Sequential Monitoring of Clinical Trials
Session 7 - Special Considerations: Futility and Non-Inferiority Decisions

Presented August 14-16, 2013

Daniel L. Gillen
Department of Statistics
University of California, Irvine

John M. Kittelson
Department of Biostatistics & Informatics
University of Colorado Denver

© 2013 Daniel L. Gillen, PhD and John M. Kittelison, PhD
RCTdesign

Designs meriting special consideration: Futility decisions

Futility decisions

- Futility designs in the unified family
- On the use of stochastic curtailment
- What is stochastic curtailment?
  - Stopping decisions based on conditional power
  - Stopping decisions based on predictive power
- Example: sepsis trial
Futility decisions in the unified family

- One-sided test \((\epsilon_{\ell} = 0, \epsilon_{u} = 1)\) means \(a_j = c_j\) and \(b_j = d_j\).
  - Set \(P_a = P_c < \infty\) and \(P_b - P_d = \infty\) to get a single \(a_j\) stopping boundary.
  - Choose \(P_a\) to control early conservatism.
- Properties evaluated as illustrated in session 3.
On the use of stochastic curtailment

Stochastic curtailment

Consider the sepsis trial:

- Suppose you observe a 5% higher mortality rate with antibody treatment at the first interim analysis ($\hat{\theta} = 0.05$). You are considering stopping for lack of effect.
- It might seem relevant to ask: “if I were to continue, would I change my mind?”
- More formally, what is the probability of reversing the current decision:
  
  $$P_{\theta}(\hat{\theta}_J < d^*_J | \hat{\theta}_1 = 0.05)?$$

  This is a form of power. If it is small, then you should stop.

The probability of changing your mind based on the data is known as **stochastic curtailment**. There are two types of approaches:

(a) **Conditional power**: Power based on some relevant value for $\theta$.

(b) **Predictive power**: Mean power integrated over a posterior distribution for $\theta$. 
### Conditional power example

- **Recall the sepsis trial:**
  - According to the monitoring plan, the trial would stop for benefit if $\hat{\theta}_J < d^*_J = -0.0435$ at $N_J = 1700$ patients.
  - If $\hat{\theta}_1 = 0.05$, then what is the chance (power) that $\hat{\theta}_J < d^*_J$?
  - If we observe $\hat{\theta}_j = x$, then the increment between the $j$th and $J$th analysis is normally distributed:

  \[ N_J \hat{\theta}_J - N_j x \sim \mathcal{N}[(N_J - N_j)\theta, (N_J - N_j)V] \]

  where $V = 0.23 \times 0.77 + 0.3 \times 0.7 = 0.387$. 
On the use of stochastic curtailment

Conditional power (example)

- It follows that:

\[ P_\theta(\hat{\theta}_J < d_j^* | \hat{\theta}_j = x) < \alpha \]

\[ \Rightarrow \frac{N_J d_j^* - N_j x - (N_J - N_j)\theta}{\sqrt{(N_J - N_j)V}} < z_\alpha \]

\[ \Rightarrow x > N_j^{-1} \left[ N_J d_j^* - (N_J - N_j)\theta - z_\alpha \sqrt{N_J - N_j} \right] \]

That is, we would be very unlikely (probability less than \( \alpha \)) to reverse a futility decision if \( \hat{\theta}_j > d_j^* \) where:

\[ d_j^* = N_j^{-1} \left[ N_J d_j^* - (N_J - N_j)\theta - z_\alpha \sqrt{N_J - N_j} \right] \]
On the use of stochastic curtailment

<table>
<thead>
<tr>
<th>Conditional power (example)</th>
</tr>
</thead>
</table>

- Suppose in the sepsis trial we used the following stopping criteria:
  - Final critical value is $d^*_J = -0.0435$.
  - Decide futility as long as the power for changing our mind is smaller than 10% ($z_{\alpha} = -1.28$).
  - Calculate this power under $\theta = 2 \times d_J = -0.087$.
- The following futility stopping criteria result from the above calculations:

\[
\begin{align*}
    d^*_1 &= 0.1539 \\
    d^*_2 &= 0.0273 \\
    d^*_3 &= -0.0161 \\
    d^*_4 &= -0.0435
\end{align*}
\]

Notice the degree conservatism. Is it ethical?
On the use of stochastic curtailment

Conditional power (example)

- Suppose we instead use $z_\alpha = 0$; i.e., a 50% chance of changing our mind. Then:

  \[
  \begin{align*}
  d_1^* &= 0.0870 \\
  d_2^* &= 0.0000 \\
  d_3^* &= -0.0290 \\
  d_4^* &= -0.0435 
  \end{align*}
  \]

- Notice the degree conservatism. Is it ethical?
- Do these stopping rules look familiar?
- Is it reasonable to use $z_\alpha = 0$?
On the use of stochastic curtailment

### Conditional power (design family)

- It is instructive to construct a conditional futility design family in the standardized scale:
  - Standardized mean: \( \delta = \frac{\theta - \theta_0}{\sqrt{V/N_J}} \)
  - Standardized statistic: \( \hat{\delta}_j \sim \mathcal{N}(\delta, \frac{1}{N_j}) \)
  - Decision criteria:
    - \( \delta_j > d_j \Rightarrow \text{Decide for superiority} \)
    - \( \delta_j < a_j \Rightarrow \text{Decide for lack of superiority} \)

Let \( d_J = a_J = G \).
On the use of stochastic curtailment

**Conditional power (design family)**

- Conditional power decision rules:
  - Stop for efficacy if $P_{\delta=0}(\hat{\delta}_J < G|\delta_j = d_j) = \alpha$, which implies
    \[ d_j = (G + z_{1-\alpha} \sqrt{1 - \Pi_j})\Pi_j^{-1} \]
  - Stop for futility if $P_{\delta=\delta_+}(\hat{\delta}_J > G|\delta_j = a_j) = \alpha$, which implies:
    \[ a_j = 2G - (G + z_{1-\alpha} \sqrt{1 - \Pi_j})\Pi_j^{-1} \]
  - As in other families, $G$ can be selected to control operating characteristics (see R-code in file session7fut.R).
On the use of stochastic curtailment

