Noninferiority Clinical Trials

Scott Evans, PhD, MS, Harvard University

Harvard Catalyst
April 4, 2016

Outline

• Concept, Rationale, and Examples
• Assumptions
  – Constancy
  – Assay Sensitivity
• Design
  – Selecting the Active Control
  – Selecting the NI Margin
  – Other issues and design alternatives
• Trial Conduct
• Analyses
  – ITT vs. Per Protocol
  – Missing Data
  – Switching between Noninferiority and Superiority
• Reporting
Abbreviations

- NI = Noninferiority
- CI = Confidence Interval
- M = Noninferiority Margin
- SOC = Standard of Care

Typical Framework

- Interest in testing new intervention

- An alternative treatment has been shown to be effective (i.e., superior to placebo) in historical trials, potentially making placebo unethical

- Thus evaluate whether the test intervention is “noninferior” in effectiveness to a active control intervention, usually the current SOC
Noninferiority Trials

- NI is assessed by evaluating if inferiority of a pre-specified magnitude (the NI margin) can be ruled-out with reasonable confidence

- The NI margin should be carefully selected to ensure that:
  1. A conclusion of noninferiority implies that the test intervention is effective compared to placebo/no therapy, and
  2. “Clinically important” levels of inferiority relative to the control intervention can be ruled out, so that clinical application of the new intervention would be ethical and clinically acceptable

- Underlying implication (hopefully)
  - The experimental therapy is thus superior to placebo
  - Therapeutic exchangeability

Warning!

- NI cannot be demonstrated with non-significant p-values from superiority tests
  - High p-value ≠ similarity
  - Remember the scientific method

- Absence of evidence is not evidence of absence
Methodological Approach

- Select NI Margin (M)

- Rule out important clinically relevant differences with reasonable confidence by showing that differences between experimental therapy and active control are ≤ M
  - Analyses consists of obtaining a (2-sided) CI for the between-group difference noting if the CI estimate is ≤ M
  - Hypotheses
    - H₀: Between group differences > M
    - Hₐ: Between group differences ≤ M (note: 1-sided alternative)

Ethical Dilemmas

- Null hypothesis is inferiority (assumed to be true)
  - Ethicists argue that this is not necessary equipoise
  - Patients often not consented in a manner that informs them that they may receive an intervention that is hypothesized to be inferior

- Why will patients volunteer to risk being randomized to a strategy that might be as good (but unproven as of yet) as a proven existing medical alternative but is not hypothesized to be better?
  - Why not simply opt for the proven alternative?
**Interpretation**

- **CLINICALLY INFERIOR**
- **CLINICALLY NONINFERIOR**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
</table>

Noninferiority:
- $H_0: p_1 - p_2 \leq -M$
- $H_A: p_1 - p_2 \geq -M$

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

**Motivation**

- Ideally the experimental therapy may be better in other ways
  - Better toxicity profile
  - Less expensive
  - Less invasive
  - Simpler regimen
  - Shorter treatment duration
  - Different resistance profile

- If new therapy can be shown to be NI with respect to efficacy and has other advantages, then it will be useful
Examples

• In HIV, less costly or less toxic regimens with similar efficacy to existing regimens are sought

• Evaluation of generics

• To show BID is NI to TID

• To identify new treatment options in case resistance develops to current alternatives

Example: ACTG 116A

• Objective
  – To show that DDI is NI to AZT

• Rational
  – In 1989, AZT was the only approved ARV and had been shown better than placebo in reducing disease progression
  – Placebo not ethical with approval of AZT
  – More treatments needed considering development of resistance

• Endpoint: time to AIDS-defining event or death

• NI if an increase in the risk is not more than 60%
  – I.e., UB of CI for HR (DDI vs. AZT) < 1.6
Example ACTG 116A

DDI (500mg/day): HR=1.02 90% CI = (0.79, 1.33)
DDI (750mg/day): HR=1.04 90% CI = (0.80, 1.34)

