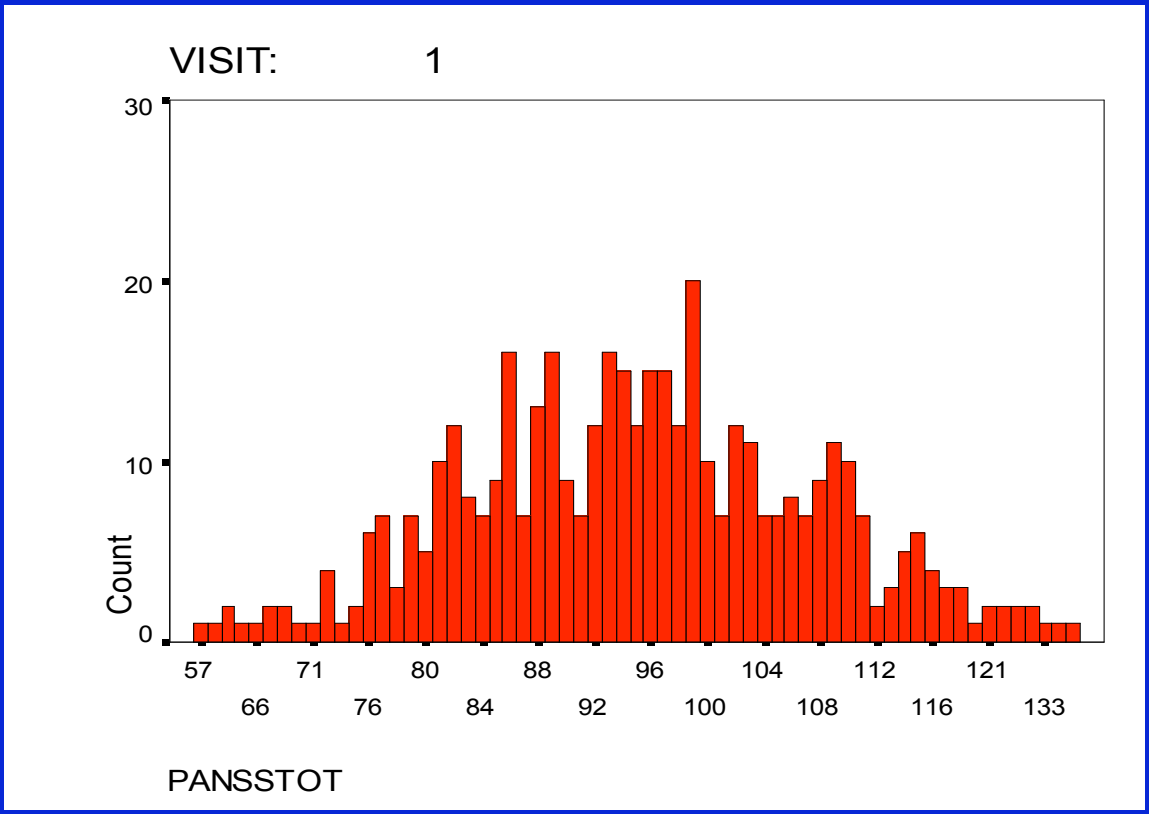
Central Raters - RCT for Schizophrenia

Study Design

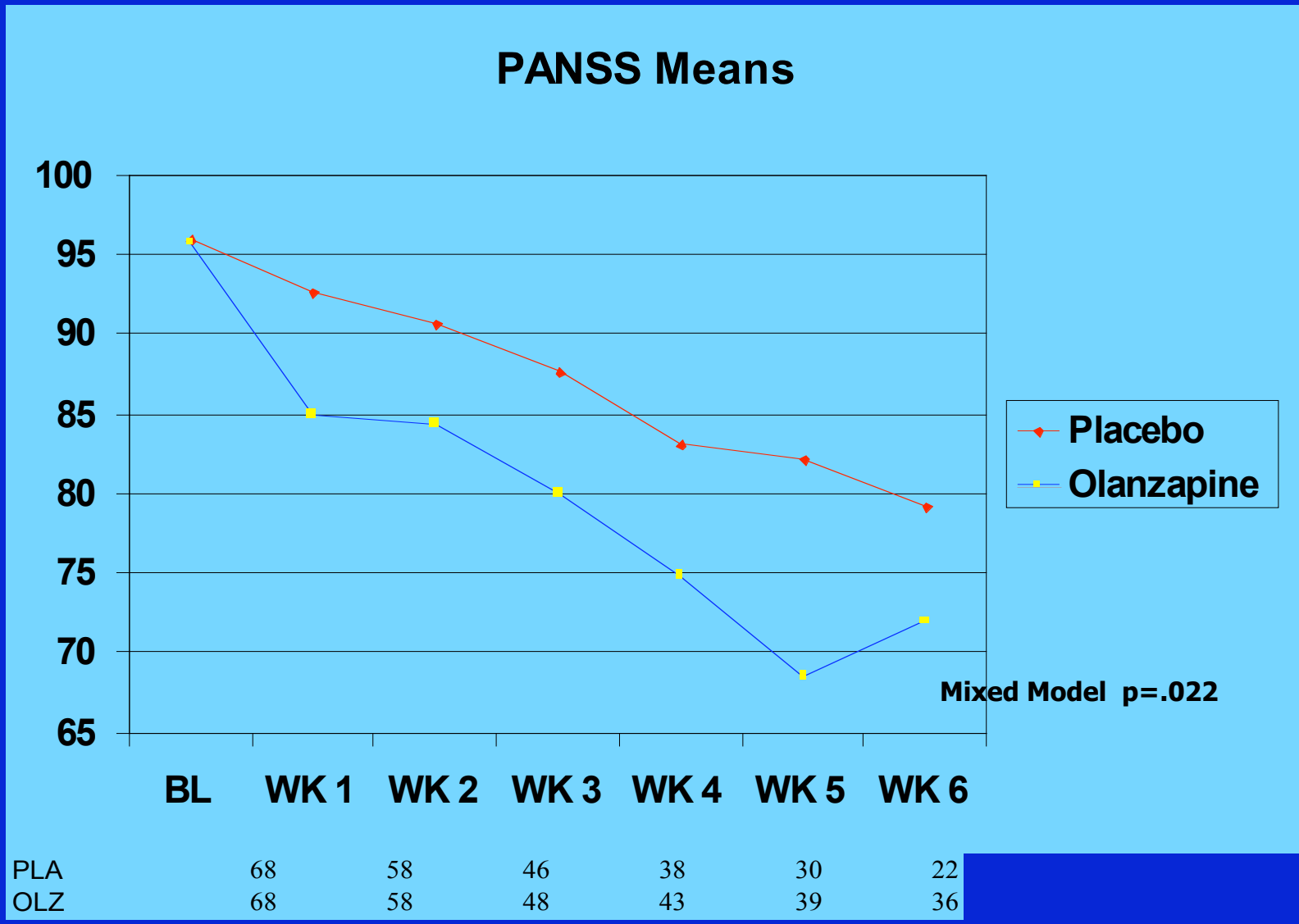
- 289 acutely psychotic, hospitalized patients
- Moderate to severely ill ($70 \leq PANSS \leq 120$)
- 35 sites
- 6 weeks of treatment
- Active comparator vs. 2 inv. doses vs. placebo
- Central Ratings were the primary outcome measure
- Sponsor only allowed publication of Central Ratings for comparator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

Centralized Raters' Score Distribution: Screen

Screening Visit: All Subjects (PANSS)