Conditional power (design family)

Example: Suppose $\Pi_j = 0.25, 0.5, 0.75, 1.0$. The values of $G$ satisfying:

\[
\sum_{j=1}^{J} P(\hat{\delta}_j > d_j|\delta = 0) = 0.025 \quad \text{(type I error)}
\]

\[
\sum_{j=1}^{J} P(\hat{\delta}_j < a_j|\delta = 2G) = 1 - 0.025 \quad \text{(type II error)}
\]

are:

<table>
<thead>
<tr>
<th>$\Pi_j$</th>
<th>$\alpha = 0.10$</th>
<th>$\alpha = 0.35$</th>
<th>$\alpha = 0.50$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>12.2815</td>
<td>9.2302</td>
<td>8.0129</td>
</tr>
<tr>
<td>0.50</td>
<td>5.7334</td>
<td>4.4926</td>
<td>4.0065</td>
</tr>
<tr>
<td>0.75</td>
<td>3.4684</td>
<td>2.8887</td>
<td>2.6710</td>
</tr>
<tr>
<td>1.00</td>
<td>1.9605</td>
<td>1.9739</td>
<td>2.0032</td>
</tr>
</tbody>
</table>
On the use of stochastic curtailment

Conditional power (design family)

▶ Stopping boundaries:

Stopping boundaries:

- CF.10
- CF.35
- CF.50

Sample size vs. Difference in Means

Fixed-sample non-Inferiority designs
Regulatory guidance
Group sequential non-Inferiority designs
On the use of stochastic curtailment

Conditional power (design family)

Sample Size:

Average Sample Size

75th percentile

Sample Size

Difference in Means

Sample Size

Difference in Means

CF.10

CF.35

CF.50

Difference in Means

Sample Size

Average Sample Size

Sample Size

Difference in Means

CF.10

CF.35

CF.50

Difference in Means
On the use of stochastic curtailment

Conditional power (design family)

- Power:

![Graph showing conditional power for different CF values (CF.10, CF.35, CF.50)]
On the use of stochastic curtailment

Conditional power (Sepsis example)
Comparing inference at lower (efficacy) boundary:

▶ Design: DSMB2

<table>
<thead>
<tr>
<th>IA (N)</th>
<th>$a_j$</th>
<th>$\hat{\delta}_j$</th>
<th>95% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ($N = 425$)</td>
<td>-0.175</td>
<td>-0.166</td>
<td>(-0.228, -0.090)</td>
<td>0.0000</td>
</tr>
<tr>
<td>2 ($N = 850$)</td>
<td>-0.087</td>
<td>-0.081</td>
<td>(-0.132, -0.027)</td>
<td>0.0020</td>
</tr>
<tr>
<td>3 ($N = 1275$)</td>
<td>-0.058</td>
<td>-0.055</td>
<td>(-0.097, -0.008)</td>
<td>0.0107</td>
</tr>
<tr>
<td>4 ($N = 1700$)</td>
<td>-0.044</td>
<td>-0.044</td>
<td>(-0.086, -0.001)</td>
<td>0.0225</td>
</tr>
</tbody>
</table>

▶ Design: Conditional Futility ($\alpha = 0.1$):

<table>
<thead>
<tr>
<th>IA (N)</th>
<th>$a_j$</th>
<th>$\hat{\delta}_j$</th>
<th>95% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ($N = 425$)</td>
<td>-0.266</td>
<td>-0.258</td>
<td>(-0.319, -0.182)</td>
<td>0.0000</td>
</tr>
<tr>
<td>2 ($N = 850$)</td>
<td>-0.124</td>
<td>-0.120</td>
<td>(-0.168, -0.064)</td>
<td>0.0000</td>
</tr>
<tr>
<td>3 ($N = 1275$)</td>
<td>-0.075</td>
<td>-0.072</td>
<td>(-0.114, -0.026)</td>
<td>0.0012</td>
</tr>
<tr>
<td>4 ($N = 1700$)</td>
<td>-0.043</td>
<td>-0.043</td>
<td>(-0.085, -0.001)</td>
<td>0.0225</td>
</tr>
</tbody>
</table>
On the use of stochastic curtailment

### Conditional power (Sepsis example)
Comparing inference at upper (futility) boundary:

#### Design: DSMB2

<table>
<thead>
<tr>
<th>IA (N)</th>
<th>$d_j$</th>
<th>$\hat{\delta}_j$</th>
<th>95% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ($N = 425$)</td>
<td>0.087</td>
<td>0.079</td>
<td>(0.003, 0.141)</td>
<td>0.9789</td>
</tr>
<tr>
<td>2 ($N = 850$)</td>
<td>0.000</td>
<td>-0.006</td>
<td>(-0.060, 0.044)</td>
<td>0.4013</td>
</tr>
<tr>
<td>3 ($N = 1275$)</td>
<td>-0.029</td>
<td>-0.032</td>
<td>(-0.079, 0.010)</td>
<td>0.0636</td>
</tr>
<tr>
<td>4 ($N = 1700$)</td>
<td>-0.044</td>
<td>-0.044</td>
<td>(-0.086, -0.001)</td>
<td>0.0225</td>
</tr>
</tbody>
</table>

