Noninferiority Trials

Scott Evans
Harvard University

Graybill Conference
June 11, 2008

Outline

• Background
• Noninferiority Trials
  – Examples
  – Assumptions
    • Constancy
    • Assay sensitivity
  – Design Issues
    • Choice of active control
    • Choice of noninferiority margin
    • Biocreep
    • Sample Size
Outline

• Noninferiority Trials
  – Design Issues
    • Fixed Margin
    • Combination Approach
    • Precision-Based Approach
    • Synthesis Method
  – Analyses and Reporting
    • Interim Analyses
    • CONSORT
    • ITT vs. PP
    • Missing Data
    • Switching between Noninferiority and Superiority

Abbreviations

• NI = Noninferiority
• CI = Confidence Interval
• M = Noninferiority Margin
• T = Experimental Intervention
• C = Active Control
• P = Placebo
Background on Scientific Designs to Evaluate an Intervention

• Single arm study
  – Administer an intervention to a group of patients and see if they improve or are cured

Single Arm Study
Single Arm Study

• Limitations:
  – Cannot control for “natural history”
    • People may have improved anyway
  – Placebo-effect (or Hawthorne effect)
    • Patients/clinicians believe that they are getting better because of treatment; manifesting itself in better outcomes
      – E.g., pain
  – Miss a “good result” when you observe no change but patients would have gotten worse if left untreated.

Control Group

• Control for natural history

• Control for placebo effect
Placebo Controlled Trial

- Randomize to one of two arms
  - New intervention or placebo
- Blind the intervention to patients and investigators when possible
- Analysis compares the two arms with respect to response
Superiority:

\[ H_0: p_1 - p_2 = 0 \]
\[ H_A: p_1 - p_2 \neq 0 \]

- \( p_1 \) = efficacy of new treatment
- \( p_2 \) = efficacy of control group

Placebo Controlled Trial

- Limitations:
  - If an *effective* standard of care (SOC) treatment exists, then randomization to a placebo may not be ethical
  - Consider using the SOC as an “active control”
    - Show noninferiority (NI) to the active control and thus significant efficacy over placebo (indirectly)
Absence of Evidence is Not Evidence of Absence

- Hypothesis testing is analogous to a court trial:
  - People are assumed innocent until proven guilty
    - $H_0$: Innocent
    - $H_A$: Guilty
  - If VERDICT = Not Guilty (i.e., do not reject $H_0$), then:
    - We cannot say that we have proven innocence
    - We say that we failed to find enough evidence to prove guilt
    - There is a subtle but important difference between the two

Thus...

- We cannot conclude equivalence simply because we fail to reject $H_0$

- Historical problems with reporting
  - Greene et.al. AIM, 2000 reviewed 88 studies claiming NI
    - 67% of them inappropriately concluded NI based on nonsignificant superiority tests
    - Only 23% pre-specified a NI margin
Noninferiority Trial

- Randomize to one of two arms
  - New intervention or active-control
- Blind the intervention to patients and investigators when possible
- Analysis compares the two arms with respect to response
- Need to show new intervention is “no worse” than the active control

Active Control

Population of Patients

Randomize to Treatment

<table>
<thead>
<tr>
<th></th>
<th>Investigational Treatment</th>
<th>Active Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAMPLE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIME</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response</td>
<td></td>
<td>YES</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NO</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YES</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NO</td>
</tr>
</tbody>
</table>
Noninferiority Trial

- Decide on a NI margin, M (more later)

- Analysis
  - Get a confidence interval (CI) for the difference between arms (new intervention minus active control with respect to efficacy) and note if lower bound of CI is within the NI region

---

Efficacy Difference

- **Superiority**:
  - $H_0$: $p_1 - p_2 = 0$
  - $H_A$: $p_1 - p_2 \neq 0$

- **Noninferiority**:
  - $H_0$: $p_1 - p_2 < -M$
  - $H_A$: $p_1 - p_2 \geq -M$

P_1 = efficacy of new treatment
P_2 = efficacy of control group

Noninferiority Margin

```
STATISTICALLY SUPERIOR
       Efficacy Difference
STATISTICALLY INFERIOR
```

- CLINICALLY NONINFERIOR
- CLINICALLY INFERIOR
Statistical vs. Clinical Significance

• Not the same thing!