Central Raters in Schizophrenia: Results



PLA	68	58	46	38	30	22
OLZ	68	58	48	43	39	36

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

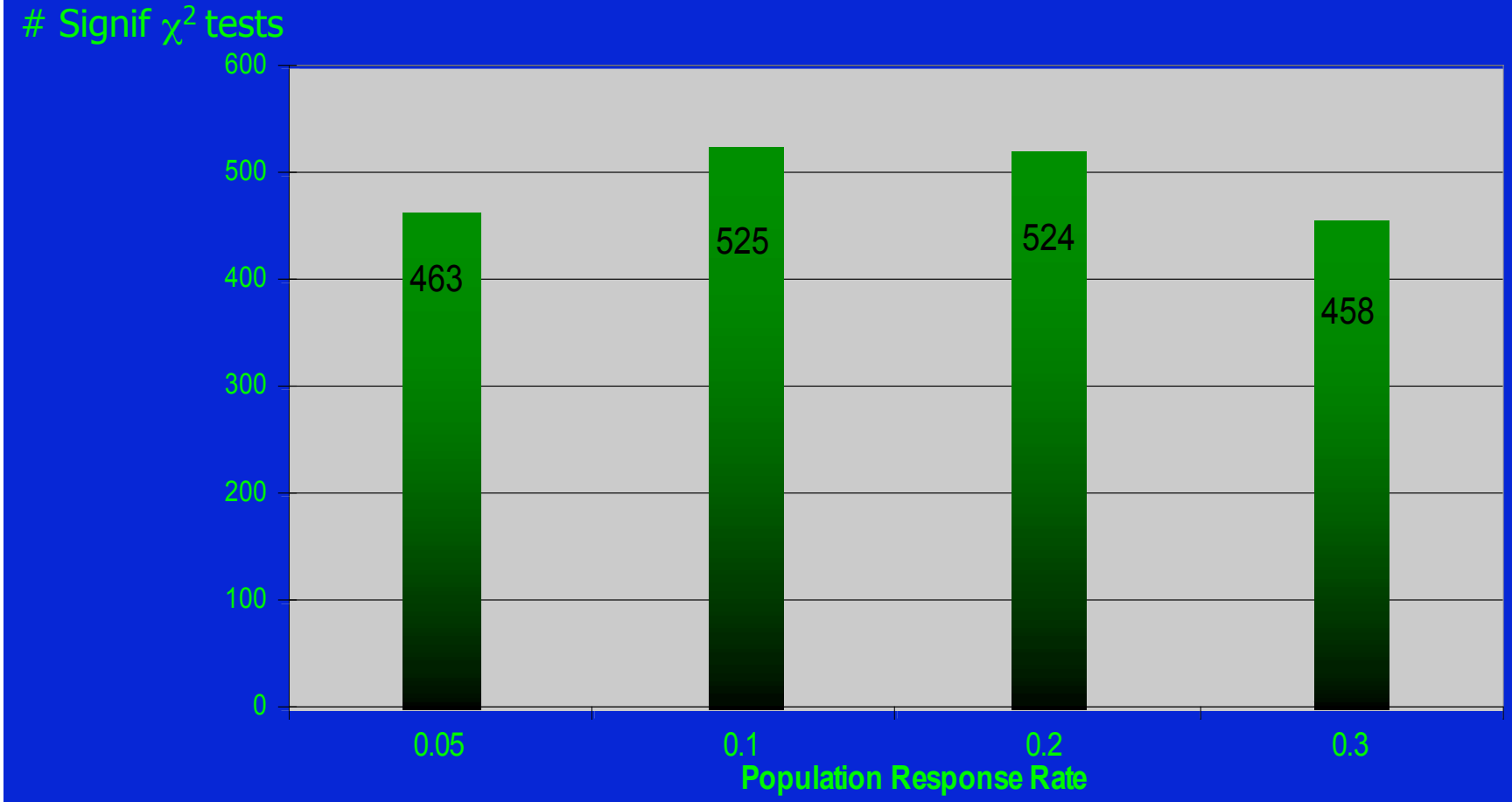
Randomized Clinical Trial Design

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 χ^2 tests/ response rate

FDA / ICH Guidance for Industry (E9):
Statistical Principles for Clinical Trials

"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


1892-1960

* Partitions the α=0.05 among the *k* tests

  α/k , for *k =1,2,3* endpoints: α^* = .05, .025, .0167…

* Sets an upper limit on *Experimentwise* Type I error (α_{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.