#### Design: Conditional Futility ($\alpha = 0.1$):

<table>
<thead>
<tr>
<th>IA (N)</th>
<th>$d_j$</th>
<th>$\hat{\delta}_j$</th>
<th>95% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ($N = 425$)</td>
<td>0.180</td>
<td>0.172</td>
<td>(0.096, 0.234)</td>
<td>1.0000</td>
</tr>
<tr>
<td>2 ($N = 850$)</td>
<td>0.039</td>
<td>0.034</td>
<td>(-0.022, 0.083)</td>
<td>0.8825</td>
</tr>
<tr>
<td>3 ($N = 1275$)</td>
<td>-0.010</td>
<td>-0.014</td>
<td>(-0.060, 0.028)</td>
<td>0.2506</td>
</tr>
<tr>
<td>4 ($N = 1700$)</td>
<td>-0.043</td>
<td>-0.043</td>
<td>(-0.085, -0.001)</td>
<td>0.0225</td>
</tr>
</tbody>
</table>
On the use of stochastic curtailment

**Conditional power (Summary)**

- When compared with an O’Brien-Fleming design, the conditional power family has the following characteristics:
  - Nearly identical power
  - Much lower stopping probability at early interim analyses
  - Much larger ASN (loss of efficiency)
  - Extreme conservatism in making any early decision

- Resulting questions:
  - Is it ethical to continue in the presence of overwhelming evidence of harm (or benefit)?
  - Does conditional power obfuscate the essential clinical/scientific issues in deciding whether to terminate a trial?
On the use of stochastic curtailment

**Conditional power (Summary)**

- **Note:** There are “fixes” for the problems with conditional power:
  - Calculate conditional power under a different value of $\delta$; e.g.,:
    - Calculate $P_{\delta=\delta+/2}(\hat{\delta}_J < G|\hat{\delta}_j = d_j)$ instead of $P_{\delta=0}(\hat{\delta}_J < G|\hat{\delta}_j = d_j)$.
    - Calculate $P_{\delta=\delta+/2}(\hat{\delta}_J > G|\hat{\delta}_j = a_j)$ instead of $P_{\delta=\delta+}(\hat{\delta}_J > G|\hat{\delta}_j = a_j)$.
  - Calculate conditional power under the MLE $\hat{\delta}_j$:
    - Calculate $P_{\delta=\hat{\delta}_j}(\hat{\delta}_J < G|\hat{\delta}_j = d_j)$ instead of $P_{\delta=0}(\hat{\delta}_J < G|\hat{\delta}_j = d_j)$.
    - Calculate $P_{\delta=\hat{\delta}_j}(\hat{\delta}_J > G|\hat{\delta}_j = a_j)$ instead of $P_{\delta=\delta+}(\hat{\delta}_J > G|\hat{\delta}_j = a_j)$.
  - (Consider predictive power.)
On the use of stochastic curtailment

**Predictive power (overview)**

- Conditional power relied on a specific choice for $\delta$:
  - Efficacy: $P_{\delta=0}(\hat{\delta}_J < G | \hat{\delta}_j = d_j)$.
  - Futility: $P_{\delta=\delta_+}(\hat{\delta}_J > G | \hat{\delta}_j = a_j)$.

- Predictive power integrates over a posterior distribution for $\delta$:
  - Let $\lambda_0(\delta)$ denote a prior distribution for $\delta$.
  - Let $\lambda(\delta | \hat{\delta}_j)$ denote the posterior distribution of $\delta$ at the $j$th interim analysis.
  - Predictive probability:
    
    $$P(\hat{\delta}_J > G | \hat{\delta}_j) = \int_u P_{\delta=u}(\hat{\delta}_J > G | \hat{\delta}_j) \lambda(u | \hat{\delta}_j) du$$

- Decision criteria can be defined based on the magnitude of the predictive probability.
On the use of stochastic curtailment

Concluding remarks

- There are foundational issues with both conditional and predictive power. Neither frequentist nor Bayesian foundations construct inference around the probability of changing our mind (see paper).

- Underlying principles illustrated with this discussion of stochastic curtailment:
  - All designs can be expressed as conditions on the observed estimate.
  - *Always* consider statistical inference on the boundary when evaluating stopping criteria.
Designs meriting special consideration:
Non-inferiority designs

Non-inferiority designs

- Regulatory guidance (FDA)
- Example 1: Daptomycin
  - Fixed-sample design
  - Trial monitoring
  - Evaluation
- Example 2: Rivaroxaban in non-valvular atrial fibrillation
Recall reasons for early termination (session 3)

- **Why would you want to stop a study early?**
  - **Superiority study:**
    - For superiority (reject $H_0 : \theta \leq 0$)
    - For lack of superiority (reject $H_A : \theta \geq \theta_+$)
  - **Approximate equivalence study:**
    - For lack of inferiority (reject $H_0 : \theta \leq \theta_-$)
    - For lack of superiority (reject $H_A : \theta \geq \theta_+$)
  - **Non-inferiority study:**
    - For lack of inferiority (reject $H_0 : \theta \leq \theta_-$)
    - For inferiority (reject $H_A : \theta \geq 0$)
  - **Equivalence (2-sided) study:**
    - For superiority (reject $\theta \leq 0$)
    - For inferiority (reject $\theta \geq 0$)
    - For both non-inferiority and non-superiority (reject both $\theta \leq \theta_-$ and $\theta \geq \theta_+$)
### III. General Consideration of NI studies:

#### A. Basic principles of a NI study
1. Superiority vs NI
2. Logic of NI trial
3. Reasons for using NI design
4. The non-inferiority margin
5. Assay sensitivity and choosing NI margin
6. Regulatory conclusions

#### B. Practical Considerations for use of NI designs
1. Consider alternative designs
2. Number of studies needed
3. Statistical inferences
4. Choice of active control
5. Choice of NI method
III. General Consideration of NI studies:

A. Basic principles

A. Basic principles of a NI study

1. Superiority vs NI

   - *Superiority study*: Interpretable in their own right ... results speak for themselves without additional assumption.
   - *Non-inferiority study*: Requires knowing something that is not measured in the study; that is:
     The active control had its expected effect in the NI study.