Example: ACTG 116A

• DDI (500 mg/day and 750 mg/day) vs. AZT for treatment of HIV infection

• Objective
  – To show DDI is noninferior to AZT

• Rational
  – In 1989, AZT was the only approved ARV and had been shown better than placebo in reducing disease progression
  – More treatments needed in case of resistance development
Example: ACTG 116A

- DDI vs. placebo is not ethical with availability of AZT

- Endpoint
  - Time to AIDS defining event or death

- Noninferiority
  - If upper bound of CI for HR (DDI vs. AZT) < 1.6
    - DDI (500mg): HR=1.02 90% CI = (0.79, 1.33)
    - DDI (750mg): HR=1.04 90% CI = (0.80, 1.34)

---

Hazard Ratio

<table>
<thead>
<tr>
<th>DDI</th>
<th>1.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 mg</td>
<td></td>
</tr>
<tr>
<td>750 mg</td>
<td></td>
</tr>
</tbody>
</table>

Exact Equality

\( \Delta = \text{range of practical equivalence} \)
Example: FDA SGE Experience

• A randomized, double-blind, multicenter study comparing the efficacy and safety of Piperacillin/Tazobactam (PT, 4G/500MG) and Imipenem/Cilastatin (IC, 500MG/500MG) administered intravenously every six hours to treat nosocomial pneumonia in hospitalized patients

Example: FDA SGE Experience

• Active Control: Imipenem/Cilastatin (IC) – 60/99 cured

• New drug: Piperacillin/Tazobactam (PT) – 67/98 cured

• Noninferiority margin = 20%
Example: FDA SGE Experience

• Lower bound of 95% CI for the difference in response rates (PT-IC) is –0.066 (> –0.20)
  – Was a margin of 20% too large?
  – Noninferiority would be shown for a margin as small as 7%

• Result
  – PT was noninferior to IC
  – Approved by the FDA

Noninferiority Trial

Compresses a new intervention (T) against an active control (C) that has been shown to be effective in the treatment of the target condition (by superiority to placebo (P)).

The comparison is based on the estimate of a parameter, $\beta_{TC}$, that measures the effect of T relative to C. $\beta_{TC}$ might be the hazard ratio, the difference in means or proportions, or relative risk.

Since C has been shown to be superior to P, we need not show that T is superior to C. Instead, we must show that it is noninferior.
Noninferiority Trial

It is impossible to show that two treatments have identical efficacy. Instead we choose a NI margin, M, and seek to prove that $\beta_{TC}$ is less than M. This is done by testing the null hypothesis:

$$H_0: \beta_{TC} \geq M$$

against the one-sided alternative

$$H_0: \beta_{TC} < M$$

Note: The traditional roles of the hypotheses are reversed.

Noninferiority Trial

- Rule out important relevant (clinical and statistical) differences (i.e., $>M$) with reasonable confidence
T is Often Better than C in Other Ways

- Better safety profile
- Less expensive
- More convenient to administer
  - Less invasive
  - Fewer pills
- Easier to comply
- Shorter treatment duration

Examples

- In HIV, we seek less complicated or toxic regimens with similar efficacy to existing regimens
- Registration of generics
- To show BID is noninferior to TID
- To show that capsule is noninferior to tablet
- To identify new treatment options in case resistance develops
Assumption: Constancy

• Historical data:
  – C showed superiority to P in historical trial ($\beta_{\text{CP,H}} > 0$)

• Constancy Assumption
  – $\beta_{\text{CP,H}} = \beta_{\text{CP,NI}}$ if placebo was present in the NI trial
    • May not be the case in the presence of resistance development
      or with differing trial conduct (e.g., administration of
      treatment, differences in populations or endpoints, etc.)
  – Not verifiable in current trial (without placebo)
  – Implication: to conduct the NI trial in the same manner
    as the trial that established $\beta_{\text{CP}} > 0$