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
- Active Control: Imipenem/Cilastatin (IC)
  - 60/99 cured
- New drug: Piperacillin/Tazobactam (PT)
  - 67/98 cured
Example: FDA SGE Experience

- NI margin = 20%
- 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?
  - NI would be shown for a margin as small as 7%
- Result
  - PT was noninferior to IC
  - Approved by the FDA

Example: LABA Asthma Trials

FDA Drug Safety Communication: FDA requires post-market safety trials for Long-Acting Beta-Agonists (LABAs)

This safety review update is in follow-up to the FDA Drug Safety Communication on LABAs on 06/22/2010.

[04-18-2011] To further evaluate the safety of Long-Acting Beta-Agonists (LABAs) when used in combination with inhaled corticosteroids for the treatment of asthma, the U.S. Food and Drug Administration (FDA) is requiring the manufacturers of LABAs to conduct five randomized, double-blind, controlled clinical trials comparing the addition of LABAs to inhaled corticosteroids versus inhaled corticosteroids alone.

Four clinical trials will be conducted in adult and adolescent patients 12 years of age and older. The adult and adolescent trials will include 11,700 patients in each trial for a total of 46,800 patients. Each trial will evaluate one of the following LABA-containing drugs: 1) Symbicort (budesonide and formoterol); 2) Advair Diskus (fluticasone and salmeterol); 3) Duksis (mometasone and formoterol); and 4) Foradil (formoterol). The Foradil trial will also include treatment with fluticasone, which will be provided in a separate inhaler.

One clinical trial will be conducted in pediatric patients aged 4 to 11 years with Advair Diskus. The pediatric trial will include 6,530 patients. Patients in all trials will be treated for six months, and the primary endpoint will be a composite of serious asthma outcomes: asthma-related death, intubation, or hospitalization. The pediatric trial will also assess other relevant quality of life endpoints such as days of school missed and emergency room visits because of asthma-related illness.
Example: LABA Asthma Trials

- Concerns of increase in asthma-related deaths or serious exacerbations for asthma patients taking medications containing LABAs

- FDA urged LABA-drug makers to design randomized NI trial comparing LABA+ICS (inhaled corticosteroids) vs. ICS alone, to rule out unacceptable risk (4 sponsors each running own trial)

- Composite event-time endpoint: asthma-related hospitalization, intubation, and death)

- Need upper bound of 1-sided 97.5 CI for the hazard ratio < 2.0

- N=11,700 for each sponsors trial
NI Complexities

- Lower scientific integrity than superiority trials
  - Prone to biases and manipulation
  - Validity relies upon several foundational requirements

- Assay sensitivity

- Constancy assumption

- Blinding provides less protection from dilution of differences, as blinded investigators can skew results toward similarity by assigning similar response ratings for all participants given knowledge that all participants receive active interventions

Assay Sensitivity

- Treatment differences can be diluted (intentionally or unintentionally) by reducing “assay sensitivity” through subtle choices in design and conduct … resulting in a NI conclusion
  - E.g., consider a case where nobody adheres to treatment … resulting in treatments appearing to be similar in effects

- Potential causes of dilution: poor adherence and treatment crossovers; inadvertent enrollment; poor diagnostic criteria; concomitant medications; LFU and missing data; poorly defined endpoints, misclassification, and measurement error

- Need sensitive instruments/processes to detect differences if they exist
  - Otherwise, interventions appear similar due to the insensitivity

- High quality trial conduct is critical (e.g., diligent patient follow-up)
### Measures Taken to Ensure Assay Sensitivity often Limit Pragmatism and Feasibility