   **Assay sensitivity** (think power):
   - Trial could have distinguished an effective from an ineffective drug.
   - Establishing NI is meaningless without assay sensitivity.
   - Knowing if the trial has assay sensitivity relies on external information.
III. General Consideration of NI studies:

2. Logic of the NI trial
   - *Superiority study:* establish significant benefit over placebo:

   Figure 1: Three Possible Results of a Placebo-Controlled Superiority Study
   (Point Estimate, 95% CI)

   1. Point estimate of effect is 2; 95% CI lower bound is 1. Conclusion: Drug is effective and appears to have an effect of at least 1.
   2. Point estimate of effect is 2; 95% CI lower bound is <0 (study perhaps too small). Conclusion: Drug is not shown to be effective.
   3. Point estimate of effect is 0; 95% CI lower bound is well below 0. Conclusion: Drug shows no suggestion of effectiveness.
III. General Consideration of NI studies:

2. Logic of the NI trial

▶ *Non-inferiority study*: benefit at least as large as control superiority:

![Figure 2: Results of NI Study Showing C-T and 95% CI (M1 = 2)](image)

1. Point estimate of C-T is 0, suggesting equal effect; upper bound of the 95% CI for C-T is 1, well below M1; NI is demonstrated.
2. Point estimate of C-T favors C; upper bound of the 95% CI for C-T is >2, well above M1; NI is not demonstrated.
3. Point estimate of C-T is zero, suggesting equal effect, but upper bound of the 95% CI for C-T is >2 (i.e., above M1), so that NI is not demonstrated.
4. Point estimate favors T; NI is demonstrated, but superiority is not demonstrated.
5. Point estimate favors T; superiority and NI are demonstrated.
6. Point estimate of C-T favors C and C is statistically significantly superior to T. Nonetheless, upper bound of the 95% CI for C-T<2 (M1), so that NI is also demonstrated for the NI margin M1. (This outcome would be unusual and could present interpretive problems.)
III. General Consideration of NI studies:

2. Logic of the NI trial
  ▶ Non-inferiority study: Choice of non-inferiority margin ($M_1$)
    (1) Treatment effect based on historical experience with the active control drug
    (2) Assessment of the likelihood that the current effect of the active control is similar to the past effect (the constancy assumption)
    (3) Assessment of the quality of the NI trial, particularly looking for defects that could produce bias toward the null.
      [Note: for this reason the size of $M_1$ cannot usually be assessed until the trial is finished.]
III. General Consideration of NI studies:

▶ “Constancy Assumption"
  ▶ Historic evidence of sensitivity to drug effects (HESDE) (ICH E-10)
    ▶ HESDE means that past trials that used a specific active treatment regularly showed this treatment to be superior to placebo.
  ▶ The conclusion that HESDE applies only when the NI study is sufficiently similar to the past studies:
    ▶ Characteristics of the patient population.
    ▶ Concomitant treatments.
    ▶ Definitions and ascertainment of study endpoints
    ▶ Dose of active control, entry criteria, analytic approaches.
III. General Consideration of NI studies:

- Summary comments
  - Beware of ‘bias toward the null’ in a non-inferiority trial:
    - Inflated variance (⇒ poor assay sensitivity):
      - Outcome ascertainment
      - Adequate blinding (central adjudication)
      - Compliance
    - Missing data!!
  - “Biocreep” or “Noninferiority creep”:
    - Repeated comparisons to latest newest treatment results in steady drift to inferiority relative to placebo.
  - Avoid NI trials whenever possible.
    - “If it is ethical to conduct a placebo-controlled trial, then it is unethical not to” (Lloyd Fisher).
Non-inferiority trials

▶ Two examples:
  ▶ Daptomycin for Staph infection
    ▶ Trial results: Fowler, VG, et.al. NEJM (2006); 355(7):653-65
    ▶ Editorial: Grayson, ML. NEJM (2006); 355(7):724
  ▶ Rivaroxaban for atrial fibrillation
    ▶ Trial results: Patel, MR, et.al. NEJM (2011); 365: 883-91
    ▶ Editorial: Fleming, TR, Emerson SS. NEJM (2011); 365(17): 1557
Non-inferiority designs

Example: Daptomycin *NEJM* (2006); 355(7):653-65

Structuring parameter space:

- **Background**: Patients with bacteremia and endocarditis caused by *S. aureus* need alternative antibiotic therapies. Current antibiotic treatments are successful in about 65% of the cases.

- **Clinical question**: Can daptomycin be used as another approach for treating *S. aureus* bacteremia?

- **Study interventions**: open label assignment to
  - Standard antibiotic care
  - Daptomycin

- **Primary outcome**: Clinical success at 42 days (failure = no response to drug based on ongoing signs and symptoms)

- **Functional**: $\theta_1, \theta_0$: population success probability with daptomycin and standard care

- **Contrast**: $\theta = \theta_1 - \theta_0$
Non-inferiority designs

Example: Daptomycin *NEJM* (2006); 355(7):653-65

**Trial information:**

- **Non-inferiority definition:** “Lower bound of the 95% confidence intervals must be within the prespecified noninferiority margin of 20 percent and the upper bounds containing 0 percent.”

- **Sample size:** 90 patients per group gives 80% power for NI if $\theta_0 = \theta_1 = 0.65$.

- **Sample size evaluation:** If $\theta_0 = \theta_1 = 0.65$, then the 95% CI has half-width:

$$1.96\sqrt{(0.65 \times 0.35) \frac{2}{90}} = 0.14$$
Structuring parameter space: Non-inferiority designs

▶ What should we decide with an infinite sample size ($\theta$ is known)?