Assumption: Constancy

• NI trials for which the constancy assumption does not hold are “failed” trials
  – This is distinct from a “negative” trial in which
    noninferiority is not demonstrated because of a
    lack of efficacy of the new intervention
Assumption: Assay Sensitivity

• Trial is able to detect differences between treatments if they exist

• Need a sensitive enough instrument to measure response and detect differences if they exist
  – Otherwise, all interventions will display similar responses due to the insensitivity of the instrument

Design Issues

• Participants, endpoints, and other important aspects of the trial should be similar to those used in the trials used to demonstrate the effectiveness of the active control over placebo
  – Early HIV studies used deaths as the primary endpoint. When deaths became uncommon, then AIDS clinical events were used. Now surrogate markers (HIVRNA viral load or CD4 counts) are utilized.
Choice of Active Control

- Must have clinical efficacy
  - Of substantial magnitude
  - That is precisely estimated
  - With estimates that are relevant to the setting in which the NI trial is being conducted
  - Constancy assumption
  - Preferably measured by multiple controlled trials

- Regulatory approval does not necessarily imply that an intervention can be used as an active control in an NI study
  - Superiority to placebo must be reliably established

Choice of the Active Control

- Must have proven superiority over placebo
  - Estimated effect size over placebo (and perhaps associated precision) is used to determine the NI margin
  - Otherwise one may be designing a trial to show NI to something that is no better than (or perhaps even worse than placebo)
Choice of the Active Control

- Must have assurance that the active control would be superior to placebo if a placebo arm was employed in the trial

Choice of Active Control

- Recently, there have been concerns over the development of NI studies using active controls that do not have proven efficacy over placebo
  - Groups claim that placebos are unethical because these active controls are standard of care practice
    - Patients are unwilling to enroll
    - IRBs question the ethics of placebos
Biocreep

• The tendency for recently demonstrated noninferior treatments to be active controls in new NI trials even though they are slightly inferior to historically proven active controls vs. placebo (D’Agostino, SIM, 2003)

• If A is noninferior to B and B is noninferior to C, then it does NOT necessarily follow that A is noninferior to C

• A slightly inferior treatment becomes the active control for the next generation of NI trials
  – Multiple generations of NI trials using active controls that were themselves shown to be effective using NI trials, could result in an active control that is not superior to placebo (and ineffective interventions could being viewed as efficacious)

• NI trials should generally choose the best available control
Choice of Active Control

• Example: OHARA is developing a treatment trial for oral candidiasis (OC) in Africa within the ACTG system
  – Fluconazole is not readily available in Africa (too expensive)
  – Nystatin is used in many places as the standard of care
  – Gentian Violet (GV, an inexpensive topical agent) showed excellent in vitro activity
  – A NI trial of GV compared to Nystatin was proposed

Nystatin as a Control?

• However, despite the standard use of nystatin, there were no published studies that could be identified for evidence that nystatin was superior to placebo

• Pushing for a 3 arm (GV, nystatin, placebo) superiority trial
  – Skepticism about the ethics of the control arm
Choice of Active Control

• Similar issues have arisen with treatments that were once shown to be effective but may no longer be effective because of the development of resistance
  – Example: MRSA

MRSA
“The Superbug”
MRSA

- Methicillin-resistant Staphylococcus aureus
  - Bacterial infection that has become resistant to antibiotics (such as penicillin, amoxicillin, methicillin)
  - Lives on the skin
  - Community acquired
    - Spread through contact

CNN – October 17, 2007

Experts: Drug-resistant staph deaths may surpass AIDS toll
DMC Discussion

• ID clinicians insist antimicrobials are "highly effective" in treating skin infections, but seem unable to site evidence to back this claim other than anecdotal (and then claim it is "unethical" to do any other trial design other than NI).