# tests	adjusted _		*d*=0.50
1	0.050		64
2	0.025		78
3	0.017		86
4	0.013		91
5	0.010		96

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.

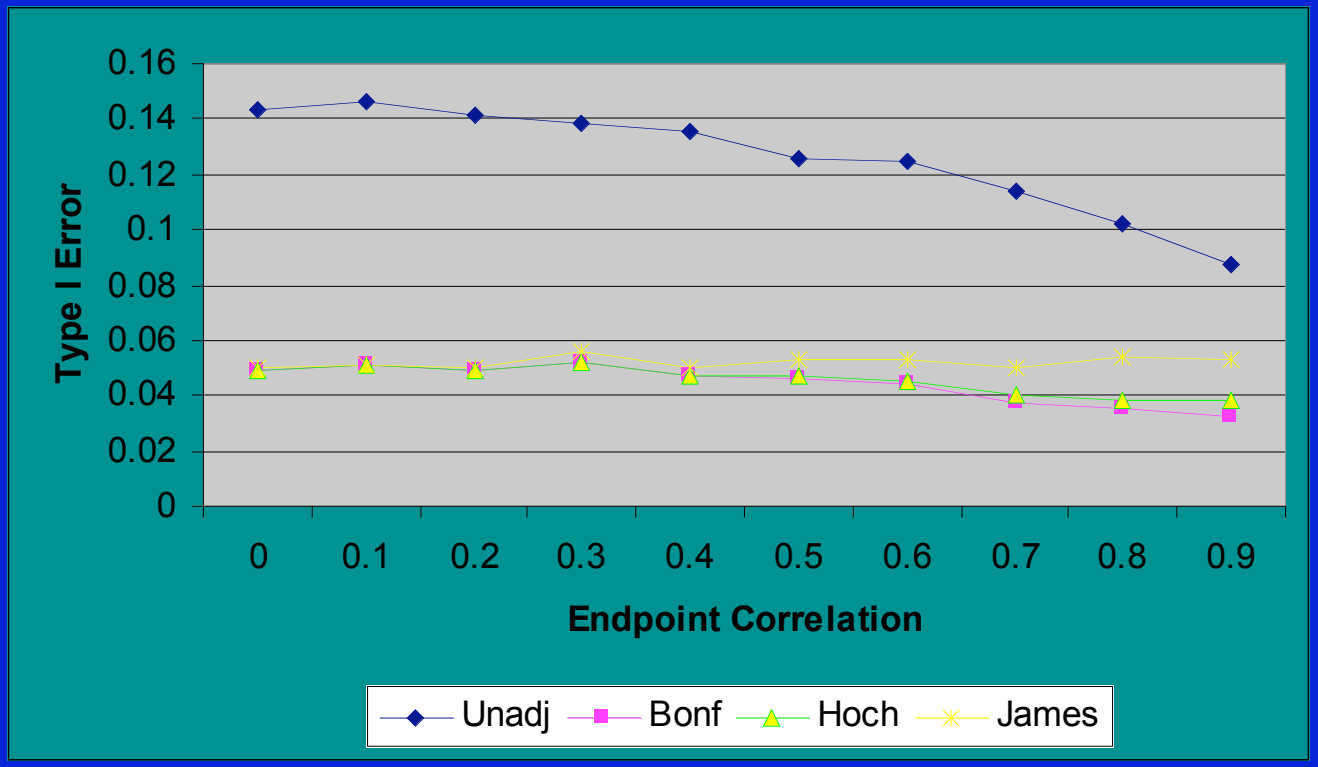
test #	*Hochberg*
1	0.0500
2	0.0250
3	0.0167
4	0.0125
5	0.0100

James adjustment (Stat Med, 1991)

