Statistical Issues in RCTs with an Active Comparator

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

Disclosures

Past 12 months

Data and Safety Monitoring Boards

AstraZeneca, Dainippon Sumitomo Pharma America, Pfizer, Vanda (2006-07)

Consultant/Advisor

Cyberonics, FDA, MedAvante, NIMH, and Takeda

Outline

- Superiority vs. Noninferiority
- Required sample sizes
- Choice of non-inferiority margin
- Unintended consequence of active comparator trials

Three Arm RCT

Investigational vs. Placebo vs. Active Comparator

Efficacy

Investigational vs. Placebo

Assay sensitivity

Active Comparator vs. Placebo

“Safety”

Investigational vs. Active Comparator

Is Investigational worse than Active Comparator?

Assay Sensitivity

2 Arm study

H_0 : Investigational = Active Comparator

If H_0 is not rejected, *No difference* can mean either:

- *Both* treatments are effective *or*
- *Neither* treatment is effective

3 Arm study

Placebo provides a context to test assay sensitivity.

- Was RCT designed and implemented so differences between effective and ineffective agents would be detected?

Statistical Power in Three Arm RCT

Sample sizes driven by comparison with smallest effect

Efficacy
Investigational vs. Placebo

Assay sensitivity
Active Comparator vs. Placebo

"Safety"
Investigational vs. Active Comparator

Sample Size Determination

4 components of power analysis in superiority RCT

- α **(0.05 - Except with Co-primaries)**
- power (0.80 or 0.90)
- Sample size
- Population effect size (*d*)

Effect Size for a t-test

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

Cohen's *d:* Group difference in standard deviation units

Sample Size Determination:

Design to Detect a Clinically Meaningful Difference

d	N/group* (from Cohen 1992)
small (.20)	393
medium (.50)	64
large (.80)	26

Meta-analysis of 38 placebo-controlled RCTs of SGAs for schiz (Leucht, 2008): $d = 0.51$ → **64 subjects/group**

N/grp = 16/d² (Lehr, 1992): 16/.4² = 100/group

*** Assume 80% power for t-test with 2-tailed alpha of .05**

Designed to Detect a Clinically Meaningful Investigational-Placebo Difference (*d*=.40)

Power for Investigational vs. Active Comparison

With100/group power is:

80% for *d*=.40
69% for *d*=.35
56% for *d*=.30
42% for *d*=.25
29% for *d*=.20

Superiority trial must be designed with power to detect Investigational vs. Active difference.

Non-significant Superiority Result

A negative superiority trial does not demonstrate equivalence or non-inferiority.

A study can reject the null hypothesis. We cannot accept the null.

Non-significant result could stem from problems with design or implementation of study.

Three Arm RCT

Investigational vs. Placebo vs. Active Comparator

Superiority comparison

Design to detect smallest clinically meaningful difference

- Investigational vs. Placebo
- Active Comparator vs. Placebo

Non-inferiority comparison

Is Investigational worse than Active Comparator?

Protocol: Margin of non-inferiority (δ)

- Clinically acceptable difference

Equivalence vs. Non-inferiority

- **Equivalence trials** objective: To show that the effects differ by no more than a specified amount (equivalence margin) in either direction. (2-sided CI)

- **Non-inferiority trials** objective: to show that investigational agent is not worse than an active control by more than a specified amount. (1-sided CI)

- The non-inferiority margin (**δ**) should be much smaller than a clinically meaningful difference (*d*) used in superiority trial.

Non-Inferiority Trial

H_0: Active – Investigational ≥ Non-Inferiority Margin
(Active is superior to Investigational)

H_1: Active – Investigational < Non-Inferiority Margin
(Investigational is not inferior to Active)