• Without the data to support one of the drugs as the "control" based upon reliable and reproducible evidence of the magnitude of the benefit of the control compared to placebo, the results of an NI trial are not meaningful

FDA Guidance (released October, 2007)
Use of Noninferiority Studies to Support Approval

• NI study designs may be appropriate when there is adequate evidence of a defined effect size for the control treatment so that the proposed NI margin can be supported. For an NI study, having an adequately justified NI margin is essential to having an informative study. If NI studies are being considered, a comprehensive synthesis of the evidence that supports the effect size of the active control and the proposed NI margin should be assembled during the period of protocol development and provided to the FDA along with the protocol. We are asking sponsors to provide adequate evidence to support the proposed NI margin for any indication being studied using active-controlled studies designed to show NI. It is likely, however, that for some indications, such as acute bacterial sinusitis (ABS), acute bacterial exacerbation of chronic bronchitis (ABECB) and acute bacterial otitis media (ABOM), available data will not support the use of an NI design. We recommend that sponsors consider other study designs (e.g., superiority designs) to provide evidence of effectiveness in these three indications.
Choosing the Noninferiority Margin (M)

Traditional Approach

- Subjective but structured
  - Combination of statistical reasoning and clinical judgment (CHMP Guidance)

- Concepts
  - “Maximum treatment difference that is clinically irrelevant”
  - “Largest treatment difference that is acceptable in order to gain other advantages of the experimental intervention”

- Design parameter is not present in superiority trials

Choosing the Noninferiority Margin (M)

- Must be smaller than the effect size of C over P
  - Use historical data
  - Account for within trial and across trial variability

- Theoretically should be chosen independent of considerations of study power

- Directly impacts study conclusions

- Context-dependent
Choosing the Noninferiority Margin (M)

Investigators frequently set M equal to half of the estimated effect of the active control (C) relative to placebo (P) from the historical evidence. That is, set

\[ M = \frac{\hat{\beta}_{CP}}{2} \]

This is known as “preserving a fraction of the effect”. Note that the method does not consider the fact that the estimate from historical data is subject to uncertainty.

STAR Trial

- Noninferiority analysis of Raloxifene vs. Tamoxifen
  - Primary endpoint: invasive breast cancer
  - Raloxifene is test agent
  - Tamoxifen is an `active control’
Tamoxifen vs. Placebo: NSABP P1 Trial
Subset of Women ≥ 50 years old

RR = 2.12 (1.52 - 3.03)
Favors Tamoxifen
Favors Placebo

Interpretation: P increases the rate of invasive breast cancer incidence compared to Tam by 112% (CI: 52% to 303%)

Relative Risk for Invasive Br Ca: Placebo / Tamoxifen

NI margin: 50% of Active control effect retained.
56% increased risk on P

RR = 1.56
STAR Trial: Claiming NI

• Thus the upper bound of the 95% CI estimate for the relative risk needs to be less than 1.56

The Two CI (95-95) Method

M is set equal to the lower bound of the 95% CI for the effect of C relative to P in the placebo controlled trials (addresses the issue of the variability of the effect estimate).

In the NI trial, the upper confidence limit for the effect of T relative to C must be less than this value.

This criterion is stringent and depends directly on the strength of the evidence in the historical trials.
**Tamoxifen vs. Placebo: NSABP P1 Trial**

*Subset of Women ≥ 50 years old*

<table>
<thead>
<tr>
<th>RR</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
<th>1.2</th>
<th>1.4</th>
<th>1.6</th>
<th>1.8</th>
<th>2.0</th>
<th>2.2</th>
<th>2.4</th>
<th>2.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favors Placebo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Favors Tamoxifen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Relative Risk for Invasive Br Ca: Placebo / Tamoxifen**

 RR = 2.12 (1.52 - 3.03)

NI margin: Lower bound of the 95% CI estimate for the RR of Tamoxifen vs. placebo

RR = 1.52

**Failed Noninferiority Trials**

**Poor Choice of Noninferiority Margin**

In the TARGET trial, two glycoprotein IIa/IIIb inhibitors, tirofiban and abciximab, were studied to establish the NI of tirofiban for treatment of coronary syndromes.

The NI margin for the hazard ratio was chosen as 1.47, half the effect of abciximab in the EPISTENT trial.