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Explanatory</th>
<th>Pragmatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>Can it work?</td>
<td>Will it work?</td>
</tr>
<tr>
<td>Goal</td>
<td>Evaluate mechanisms.</td>
<td>Improve practice / policy.</td>
</tr>
<tr>
<td>Relevance to Practice</td>
<td>Indirect.</td>
<td>Direct.</td>
</tr>
<tr>
<td>Participants</td>
<td>Selected. Defined criteria.</td>
<td>Limited restrictions.</td>
</tr>
<tr>
<td>Intervention Application</td>
<td>Strict instructions.</td>
<td>Flexible.</td>
</tr>
<tr>
<td>Participant Adherence</td>
<td>Enforced and monitored.</td>
<td>Not enforced.</td>
</tr>
<tr>
<td>Analyses</td>
<td>mITT and PP.</td>
<td>ITT. As seen in practice.</td>
</tr>
</tbody>
</table>

E.g., limited to patients without prior therapy
Generalizability is limited given common use of prior therapy.

### Limiting Pragmatism and Feasibility

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Explanatory</th>
<th>Pragmatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>Can it work?</td>
<td>Will it work?</td>
</tr>
<tr>
<td>Goal</td>
<td>Evaluate mechanisms.</td>
<td>Improve practice / policy.</td>
</tr>
<tr>
<td>Relevance to Practice</td>
<td>Indirect.</td>
<td>Direct.</td>
</tr>
<tr>
<td>Participants</td>
<td>Selected. Defined criteria.</td>
<td>Limited restrictions.</td>
</tr>
<tr>
<td>Intervention Application</td>
<td>Strict instructions.</td>
<td>Flexible.</td>
</tr>
<tr>
<td>Participant Adherence</td>
<td>Enforced and monitored.</td>
<td>Not enforced.</td>
</tr>
<tr>
<td>Analyses</td>
<td>mITT and PP.</td>
<td>ITT. As seen in practice.</td>
</tr>
</tbody>
</table>

Not directly helpful for decision-making since a patient’s subgroup-status (i.e., mITT) is unknown until after treatment has already been initiated.
Trial Conduct

- High quality trial conduct is critical
  - Minimize drop-out and poor adherence

- Careful planning with diligent patient follow-up and monitoring is important

Constancy

- Premise
  - Active control is effective (i.e. superior to placebo/sham supported by historical trial data)

- Constancy Assumption
  - The effect of active control relative to placebo is unchanged
    - Otherwise may be unable to show retention of some of the active control effect vs. placebo
    - May not be the case in the presence of changing medical practice, development of resistance, etc.
  - Not verifiable internally (without placebo/sham) but can compare control rates to that from historical studies
Control Selection

- The control should have established superiority over placebo / sham (with respect to the same endpoint / setting used in the new trial)
  - Regulatory approval may not be sufficient
  - Issue with belief of effectiveness with off-label use

- “Biocreep”
  - Can be problematic when a therapy shown to be noninferior is selected as the active control for the next generation of NI trials
  - NI is not transitive
    - If A is noninferior to B and B is noninferior to C, then it does NOT necessarily follow that A is noninferior to C
    - Could be a problem with iterative generations of predicate devices

- Consider selecting the best available control

Is a NI trial appropriate?
MRSA: The Superbug
CNN – October 17, 2007

FDA Guidance on use of NI trials
(October 2007)

- 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.
Acute Bacterial Sinusitis (ABS)

- One of most common indications for prescribing antimicrobials
- FDA approved > 20 new drugs in ABS based on NI of new drug to old drug when not clear old drug was superior to placebo
  - 12 of 17 randomized placebo-controlled trials show no benefit of the antimicrobial used as a control in the NI trial
  - 9 of 17 RCTs show statistically significant increased harms with the control antimicrobial compared to placebo
  - No valid standardized outcome measures (17 placebo-controlled trials used 15 different outcome definitions)