Potential trial results (infinite sample size)

- No Difference
- Clinically Important Benefit
- Clinically Important Harm
- Clinical Inferiority
- Clinical Superiority

- $\theta_-$
- $\theta_+$
- $\theta_0$
- $\theta^*$

---

W UW Grp Seq - Sec 7 - sld 34
Structuring parameter space: Non-inferiority designs

**Example: Daptomycin** *NEJM (2006); 355(7):653-65*

- What should we decide with an infinite sample size ($\theta$ known)?
  - A, B, C, D $\Rightarrow$ Superior (use daptomycin)
  - E $\Rightarrow$ Unimportant inferiority (use daptomycin?)
  - F, G, H $\Rightarrow$ Inferior (do not use daptomycin)

- May be willing to accept some inferiority:
  - Diversity of antibiotics may avoid evolution of resistant strains
  - Unresponsive patients may respond to a different treatment
Structuring parameter space: Non-inferiority designs

Example: Daptomycin *NEJM (2006); 355(7):653-65*

- Other considerations
  - In the marginal situations the acceptability of daptomycin would also consider:
    * The clinical “importance” of a difference, which is difficult to quantify.
    * Nature of the primary endpoint (e.g., survival vs treatment success).
    * Effects on secondary endpoints (toxicity, rescue therapies)?
    * Potential for long-term effects that could not be measured in this trial.
  - Bottom line:
    * There are grey-areas in parameter space.
    * These will affect the design (e.g., particularly in non-inferiority trials).
  - Remark: A question that cannot be answered with an infinite amount of data cannot be answered with a finite amount of data.
Structuring parameter space: Non-inferiority designs

▶ What should we decide with a finite sample size
(θ is estimated with uncertainty)?

Potential trial results (confidence intervals)

<table>
<thead>
<tr>
<th></th>
<th>No Difference</th>
<th>Clinically Important Benefit</th>
<th>Clinically Important Harm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inferiority</td>
<td>θ &lt; θ -</td>
<td>θ &gt; θ +</td>
<td>θ &lt; θ -</td>
</tr>
<tr>
<td>Superiority</td>
<td>θ &gt; θ +</td>
<td>θ &gt; θ +</td>
<td>θ &lt; θ -</td>
</tr>
</tbody>
</table>

A: No Difference
B: Clinically Important Benefit
C: Clinically Important Harm
D: Inferiority
E: Superiority
F: θ < θ -
G: θ > θ +
H: Clinical Inferiority
I: Clinical Superiority

Potential trial results (confidence intervals)
**Structuring parameter space: Non-inferiority designs**

<table>
<thead>
<tr>
<th>Example: Daptomycin</th>
<th>NEJM (2006); 355(7):653-65</th>
</tr>
</thead>
</table>

- What should we decide with an infinite sample size ($\theta$ known)?
  - A, B, C, D $\Rightarrow$ Superior (use daptomycin)
  - E, F $\Rightarrow$ Unimportant inferiority (use daptomycin?)
  - G, H $\Rightarrow$ Significant inferiority (do not use daptomycin)
Structuring parameter space: Non-inferiority designs

Superiority/non-superiority studies

▶ **Application**: Evaluating new therapies to determine if they increase benefits.

▶ **Defining Hypotheses**:
  - **Inferiority**: $\theta \leq \theta_\emptyset$
    - (Decide for superiority if inferiority is rejected)
  - **Clinical Superiority**: $\theta \geq \theta_+$
    - (Decide against superiority if clinical superiority is rejected)

▶ **Example**: Suppose we hypothesize daptomycin is *superior* to standard care and we want to discriminate between $\theta \leq \theta_\emptyset$ and $\theta \geq \theta_+$:

  - Inferiority: $\theta \leq \theta_\emptyset = 0.0$
  - Clinical superiority: $\theta \geq \theta_+ = 0.28$
Structuring parameter space: Non-inferiority designs

Non-inferiority/inferiority studies

► Applications:
  ► Evaluating new therapies to determine if they are non-inferior to existing therapy.
  ► Evaluating existing therapies to determine if they are harmful.

► Defining Hypotheses:
  ► Clinical Inferiority: \( \theta \leq \theta_\cdot \)  
    (Decide for non-inferiority if clinical inferiority is rejected)
  ► Superiority: \( \theta \geq \theta_\emptyset \)  
    (Decide against non-inferiority if superiority is rejected)

► Example: Suppose daptomycin is in routine use, but there is some evidence it may be harmful, and we want to discriminate between \( \theta \leq \theta_\cdot \) and \( \theta \geq \theta_\emptyset \):

Clinical Inferiority: \( \theta \leq \theta_\cdot = -0.28 \)
Superiority: \( \theta \geq \theta_\emptyset = 0.0 \)
Structuring parameter space: Non-inferiority designs

Equivalence studies

- **Application:**
  - Evaluating therapies to determine if they can be used interchangeably.
  - (If not, then to select the best therapy.)

- **Defining Hypotheses:**
  - Clinical inferiority ($\theta \leq \theta_-$)
  - Clinical superiority ($\theta \geq \theta_+$)
  - Equality ($\theta = \theta_0$)

- **Decisions**
  - Decide equivalence if reject both clinical inferiority and superiority.
  - Decide treatment A better than B if reject inferiority (of A).
  - Decide treatment B better than A if reject superiority (of A).