The trial was viewed as poorly designed because an agent with a hazard ratio of 1.47 would not have been considered to be therapeutically equivalent to abciximab.
Failed Noninferiority Trials
Poor Choice of Noninferiority Margin

In the SPORTIF trials, ximelegatran was compared to warfarin for stroke prevention in atrial fibrillation patients. The event rates in the warfarin group (control) were 2.3% (Sportif III) and 1.2% (Sportif V). Based on the historical evidence, the sponsor chose an absolute NI margin of 2%. Because of the low event rates in the control arm, this resulted in a NI margin that allowed the conclusion of NI even if the trial did not rule out a doubling of the event rate.

The common theme in these trials was that a NI margin based on the effect of the comparator drug was not consistent with the community standard for therapeutic equivalence.

Sample Sizes

- Sample sizes increase with decreasing M
- Stratification can help
  - Adjusted CIs are generally narrower than unadjusted CIs
- For binary and continuous data, simply reverse the roles of $\alpha$ and $\beta$ from superiority trials
- Weigh the costs of Type I (incorrectly claiming NI) and Type II errors (incorrectly failing to claim NI)
Sample Size

- Still only get $\alpha=0.025$ on the one side
  - Using 0.05 lowers the level of evidence for drawing conclusions vs. accepted practice in superiority trials

- General wisdom (including regulatory agencies) is to use 2-sided 95% CIs to evaluate both superiority and NI

- Important to power for per protocol analyses too
  - More on this later

Sample Size per Group

- Binary Data

$$n = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 (p_c q_c + p_t q_t)}{(p_c - p_t - \delta)^2}$$

- Continuous Data

$$n = 2\left[\left(Z_{1-\alpha} + Z_{1-\beta}\right)\left(s / \delta\right)\right]^2$$
NI Trials May Not Be Appropriate When:

- Constancy assumption is in doubt
- Assay sensitivity is in doubt
- Historical data of active control effect over placebo are thin
- Variability of active control effect is very large

Alternative Designs: 3-Arm Trials

- Consider 3-arm trials with test intervention, active control, and a placebo control
  - Scientifically very attractive
    - High validity
    - Within-trial validation of NI margin
    - Direct comparison of test intervention vs. placebo
  - Ethical?
- Use when the efficacy of the active control has changed, is small, is in doubt, or is volatile
Fixed Margin Method

- Define NI margin M
  - Using data from historical trials
  - Note: subject to bias and uncertainty

- Use the 2-sided 95% CI to rule out M

- Too conservative?

What is the Problem?

Failure to distinguish between two distinct objectives of NI trials:

O₁: The trial must demonstrate that T is NI to C.

O₂: The trial must demonstrate that T is superior to P, taking account of the uncertainty associated with the historical evidence.

This can be achieved by designing the trial to test two separate hypotheses. (Gau and Ware, SIM, 2007)
Testing for NonInferiority

To address O1, noninferiority, we determine a NI margin, M, which, if met will allow us to conclude that T is NI to C. M should be chosen based on clinical considerations, not the historical data regarding the effect of C relative to P.

Given the margin, M, we test the null hypothesis

\[ H_{10}: \beta_{TC} = M \]

Against the alternative hypothesis

\[ H_{1A}: \beta_{TC} < M \]

Using standard methods for NI trials.

Testing for Superiority to Placebo

A direct comparison of T to P is not possible.

If, however, we believe that the effect of C relative to P is identical in the contemporary and historical contexts (constancy), then we can represent \( \beta_{TP} \) as a function of the parameters from the placebo-controlled trial of C and the NI trial:

\[ \beta_{TP} = \beta_{TC,NI} + \beta_{CP,H} \]
Testing for Superiority to Placebo

Assuming assay constancy, we test the null hypothesis:

\[ H_{20} : \beta_{TP} (= \beta_{TC} + \beta_{CP}) = 0 \]

against the alternative \( H_A : \beta_{TP} < 0 \) with the test statistic

\[ \frac{\hat{\beta}_{TP} - \hat{\beta}_{TC} - \hat{\beta}_{CP}}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_{TP}} + \hat{\sigma}^2_{\hat{\beta}_{CP}}}} \]

If we assume approximate normality of the estimates, this can be done with standard methods.