### Analysis of Efficacy in Placebo Controlled Trials in Acute Bacterial Sinusitis

<table>
<thead>
<tr>
<th>Study</th>
<th>Treatment</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kristo et al. 2005 n=82</td>
<td>cefuroxime d14</td>
<td></td>
</tr>
<tr>
<td>Norrelund et al. 1978 n=135</td>
<td>pivampicillin d8</td>
<td></td>
</tr>
<tr>
<td>Stalman et al. 1997 n=186</td>
<td>doxycycline d10</td>
<td></td>
</tr>
<tr>
<td>Garbutt et al. 2001 n=161</td>
<td>amox or amoxicillin-clav d14</td>
<td></td>
</tr>
<tr>
<td>Lindbaek et al. 1998 n=70</td>
<td>amoxicillin d14</td>
<td></td>
</tr>
<tr>
<td>Bucher et al. 2003 n=251</td>
<td>amoxicillin-clavulanate d14</td>
<td></td>
</tr>
<tr>
<td>Merenstein et al. 2005 n=135</td>
<td>amoxicillin d14</td>
<td></td>
</tr>
<tr>
<td>van Buchem et al. 1997 n=206</td>
<td>amoxicillin d10</td>
<td></td>
</tr>
<tr>
<td>deSutter et al. 2003 n=135</td>
<td>amoxicillin d14</td>
<td></td>
</tr>
<tr>
<td>Axelsson et al. 1970 n=142</td>
<td>amoxicillin d10</td>
<td></td>
</tr>
<tr>
<td>Varonen et al. 2003 n=146</td>
<td>amoxicillin or penicillin d10</td>
<td></td>
</tr>
<tr>
<td>Wald et al. 1986 n=93</td>
<td>amoxicillin or penicillin d10</td>
<td></td>
</tr>
<tr>
<td>Kaiser et al. 2001 n=265 (77)</td>
<td>amoxicillin or doxy or penicillin d 14</td>
<td></td>
</tr>
<tr>
<td>Hansen et al. 2000 n=127</td>
<td>azithromycin d8</td>
<td></td>
</tr>
<tr>
<td>Haye et al. 1998 n=168</td>
<td>penicillin d7</td>
<td></td>
</tr>
<tr>
<td>Lindbaek et al. 1996 n=127</td>
<td>azithromycin d14</td>
<td></td>
</tr>
<tr>
<td>Ganaca et al. 1973 n=50</td>
<td>amoxicillin or penicillin d14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cyclacillin (not specified)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Favors study drug</th>
<th>Favors placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>10</td>
</tr>
</tbody>
</table>

Favors placebo
A5265: Treatment for Oral Candidiasis in Africa

- Fluconazole unavailable (expensive)
- Nystatin is used as the SOC
- Gentian Violet (GV), an inexpensive topical agent, showed excellent in-vitro activity
- A NI trial of GV compared to Nystatin was proposed
- But published studies showing the superiority of nystatin to placebo could not be identified (no nystatin effect to retain)
- Running superiority study (nystatin acting as placebo)
  - But may be unable to claim superiority to placebo upon conclusion
NI Margin

- Should be carefully selected to ensure that a NI conclusion implies:
  1. the test intervention is effective compared to placebo/no therapy, and
  2. “clinically important” levels of inferiority to the control intervention can be ruled out, implying therapeutic exchangeability

- Unfortunately, in practice, the selection of the NI margin can ignore both criteria and is based on sample size considerations, or is based solely on #1 (demonstration of effectiveness compared to placebo) with little attention paid to criteria #2 (clinical importance). Limited work has been conducted regarding what levels of inferiority are inconsequential to patients. Increased emphasis on the clinical importance considerations is needed in future trials.