Confidence Intervals of Act-Inv Difference

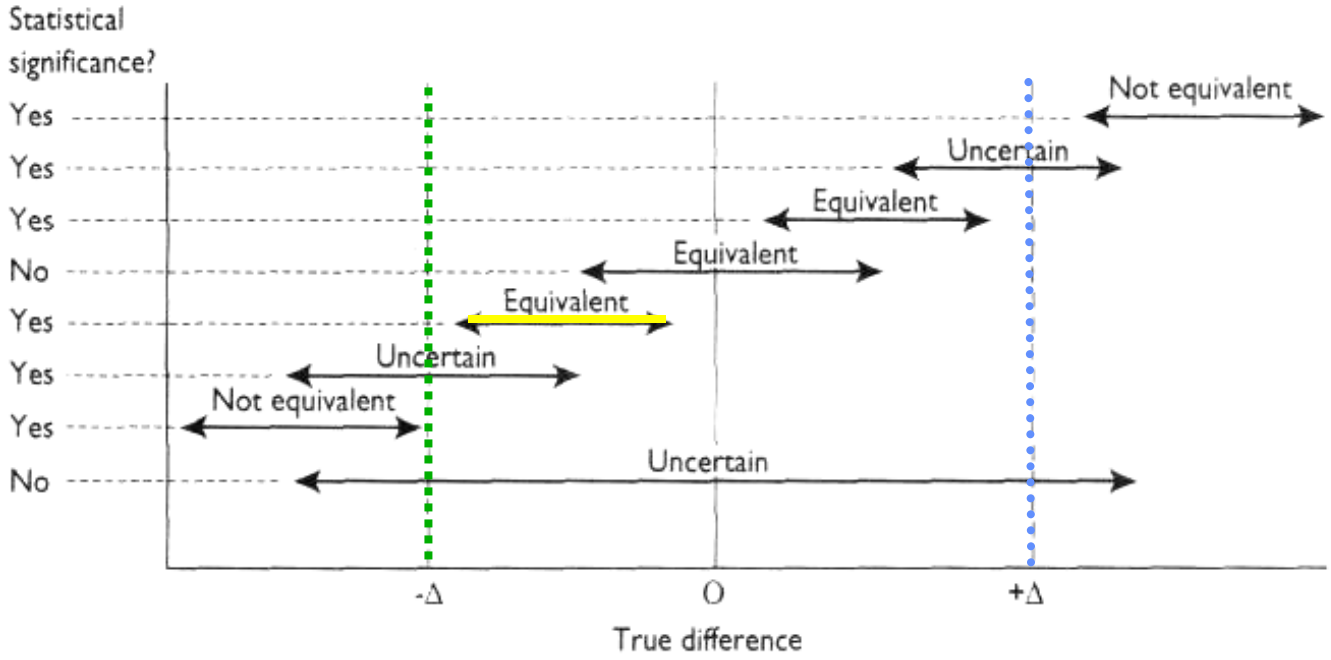

Fig 1—*Examples of possible results of using the confidence interval approach: −Δ to +Δ is the prespecified range of equivalence; the horizontal lines correspond to possible trial outcomes expressed as confidence intervals, with the associated significance test result shown on the left; above each line is the decision concerning equivalence*

B. Jones et al., BMJ, 1996

Errors in Non-Inferiority Tests

Type I error: falsely conclude non-inferiority

Type II error: failure to conclude non-inferiority when treatments are similar

Bias in a non-inferiority trial will favor conclusion of non-inferiority (ICH-E10).

- Inclusion criteria, dosing, small Ns, non-compliance, attrition all can contribute to “non-inferiority”.

Choosing the Non-Inferiority Margin (δ)

Defining "not worse than". Largest difference that is clinically acceptable. Involve both clinical and statistical input.

Margin (δ) should be smaller than smallest effect of active vs. placebo.

Meta-analysis SGAs (Leucht, 2008): $d = 0.51$

If PANSS change sd=20, $d = .51 \cong 10$ PANSS units
Non-inferiority $\delta = .25 = 5$ PANSS units.

Is 5 PANSS points clinically acceptable difference?

Non-inferiority RCT

Required Sample Size > Superiority N

If non-inferiority margin is half of the difference expected in superiority RCT, non-inferiority N is fourfold higher (binary & continuous outcomes)

Larger N provides more safety data

Active comparator could attract potential participants.

Non-Inferiority Sample Sizes: Continuous Outcome

N/group	Delta*
6280	0.05
1570	0.10
698	0.15
393	0.20
252	0.25

$252 \cong 4(64) = 256$ for $d = .50$

* sd units

NCSS/PASS

Assumes: α= .025; β= .20

Unintended Consequence of Active Comparator

Meaningful differences between investigational and active comparator are much smaller than with placebo

Sample size required in RCT of active comparator is larger than a placebo-controlled RCT

By virtue of the larger sample sizes, active comparator RCTs have more non-responders and more Ss exposed to risks of experiment

Applies to both superiority and non-inferiority designs.

Leon AC, Arch Gen Psych 2000; Leon AC et al, Eval and Health Prof 2003

Planning Superiority RCT to compare: *Investigational Drug* vs. *Placebo*

Expected *Response* rates:

Investigational Drug (40%) vs. Placebo (10%)

Sample size requirement: 38 per group*

*For power of 0.80 for a two-tailed chi square test with the continuity correction and alpha of 0.05.

Investigational Drug vs. Placebo

Expected *non*-response rates:

Expected number of *non*-responders:

Investigational:	60% of 38:	23
Placebo:	90% of 38:	34

Total number of *non*-responders: 57

Planning Superiority RCT comparing *Investigational Drug vs. Active Comparator*

Expected *Response* rates:

Investigational (40%) vs. Active (50%)

Sample size requirement: 408 per group

A *larger N* is needed to detect *smaller differences* in response rates.

Investigational Drug vs. Active Comparator

Expected *non*-response rates:

Expected number of *non*-responders:

Investigational:	60% of 408:	245
Comparator:	50% of 408:	204

Total number of *non*-responders: 449

More Non-responders with Active Comparator

The number of non-responders:

Investigational (40%) vs. Placebo (10%) :	57
Investigational (40%) vs. Active (50%) :	449

Ideal Circumstances of Non-inferiority RCT

Use active control superiority over placebo is well documented and convincing

Non-inferiority trial conducted under similar conditions of superiority trials (e.g., subject selection, dosing, trial duration, outcome)

Choice of δ is small enough that clinicians convinced new treatment (passing such non-inferiority test) is therapeutic

Pocock, 2003

Summary

Investigational vs. Placebo vs. Active Comparator

Superiority Contrast: Investigational vs. Placebo

- Efficacy: Design to detect clinically meaningful difference (d)

Superiority Contrast: Active Comparator vs. Placebo

- Assay sensitivity

Non-inferiority Contrast: Investigational vs. Active

- Is Investigational worse than Active?
- Design specifies clinically acceptable difference (δ)

Summary

A negative superiority trial does not demonstrate equivalence or non-inferiority.

Evidence of non-inferiority comes from well-designed and well-conducted RCT. Bias in a non-inferiority trial will favor conclusion of non-inferiority.

Sample size for non-inferiority much larger than superiority

As a result, non-inferiority RCT is more costly, longer duration, and more Ss exposed to risk.