- **Example:** Daptomycin vs standard therapy:
  - Clinical Inferiority (of daptomycin): $\theta \leq -0.28$
  - Clinical Superiority (of daptomycin): $\theta \geq 0.28$
  - Equality: $\theta = 0$
Types of equivalence (ironically: not all equivalents are the same)

- Categories of equivalence studies include:
  - Bioequivalence: showing that two formulations of the same drug are equivalent.

\[
\begin{align*}
\text{Reject } \theta & \geq \theta_0 \quad \Rightarrow \quad \text{Conclude inferior} \\
\text{Reject } \theta & \leq \theta_0 \quad \Rightarrow \quad \text{Conclude superior} \\
\text{Reject } \theta & \leq \theta_- \and \theta \geq \theta_+ \quad \Rightarrow \quad \text{Conclude equivalence}
\end{align*}
\]

In some settings FDA defines bioequivalence as \(0.8 < \theta_1/\theta_0 < 1.25\).
Structuring parameter space: Non-inferiority designs

Types of equivalence (ironically: not all equivalents are the same)

- Categories of equivalence studies include:
  - Non-Inferiority:
    - Reject $\theta \geq \theta_0 \Rightarrow$ Conclude inferior
    - Reject $\theta \leq \theta_-$ $\Rightarrow$ Conclude non-inferior
  - Approximate equivalence: Require that observed effect be beneficial and important inferiority is rejected:
    - Observe $\hat{\theta} > \theta_0 \Rightarrow$ Conclude non-inferior
    - Reject $\theta \leq \theta_- \Rightarrow$ Conclude non-inferior
Notes (see figure below):

Non-inferiority creep:

- Suppose treatment A replaces an existing treatment in a strict non-inferiority design (as above).
  The point estimate may indicate inferior ($\hat{\theta} < \theta_0$)
- Now suppose treatment B replaces treatment A in a strict non-inferiority design.
  The point estimate may indicate inferior ($\hat{\theta} < \theta_0$)
- If this process continues, there may be a steady creep that results in increasingly inferior treatments.
- Confidence interval A on the following figure illustrates how an inferior point estimate could produce a non-inferior decision.
Non-inferiority trials of new therapeutics: public health implications

- Notes (see figure below):
  - Approximate equivalence designs:
    - Requiring a positive point estimate (confidence interval B) would avoid non-inferiority creep.
    - Requiring a positive point estimate and requiring rejection of $\theta \leq \theta_-$ would allow a smaller sample size (confidence interval C).
    - Note: Sample size may be relaxed somewhat, but not as much as the difference illustrated in confidence intervals B and C.
    - Note: Requiring a positive point estimate means that there is only 50% power for a treatment that is truly equivalent ($\theta = \theta_0$).
Structuring parameter space: Non-inferiority designs

Non-inferiority trials of new therapeutics: public health implications

- Daptomycin example (see figure below)
  - Approximate equivalence design (confidence interval B)
    - Non-inferiority bound = −14%.
    - Sample size: \( N = 180 \) (90 per group)  
      \( \Rightarrow \) critical value = 0%.
  - Reported trial design (confidence interval D)
    - Non-inferiority bound = −20%.
    - Sample size: \( N = 180 \) (90 per group)  
      \( \Rightarrow \) critical value = −6%.
  - Actual trial results (confidence interval E)
    - \( N = 235 \) (\( N = 120 \) daptomycin; \( N = 115 \) standard care)
    - Point estimate: 2.4%
    - CI: -10.2% to 15.1%
Structuring parameter space: Non-inferiority designs

- Potential ways to specify non-inferiority decisions

**Potential trial results**

- **Daptomycin trial**

**Clinical Superiority**

- \( \theta > \theta^+ \)

**Clinical Inferiority**

- \( \theta < \theta^- \)

**No Difference**

- \( \theta = \emptyset \)

**Clinically Important Harm**

- \( \theta^- \)

**Clinically Important Benefit**

- \( \theta^+ \)
## Structuring parameter space: Non-inferiority designs

### Bottom line

- Design of equivalence trials (including non-inferiority trials) requires careful consideration of the standards that define the decision.
- Poorly run equivalence trials can introduce bias toward the null hypothesis (an equivalence conclusion). *(See editorial with daptomycin paper.)*
Comparison of designs (illustrated using daptomycin):
- Two-sided bioequivalence design
- Shifted superiority design
- Approximate equivalence design
- Approximate equivalence with a single boundary
Monitoring plan for non-inferiority trial

A not-so-hypothetical example

Stage 1: Investigator discovers bioequivalence

- Investigator hypothesizes that daptomycin might work for sepsis:
  - “Knows" it cannot be superior, so wants to show it is equivalent to standard care
  - Finds a website about bioequivalence and when he plugged in his numbers it told him:
    * Use 90 patients per group and measure the difference in success proportions.
    * Conclude equivalence if the 95% CI for the difference in proportions is contained entirely between -0.28 and 0.28.

- A research assistant says he should really use a non-inferiority design.

- A clinical trialist says he should worry about non-inferiority creep.

- Investigator is seeking help from the biostatisticians - everyone is confused ...
Monitoring plan for non-inferiority trial

A not-so-hypothetical example

Stage 2: The research assistant says...

- A non-inferiority website says the design has 80% power for equivalence.
- You end up having to explain/compare these approaches, and begin with the fixed-sample design:
  - 2-sided bioequivalence design
    (dap.2side)
  - 1-sided superiority design
    (dap.fixSup)
  - 1-sided shifted superiority design (eventually you reproduce their design)
    (dap.fixShift)
  - 1-sided “approximate equivalence” design
    (dap.fixNI)
Monitoring plan for non-inferiority trial

A not-so-hypothetical example

Stage 3: Enter the biostatistician

dap.2side <- seqDesign(prob.model='proportions', arms=2, null.hypothesis=c(0.65, 0.65), alt.hypothesis="calculate", test.type='two.sided', variance="null", alpha=0.025, sample.size=180, power=0.975, nbr.analyses=1)

dap.fixSup <- seqDesign(prob.model='proportions', arms=2, null.hypothesis=c(0.65, 0.65), alt.hypothesis="calculate", test.type='greater', variance="null", alpha=0.025, sample.size=180, power=0.975, nbr.analyses=1)

e <- 0.715
dap.fixShift <- update(dap.fixSup, epsilon=c(e, 1-e))

dap.fixNI <- update(dap.fixSup, epsilon=c(0.5, 0.5))
Monitoring plan for non-inferiority trial