NI Margin: Antibiotics

- In many cases, reliable data to justify a NI margin often does not exist or is no longer applicable due to medical practice advances or evolution of antibiotic resistance
  - Selections based on studies from the 1930s - 1950s
  - The validity of some trials were questioned because there was no reliable evidence to justify a margin (AIDAC 2012)
Selection of the NI Margin

- Combination of statistical reasoning and clinical judgment
  - Must be smaller than the effect size of active control over placebo to retain effect
  - Context: consideration of disease severity as well as availability and costs for alternative therapies
    - “Maximum difference that is clinically irrelevant”
    - “Largest treatment difference that is acceptable in order to gain other advantages of the experimental intervention”

- FDA
  - M1: effect of active control (recommend bound of 95% CI vs. placebo, acknowledging conservative)
  - M2: largest clinically acceptable difference

- Pre-specification is important

- Directly impacts study conclusions

Choosing the Noninferiority Margin

- No statistical formula

- One strategy: Fixed margin approach (preserve a fraction of the effect)
  - E.g., set the margin to be half of the estimated effect that the active control had over placebo
    - Note that this approach does not consider the fact that the estimate from historical data is measured with uncertainty

- STAR Trial
  - NI evaluation of Raloxifene vs. Tamoxifen
    - Primary endpoint: invasive breast cancer
    - Raloxifene is test agent
    - Tamoxifen is the active control
Tamoxifen vs. Placebo: NSABP P1 Trial Subset of Women ≥ 50 years old

Favors Placebo Favors Tamoxifen

RR = 2.12 (1.52 - 3.03)

Interpretation: P increases the rate of invasive breast cancer incidence compared to Tam by 112% (CI: 52% to 203%)

Relative Risk for Invasive Breast Cancer: Placebo / Tamoxifen

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

- Another strategy: two 95%-95% CI method
  - Set the NI margin = lower bound of the 95% CI for the effect of the placebo relative to the active control in the placebo controlled trial
    - Addresses the issue of the variability of the effect estimate
    - This criterion is stringent and depends directly on the strength of the evidence in the historical trial

Tamoxifen vs. Placebo: NSABP P1 Trial Subset of Women ≥ 50 years old

RR = 2.12 (1.52 - 3.03)

Relative Risk for Invasive Breast Cancer: Placebo / Tamoxifen
STAR Trial: Claiming NI

- Using “preservation of effect” (50%) method
  - The upper bound of the 95% CI estimate for the relative risk needs to be less than 1.56

- Using 95-95 method
  - The upper bound of the 95% CI estimate for the relative risk needs to be less than 1.52

Examples of Poor Selection of NI Margin

- TARGET Trial
  - Evaluated if tirofiban was NI to abciximab for coronary syndromes
    - NI margin was a HR = 1.47 (half of the effect of abciximab in the EPISTENT trial)
    - Problem: agent with a HR of 1.47 would not have been considered therapeutically NI to abciximab

- SPORTIF Trials
  - Ximelegatran compared to warfarin for stroke prevention in atrial fibrillation patients
    - Warfarin event rates in prior trials were 2.3% and 1.2% (SPORTIF III & V)
    - NI margin was selected an absolute 2% difference
    - However this may not rule out a doubling of the event rate
Major Journals Highlighting Research of Less Care

• JAMA Internal Medicine
  – Collection: Less Is More
  – Manuscripts are designated as “Less Is More®” by the editors if the subject highlights the value of improved patient-centered outcomes associated with lesser intensity or quantity of interventions.

• BMJ
  – Too Much Medicine Campaign
  – http://www.bmj.com/too-much-medicine

Lower Respiratory Tract Infection (LRTI)

• Antibiotics are frequently prescribed without proper rationale
  – Leads to avoidable AEs
  – Drives antibacterial resistance

• Majority of acute respiratory tract infections presenting to outpatient settings are suspected to be of viral etiology
  – But often treated with antibiotics (which treat bacterial infections, not viral infections)
Developing RCT

- Procalcitonin (PCT): A biomarker for non-bacterial infections
- Enrichment trial to test the hypothesis that antibiotics can safely be withheld in this biomarker-defined population
- Double-blind RCT evaluating “NI” of placebo vs. azithromycin in adults presenting as outpatients with suspect LRTI and a PCT level of <0.1 ng/mL, as a strategy for reducing antibiotic prescriptions
- NI is meant to ensure “retention of some of the active control effect”
  - How to retain any of the effect of antibiotics over placebo by using placebo?