Incorporates correlations among endpoints

Simulation Studies

Adjustment Strategies for Multiple χ^2 Tests: Type I Error

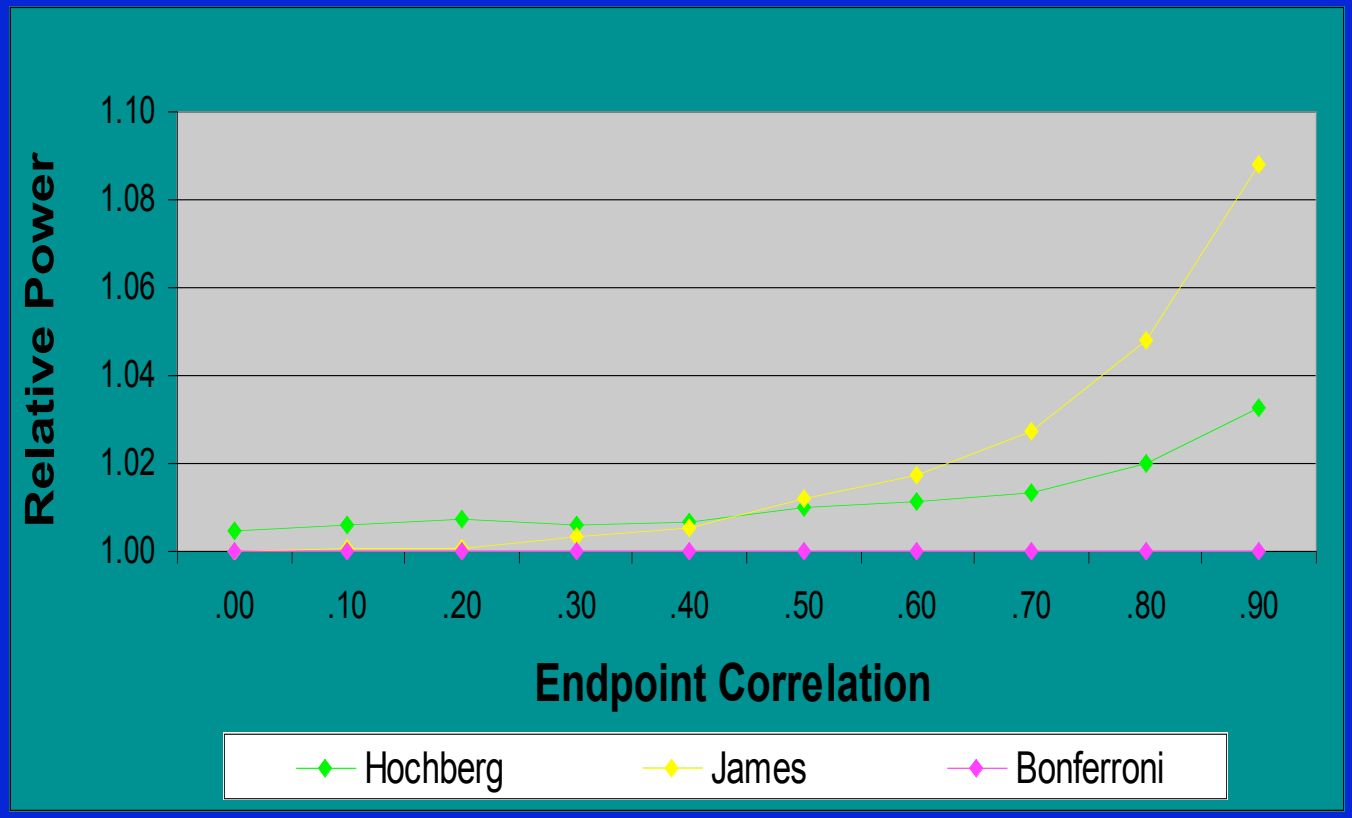


Endpoint rates = 30% vs. 30%; and $k = 3$

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

Multiplicity: Recommendations

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