- Design critical values (fixed sample designs):

<table>
<thead>
<tr>
<th>Design Name</th>
<th>$a_J$</th>
<th>$b_J$</th>
<th>$c_J$</th>
<th>$d_J$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(dap.2side)</td>
<td>-0.14</td>
<td>-0.14</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>(dap.fixSup)</td>
<td>0.14</td>
<td>NA</td>
<td>NA</td>
<td>0.14</td>
</tr>
<tr>
<td>(dap.fixShift)</td>
<td>-0.06</td>
<td>NA</td>
<td>NA</td>
<td>0.06</td>
</tr>
<tr>
<td>(dap.fixNI)</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>0</td>
</tr>
</tbody>
</table>

- Inference:
  95% CI have width ±0.14 centered at the critical value
Monitoring plan for non-inferiority trial

- Power curves (fixed-sample designs:

\[
\begin{align*}
\text{dap.fixSup} & \quad \text{(blue line)} \\
\text{dap.fixShift} & \quad \text{(red line)} \\
\text{dap.fixNI} & \quad \text{(black line)}
\end{align*}
\]

\[(\text{dap.2side and dap.fixSup have same upper power curve})\]
Monitoring plan for non-inferiority trial

A not-so-hypothetical example

Stage 4: Study section requests interim analyses

- Add interim analyses:

```r
dap.2sideIA <- update(dap.2side,sample.size=c(60,120,180))
dap.SupIA <- update(dap.fixSup,sample.size=c(60,120,180))
dap.ShiftIA <- update(dap.fixShift,sample.size=c(60,120,180))
dap.NI.IA <- update(dap.fixNI,sample.size=c(60,120,180))
```
Monitoring plan for non-inferiority trial

A not-so-hypothetical example

Stage 5: Focus switches to inferiority decision

- Ethics demands rapid termination for inferiority
- Increasing sensitivity of lower ($a_j$) boundary

```r
dap.NI.IAasym <- update(dap.fixNI, sample.size=c(60,120,180),
P=c(0.8,Inf,Inf,1))
```

- Now the steering committee decides it is only necessary to stop for inferiority

```r
dap.NI.IAinf <- update(dap.fixNI, sample.size=c(60,120,180),
P=c(0.8,Inf,Inf,Inf))
```
Monitoring plan for non-inferiority trial

A not-so-hypothetical example

Stage 6: Evaluate design properties

- Stopping boundaries, power, ASN, stopping probabilities, inference at boundaries

```r
seqPlotBoundary(dap.NI.IA, dap.NI.IAasym, dap.NI.IAinf, fixed=F, col=c(1,2,4))
seqPlotPower(dap.NI.IA, dap.NI.IAasym, dap.NI.IAinf, fixed=F, col=c(1,2,4))
seqPlotASN(dap.NI.IA, dap.NI.IAasym, dap.NI.IAinf, fixed=F, col=c(1,2,4), prob=NULL)
```
Monitoring plan for non-inferiority trial
Properties: stopping boundaries

- dap.NI.IA
- dap.NI.IAasym
- dap.NI.IAinf

Sample Size

Difference in Proportions

-0.4 -0.2 0.0 0.2 0.4 0.6
Monitoring plan for non-inferiority trial
Properties: power

Difference in Proportions

Power (Upper)

- dap.NI.IA
- dap.NI.IAasym
- dap.NI.IAinf
Monitoring plan for non-inferiority trial
Properties: ASN

Average Sample Size

Difference in Proportions
Example 2: Rivaroxaban is a new anticoagulation therapy

Rivaroxaban

- Anticoagulation therapy is used in a wide range of settings:
  - Atrial fibrillation for prevention of thrombus and/or stroke
  - Thromboprophylaxis in patients undergoing orthopedic surgery (knee or hip replacement).
  - Cancer chemotherapy patients
- Standard anticoagulation therapy = warfarin (coumadin)
  - Warfarin has a narrow therapeutic range:
    - Too much causes bleeding, too little does not prevent clots.
    - Patients on warfarin must monitor their clotting (INR)
- Rivaroxaban is a new oral anticoagulant
  - Easier to control amount of anticoagulation
  - Might be preferred over warfarin if it:
    - Has equal ability to reduce blood clots/strokes.
    - Does not increase bleeding risk.
    - Clinical trials for a wide range of indications are ongoing.
Example 2: ROCKET-AF trial

Rivaroxaban for non-valvular atrial fibrillation

- **Atrial Fibrillation:**
  - AF is the most common cardiac arrhythmia; symptoms include a “racing” heart.
  - Can cause incomplete emptying of the left atrium.
  - Pooled blood in the atrium can lead to clots and stroke.
  - It is common to treat atrial fibrillation with anticoagulation therapy to prevent clots.

- **Standard anticoagulation therapy = warfarin (coumadin)**

- **Can Rivaroxaban be used instead of warfarin for AF?**
  - ROCKET-AF: large clinical trial to evaluate noninferiority of rivaroxaban when compared with warfarin.
  - **Key references:**
    - Trial design: *Am Heart J* 2010; 159: 340-347.e1.
    - Trial results: *Patel, MR, et.al. NEJM* (2011); 365: 883-91
    - Perspective: *Fleming, TR, Emerson SS. NEJM* (2011); 365(17): 1557
Example 2: ROCKET-AF trial

ROCKET-AF design synopsis

- **Patient population:**
  - Patients with non-valvular AF at risk for stroke.