Other Design Considerations

- 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
  - Can evaluate constancy assumption with active control
  - If endpoints have changed then how do we justify the NI margin and ensure that there is an effect to be retained?
  - This limits innovation

- Take measures to minimize variation (e.g., central labs, objective endpoints, standardized definitions, methods of application and collection) as wide CIs make it difficult to reach conclusions
Good News or Too Good to be True?

• Reviews of PubMed publications of NI trials
  – 55/57 (96.5%) of industry-funded NI trials published in 2011 had positive results (Flacco et. al., 2015)
  – 325/337 (96.4%) of NI trials had positive results (systematic review of 1992–2011 publications) (Li et. al., 2013)

• May be good news but raises concerns about the ability of many NI trials to objectively evaluate therapies, or whether such results could be due to poor assay sensitivity, lack of constancy, ineffective control groups, and overly generous NI margins

Composite Endpoints: A Noninferiority Trial

• Affymax received approval for peginesatide injection for the treatment of anemia due to chronic kidney disease (CKD) in adult patients on dialysis

• FDA required post-approval randomized controlled NI study to rule out increase risk of CV outcomes in incident dialysis patients (compared to Epoetin alpha)
  – Composite event time endpoint consisting of: death, MI, stroke
  – Desire to show RR<1.3

• Questions
  1. Should they use all-cause death or CV death only?
  2. Should they add additional components (unstable angina requiring hospitalization and peripheral vascular disease) to the composite?
Composite Endpoints: A Noninferiority Trial

- Argument for all-cause death and adding other components
  - More events implies smaller sample size in event-time trials
  - Non-CV death is competing risk (informative censoring) for CV death

- Argument against
  - Less serious events could dominate outcome
  - Including events unrelated to treatment would artificially trend towards NI

Alternative Design: 3-Arm Trial

- Consider a 3-arm trial 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
    - Can evaluate constancy assumption
  - Ethical?
  - Higher sample size

- Consider using when the efficacy of the active control has changed, is small, is in doubt, or is volatile
Alternative Design: Estimation-Based

• 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

• Particularly useful when the acceptable NI margin is not universal

Alternative Design: Estimation-Based
(Failed) Example: A5263

• RCT evaluating NI of: (1) ARV+oral etoposide, and (2) ARV+BV to (3) ARV+liposomal anthracycline for advanced Kaposi’s sarcoma (KS)

• Primary endpoint: death or KS progression at 1 year

• Multinational in resource-limited settings (with different availabilities of treatment alternatives)
  – Each country has there own view of an appropriate NI margin

• Suggested sizing trial based on estimation
  – Allow country-specific interpretation of NI

• Team liked it but faced other resistance
  – Less familiarity with estimation based design
  – Uncomfortable with not pre-specifying how to interpret results
Breaking Down the Research Question

- Two distinct objectives of NI trials
  1. Evaluate if new intervention is NI to active control
  2. Evaluate if new intervention is superior to placebo (accounting for uncertainty of historical evidence)

- Can design trial to address both of these objectives simultaneously or either objective individually (Gau & Ware, *SIM*, 2007)

Two Distinct Objectives

- Evaluating NI to Active Control
  - May be of interest for assessing “comparative effectiveness” (i.e., common in government funded trials) or for marketing purposes

- Evaluating superiority to placebo
  - Considered the usual criteria for regulatory approval
  - Perhaps these trials should not be called “noninferiority trials”
  - Can be accomplished using the “synthesis method”
When One Objective is Achieved

- Superiority to placebo but not NI to active control
  - The new therapy may be useful for patients in which the active control is contraindicated or not available

- NI to active control but not superior to placebo
  - May indicate lack of constancy or that evidence for active control effect is weak