Constancy

To address concerns about constancy, we consider a variant of this method based on the discounted synthetic estimate. Suppose we assume that the effect of C relative to P in the contemporary setting is only a fraction \((1 - \lambda)\) of the effect in the historical setting. Then

\[ \beta_{TP,\lambda} = \beta_{TC} + (1 - \lambda)\beta_{CP} = 0 \]

We then test \( H_{20} : \beta_{TP,\lambda} = 0 \) with the test statistic

\[ \frac{\tilde{\beta}_{\lambda, \hat{\beta}_{TP} - \hat{\beta}_{TC} - \hat{\beta}_{CP}}}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_{TP}} + \hat{\sigma}^2_{\hat{\beta}_{CP}}}} \]
Design and Analysis

Since the goal of the NI design is to reject both $H_{01}$, the hypothesis of inferiority, and $H_{02}$, the hypothesis of no difference relative to placebo, we perform sample size calculations for each test and choose $N$ equal to the larger of the two sample sizes.

The analyses do not require two unrelated tests. Since the results from the historical trials are known at the time the NI trial is conducted, the two hypothesis tests can be reduced to two tests involving the estimated treatment effect of $T$ relative to $C$ in the NI trial. The two tests can be expressed as inequalities involving

$$\hat{\beta}_{NI}$$

If the estimate satisfies the more stringent of the two inequalities, we reject both $H_{01}$ and $H_{02}$ and achieve both objectives.

Implications

A particular clinical trial might achieve only one of the two objectives.

Suppose that we demonstrate that $T$ is superior to placebo ($O_2$) but fail to show that $T$ is therapeutically equivalent to $C$ ($O_1$). The intervention might be of use in the treatment of patients for whom $C$ is contraindicated or not available.

In contrast, if we achieve $O_1$ but not $O_2$, we can conclude that $T$ is therapeutically equivalent to $C$ but may not be superior to $P$. This is likely to arise when the evidence supporting $C$ is weak.
**Designs Based on Precision**

- Based on principles of estimation rather than hypothesis testing
  - Do not specify explicit hypotheses
    - Avoid making the distinction between superiority and NI

- Specify the precision with which you want to estimate treatment differences (e.g., the maximum length of a CI for the treatment effect difference)
  - Power the study to estimate the effects with this precision
  - Interpret the CI as usual (ruling out effects in either direction as appropriate)
  - May wish to pre-specify a NI margin for regulatory purposes

**Rethinking the Objective?**

- Objective for industry (for approval)?
  - Show superiority to placebo (but via use of a NI trial)
  - Not concerned with showing NI to active control

- Objective is not to show statistical significance, but to obtain statistically reliable evaluation of safety and whether the intervention provides a clinically meaningful benefit
Inconsistency with Preservation of Effect?

• New intervention could look better than active control (point estimates) but not meet the preservation of effect condition

• Two trials with different active controls have different standards for success

• If new intervention is better than active control, should the active control be withdrawn?

Synthesis Method

• Theory
  – Required degree of efficacy should be independent of design
  – Demonstration of “any efficacy” is sufficient
    • Goal: $\beta_{TP} > 0$ (or equivalently $\beta_{CP} + \beta_{TC} > 0$)
    • Estimated by $B_{CP} + B_{TC}$, with variance $V_{CP} + V_{TC}$
  – “noninferiority” term is inappropriate
    • Only superiority to P is needed
Synthesis Method

• Combines standard errors from historical trials and NI trial
  – Note: not the standard error from a randomized comparison

• NI is not considered
  – M is irrelevant

Synthesis Method

• Efficient Z-test
  \[ \frac{(B_{CP} + B_{TC})}{\sqrt{V_{CP} + V_{TC}}} > 1.96 \]

• Discounting factors for non-constancy
  – Subjective and no supporting data
  \[ \frac{v(1-w)(B_{CP} + B_{TC})}{\sqrt{(1-w)^2(V_{CP} + V_{TC})}} > 1.96 \]
Issues with Synthesis Method

• Proper control of conditional type I error is not warranted

• Assumes no bias in historical point estimate
  – Often not the case with publication bias
  – Problematic when constancy assumption does not hold

• Cannot produce a fixed NI margin

• No consensus on role of across-trial error

• Too liberal?