- **Treatment groups:**
  - **Warfarin (standard care):**
    - Titrated to target INR of 2.5 (range 2.0-3.0).
    - Plus oral rivaroxaban placebo
  - **Rivaroxaban: 20mg od:**
    - Warfarin placebo od titrated to sham INR of 2.5 (range 2.0-3.0).

All patients have screening period, double-blind treatment period, posttreatment observation period
All patients treated according to local standard of care.

- **Endpoints**
  - Primary efficacy endpoint: composite of all-cause stroke and systemic embolism.
  - Primary safety endpoint: composite of major and clinically relevant non-major bleeding.
Example 2: ROCKET-AF trial

Analysis and sample size

- Analysis plan:
  - Per protocol population (not intention to treat)
    - PP instead of ITT is not uncommon in noninferiority trials.
    - Lack of compliance or failure to complete treatment can introduce bias toward the null.
  - JK: both PP and ITT results should be reported.
  - JK: missing data *must* be avoided!
  - Primary efficacy analysis seeks 363 events:
    - 363 events gives 95% power for hazard ratio of $\theta = 1.46$ in a one-side level $\alpha = 0.025$ logrank test:
      \[ 363 = \left[ \frac{1.96 + 1.645}{\log(1.46)} \right]^2 \times 4 \]
  - Plan to enroll 14,000 patients to get 405 events:
    - Allow for 14% dropout
    - Allow for “robust evaluation across subgroups”
Example 2: ROCKET-AF trial

Noninferiority margin

- Noninferiority margin:
  - Previous meta-analysis of 6 trials comparing warfarin to placebo:
    - Relative risk (warfarin/placebo): 0.38 (95% CI: 0.28-0.52).
    - Relative risk (placebo/warfarin): 2.63 (95% CI: 1.92-3.57).
  - Seek to preserve 50% of previous warfarin benefit:
    - Calculation of previous benefit (subtract 1 from above RR):
      $$1.63 \ (95\% \ CI \ : \ 0.92, 2.57)$$
    - Half of previous benefit (add 1 to half of above):
      $$1.82 \ (95\% \ CI \ : \ 1.46, 2.29)$$
    - “The most conservative approach was chosen, selecting the lower limit, 1.46, of this confidence boundary.”
**Example 2: ROCKET-AF trial**

### Sample size evaluation

- **Asymptotic standard error** is $se = \sqrt{\frac{4}{363}} = 0.105$:

- **Superiority decision**:
  - Critical value: $e^{-1.96 \times 0.105} = 0.814$
  - CI at critical value: $e^{-3.92 \times 0.105, 0} = (0.661, 1.0)$
  - Concluding non-superiority rules out HR ($\theta$) less than 0.661.
  - Concluding superiority rules out HR ($\theta$) larger than 1.0.

- **Noninferiority decision**:
  - CI when $\hat{\theta} = 1.0$: $e^{\pm 1.96 \times 0.105} = (0.814, 1.23)$
  - CI when $\hat{\theta} = 1.123$: $e^{\log(1.89) \pm 1.96 \times 0.105} = (0.914, 1.380)$

- **Inferiority decision**:
  - CI when $\hat{\theta} = 1.228$: $e^{\log(1.228) \pm 1.96 \times 0.105} = (1.00, 1.509)$
Example 2: ROCKET-AF trial results

**Trial results:** *Patel, MR, et.al. NEJM (2011); 365: 883-91*

- **Primary analysis event rates:**
  - Rivaroxaban: (188 patients) 1.7% per year
  - Warfarin: (241 patients) 2.2% per year
  - Hazard ratio: 0.79 (95% CI: 0.66 to 0.96); \( p < 0.001 \) for noninferiority.

- **Intention-to-treat event rates:**
  - Rivaroxaban: (269 patients) 2.1% per year
  - Warfarin: (306 patients) 2.4% per year
  - Hazard ratio: 0.88 (95% CI: 0.74 to 1.03); \( p = 0.12 \) for superiority.

- **Bleeding risk (major and non-major clinically relevant):**
  - Rivaroxaban: (1475 patients) 14.9% per year
  - Warfarin: (1449 patients) 14.5% per year
  - Hazard ratio: 1.03 (95% CI: 0.96 to 1.11); \( p = 0.44 \).
### Example 2: ROCKET-AF Regulatory considerations

**Perspective:** *Fleming and Emerson. NEJM (2011); 365(17): 1557*

- **Design synopsis:**
  - NI margin = 1.38: rivaroxaban be superior to placebo by at least 50% of the margin by which warfarin is superior to placebo.
  - Per-protocol “on-treatment” analysis pre-specified because of concerns about assay sensitivity (bias toward null).

- **FDA issues in interpretation:**
  - Constancy assumption must be satisfied: control treatment with same magnitude of benefit as in reference trials.
  - Concerns over non-constancy:
    - Higher risk patients enrolled in ROCKET-AF
    - More than 5% who discontinued due to withdrawal of consent.
    - INR in therapeutic range (between 2 and 3) only 55% of the time; other trials have 62-73% in therapeutic range.
  - Per-randomization analysis should also be reported. On treatment analysis is subject to bias:
    - Observations truncated at 2 days after treatment discontinuation.
Summary Remarks: Non-inferiority Trials

Non-Inferiority Trials

- General principles in non-inferiority trials
  - Constancy assumption
    * The effect of the comparator relative to placebo remains constant across all trials
    * May break down because of changes other aspects such as:
      - ancillary care
      - diagnostic methods and patient populations
  - Non-inferiority is difficult to objectively define
    * The “non-inferiority bound“:
      - depends on the setting (population, treatment, endpoints)
      - may be driven by cost/feasibility as opposed to public health
  - Sloppy science my bias toward non-inferiority
    * Assure compliance
    * Endpoints must be clinically important/relevant
    * Avoid surrogates

- Design principles
  - Designs should consider potential inference (CI) rather than power
  - Monitoring designs should evaluate inference on the boundaries