Sample Size

- Important to power for both ITT and per protocol analyses

- Consider/weigh the costs of potential errors
  - Type I (incorrectly claiming NI) vs. Type II errors (incorrectly failing to claim NI)
  - For regulatory submissions (in US)
    - 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
    - Envision using 2-sided 95% CI to evaluate NI
Sample Size

- Can be very large
- Smaller M implies larger sample size
- Generally believed to be larger than most superiority trials but this really depends on assumptions
- Stratification may help
  - Adjusted CIs are generally narrower than unadjusted CIs

Interim Analyses

- Reasonable to stop for futility (i.e., unable to show NI), superiority, inferiority, or ineffective active control
- Stopping for NI less likely
  - Difficult to show statistically without a lot of data (CIs need to be narrow)
  - Even if NI is observed, you may want to continue to see if superiority can be shown
  - No ethical reason to stop for NI (unless clear superiority is demonstrated)
- We often use repeated CIs (with group sequential error spending principles) and predicted intervals (Evans et.al., DIJ, 2007)
Analyses

- Confidence intervals used to “rule out” effects
  - Don’t rely on p-values
- NI margin plays a direct role in interpretation
- Check constancy assumption
- ITT and PP are important
**ITT vs. Per Protocol**

- ITT is not necessarily conservative
  - E.g., random lack of adherence can make treatments appear more similar
- For this reason, it is suggested that both ITT (evaluating strategy) and Per Protocol (evaluating biological effect) analyses be conducted
  - Sensitivity analyses are important

**Missing Data**

- Assess direction of bias from imputation
  - May wish to utilize a creative (biased) imputation method
    - 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 (maybe too much)
    - Some suggest imputation consistent with the null hypothesis
      - 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
In-Flight Design Changes?

- Switching from noninferiority to superiority
  - Generally OK to test for superiority after showing NI
  - No multiplicity adjustment necessary (closed testing procedure)
  - Use ITT (consistent with superiority evaluation)
  - Define plans \textit{a priori} when possible
- Switching from Superiority to noninferiority
  - Generally not acceptable to go from failing to demonstrate superiority to then evaluating NI
  - Post-hoc definition of NI margin is difficult to justify
  - Multiple testing concern too
- Changing the NI margin
  - Generally okay to decrease
  - Increases are difficult to justify

Interpretation

- A troubling observation is that levels of inferiority deemed to be clinically insignificant in the context of NI trials are often considered clinically significant in superiority trials

- For example, a NI margin of 10\% should imply that any difference less than 10\% between the test and control interventions has been predetermined to be clinically inconsequential. However, when a difference of less than 10\% is observed in a superiority trial, it can be portrayed as clinically meaningful.
“Equivalence” Trials

- Not too much more AND not too much less
  - 2-sided

- Common in PK studies (bioequivalence of drug levels)
  - Too little $\rightarrow$ not efficacious
  - Too much $\rightarrow$ toxic

Reporting

- Historical problems with reporting
  - Greene et.al. *AIM*, 2000
    - Reviewed reporting of 88 trials
      - 67% inappropriately concluded NI based on non-significant superiority tests
      - Only 23% pre-specified a NI margin

- Use CONSORT Statement Extension
  - Piaggio et.al., *JAMA*, 2006
Summary

- NI trials are complex and require careful design, conduct, analyses, and reporting
- NI CANNOT be concluded from nonsignificant tests for superiority
- The NI margin must be selected very carefully
- NI trials should not be conducted when the assumptions do not hold
  - Superiority trials are preferred
  - A new design strategy (DOOR) may avoid NI complexities in some settings

Useful References

- General Guidance
  - FDA Draft Guidance (March 2010).
- Discussion of Issues
  - Gau and Ware, SIM, 2007.
- Reporting
  - Greene et.al., AIM, 2000.
  - Piaggio et.al., JAMA, 2006.