Interim Analyses

• May wish to stop of futility

• No ethical reason to stop for NI (unless clear superiority is already demonstrated)
  – May want to continue to see if superiority can be shown

• Use repeated CIs (with group sequential error spending principles) and predicted intervals
  – Evans et.al., *DIJ*, 2007
Interim Result: What Would You Recommend?

Noninferiority:
H₀: p₁ - p₂ > - Δ
Hₐ: p₁ - p₂ ≥ - Δ

CLINICALLY INFERIOR \[ \rightarrow \] CLINICALLY NONINFERIOR

A [ ]
B [ ]
C [ ]
D [ ]
E [ ]
F [ ]

STATISTICALLY INFERIOR \[ \rightarrow \] STATISTICALLY SUPERIOR

Noninferiority Margin

Interim Result: What Would You Recommend?

• Extension of the CONSORT statement
  – Piaggio et.al., *JAMA*, 2006.
## CONSORT

### Table: Checklist of Items for Reporting Non-inferiority of Equivalence Trials

<table>
<thead>
<tr>
<th>Paper Section and Topic</th>
<th>Item Number</th>
<th>Description (Adapted for Non-inferiority or Equivalence Trials)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title and abstract</td>
<td>1*</td>
<td>Item 1: Exact number of participants (i.e., randomization, randomization or randomization) ensuring that the sample is representative of the population of interest.</td>
</tr>
<tr>
<td>Introduction</td>
<td></td>
<td>2* Scientific background and explanation of rationale, including the rationale for using a non-inferiority or equivalence design.</td>
</tr>
<tr>
<td>Methods: Participants</td>
<td>3*</td>
<td>3* Eligibility criteria for participants (including whether participants are randomized to the intervention and whether the intervention is part of a randomized controlled trial, if applicable).</td>
</tr>
<tr>
<td>Interventions</td>
<td>4*</td>
<td>4* Description of the intervention(s) and the key characteristics of the intervention(s) (including treatment group, and treatment group and intervention).</td>
</tr>
<tr>
<td>Objectives</td>
<td></td>
<td>5* Specific objectives and hypotheses, including the hypothesis concerning non-inferiority or equivalence.</td>
</tr>
<tr>
<td>Sample size</td>
<td>6*</td>
<td>6* Sample size determined (e.g., using a non-inferiority or equivalence criterion) and specifying the margin of equivalence with the sample size for each group.</td>
</tr>
<tr>
<td>Randomization and allocation</td>
<td></td>
<td>7* Randomization method used to generate the randomization sequence, including details of any restriction (i.e., blocking, stratification).</td>
</tr>
<tr>
<td>Blinding</td>
<td>8*</td>
<td>8* Whether or not participants, clinicians, and outcome assessors were randomized to group assignment.</td>
</tr>
<tr>
<td>Statistics and baselines</td>
<td></td>
<td>9* Analysis plan (i.e., statistical methods used to compare groups, specifying whether a 1% or 2-sided confidence interval was used).</td>
</tr>
<tr>
<td>Results: Patient flow</td>
<td>10*</td>
<td>10* Flow of participants through each stage of the study (i.e., randomization, treatment, follow-up, and analysis).</td>
</tr>
<tr>
<td>Baseline data</td>
<td>11*</td>
<td>11* Baseline demographic and clinical characteristics of the group.</td>
</tr>
<tr>
<td>Numbers analyzed</td>
<td>12*</td>
<td>12* Number of participants included in each analysis, and whether intention-to-treat analysis was conducted.</td>
</tr>
<tr>
<td>Outcome and estimation</td>
<td>13*</td>
<td>13* Outcome measures, including the number of participants and the sample size for each group (i.e., number of participants).</td>
</tr>
<tr>
<td>Adverse events</td>
<td>14*</td>
<td>14* Description of adverse events and their relationship to intervention.</td>
</tr>
<tr>
<td>Patient follow-up</td>
<td></td>
<td>15* Description of adverse events and their relationship to intervention.</td>
</tr>
<tr>
<td>Consent and participation</td>
<td></td>
<td>16* Interpretation of the results (i.e., including non-inferiority or equivalence hypothesis and any other hypothesis, sources of potential bias).</td>
</tr>
<tr>
<td>Generalizability</td>
<td>17*</td>
<td>17* Generalizability (i.e., potential validity of the trial findings).</td>
</tr>
<tr>
<td>Overall evidence</td>
<td>18*</td>
<td>18* Overall evidence (i.e., including non-inferiority and equivalence).</td>
</tr>
</tbody>
</table>
Analyses

• 2-sided CIs are generally used
  – Consistent between significance testing and subsequent estimation
  – Consistent with ICH E-9
  – P-values are not appropriate

• Choice of NI margin plays a direct role and thus must be justified

Analyses

• Analysis uses information outside of current trial to infer the effect of new intervention vs. placebo in absence of a direct comparison
  – Compare response rate, adherence, etc. of control relative to historical trials which provided evidence of the efficacy of the control
  – Can use meta-regression to determine the estimate of $\beta_{TP}$ using historical data
    • Rarely done
    • Assumes constancy
Interpretation Issues

• If active control displays different efficacy than in prior trials vs. placebo, then the validity of the pre-defined NI margin may be suspect and the interpretation of treatment differences is challenging
  – Can we model the changes in the effect of C vs. P
• May not have appropriate data as we need a comparison of C vs. P from multiple trials conducted at different times

ITT vs. Per Protocol

• Non-adherence, missing data, inadvertent enrollment, misclassification, measurement error and treatment crossovers can bias the study
  – Can undermine the validity of the trial
  – Unclear of the direction of the induced bias
    • Can bias towards NI
  – ITT not necessarily conservative
  – Careful planning of trial and patient follow-up is critical
  – Minimize dropout and non-adherence in design
  – Diligent monitoring necessary
ITT vs. Per Protocol

• Sensitivity analyses are important
  – ITT and Per Protocol should both be conducted
  – Need consistency of qualitative result (CPMP, 2000, points to consider)

• PP usually results in a larger effect size
  – ITT dilutes effects

• PP usually has wider CIs
  – Fewer patients than ITT

Analysis of Missing Data

• May wish to utilize a creative (biased) imputation method for missing data
  – E.g., binary endpoint: impute success for the active control and failure for new intervention
    • If you still show NI, then it is not because of missing data
    • Very conservative
Idea: Analysis of Missing Data

• Consistent with the null hypothesis of the NI trial
  – E.g., continuous endpoint: impute reasonable expected value (i) for the active control and \((i-M)\) for the new intervention
  – E.g., binary data: impute expected proportion \((p)\) for active control and \((p-M)\) for new intervention
• Then analyze using analysis of means

Switching from Noninferiority to Superiority

• Generally appropriate to calculate p-value for superiority after showing NI
  – Use ITT
  – Define plans apriori when possible
  – No multiplicity adjustment necessary
    • Closed test procedure
Switching from Superiority to Noninferiority

• Generally not acceptable to go from failing to demonstrate superiority to NI
  – Unless a NI margin was pre-specified
  – Post-hoc definition of NI margin is difficult to justify
    • Choice of NI margin needs to be independent of trial data

Switching from Superiority to Noninferiority

• Concerns:
  – Is the control group an appropriate control for a NI study
    • i.e., with demonstrated superiority to placebo
  – Was the efficacy displayed by the control group similar to that shown in trials vs. placebo (constancy)
  – ITT and PP become equally important
  – Trial quality must be high
    • Poor adherence, drop-out, etc. could bias towards NI
Switching from Superiority to Noninferiority

- May be feasible if:
  - Noninferiority margin was pre-specified (or can be justified, which is difficult to do)
    - Must be based on external information and not chosen to fit the data
  - ITT and PP show similar results
  - The trial was of high quality with few drop-outs and good adherence
  - The control group displayed similar efficacy to trials vs. placebo
  - The trial was sensitive enough to detect effects

Switching between Noninferiority and Superiority

- Can always interpret the CI as excluding an effect of a pre-stated size

- Changing the NI margin?
  - Generally okay to decrease
  - An increase should be justified from external data (independent of trial)