4. Issues in Trial Monitoring

4.1 Elements of Trial Monitoring

- Monitoring trial process and quality
- Infrastructure requirements and DSMBs

4.2 Interim analyses: group sequential trial design

4.3 Group sequential design families

4.4 Frequentist evaluation of sequential trials

- Evaluation criteria:
  - Expected sample size (ASN)
  - Power
  - Stopping probabilities
  - Inference at the boundary

- Example: Sepsis Trial

4.5 On the use of stochastic curtailment

4.6 Case study
4.4 Frequentist evaluation (inference at the boundary)

- Recall the discussion of section 3.2 (fixed-sample design evaluation) when information is pre-specified.
- In this setting we evaluated the information by examining the inference at the boundary. Specifically:
  - Critical value: Threshold for declaring that the treatment works.
  - Hypotheses discriminated (interval estimate):
    - What hypothesis will be rejected if the results are significant?
    - What hypothesis will be rejected if the results are not significant?
  - P-value at the boundary?
- Inference at the boundary is also of interest when evaluating a group sequential design.
  - This inference must be adjusted for the bias induced by the stopping rule.
  - The function `seqInference` gives bias-adjusted inference at the boundary.
Example:

- One-sided test with 2-sided stopping (Pocock vs OBF boundaries):

  poc <- seqDesign(prob.model = "normal", arms = 1,
                  null.hypothesis = 0., alt.hypothesis = 3.92, variance = 1.,
                  nbr.analyses = 5, test.type = "greater",
                  power = 0.975, alpha = 0.025, epsilon = c(0., 1.),
                  early.stopping = "both",
                  display.scale = seqScale(scaleType = "X"),P=0.5)
  obf <- update(poc, P=1)
  poc.infr <- seqInference(poc)
  obf.infr <- seqInference(obf)
### 4.4 Frequentist evaluation (inference at the boundary)

**Example (Pocock-type decision criteria)**

- **Lower boundary:**

<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 ($N = 0.293$)</td>
<td>-0.463</td>
<td>0.026</td>
<td>(-2.749, 3.189)</td>
<td>0.495</td>
</tr>
<tr>
<td>2 ($N = 0.587$)</td>
<td>0.821</td>
<td>1.150</td>
<td>(-1.020, 3.656)</td>
<td>0.146</td>
</tr>
<tr>
<td>3 ($N = 0.880$)</td>
<td>1.390</td>
<td>1.565</td>
<td>(-0.290, 3.826)</td>
<td>0.047</td>
</tr>
<tr>
<td>4 ($N = 1.173$)</td>
<td>1.729</td>
<td>1.801</td>
<td>(-0.036, 3.899)</td>
<td>0.027</td>
</tr>
<tr>
<td>5 ($N = 1.466$)</td>
<td>1.960</td>
<td>1.960</td>
<td>(0.000, 3.920)</td>
<td>0.025</td>
</tr>
</tbody>
</table>

- **Upper boundary:**

<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 ($N = 0.293$)</td>
<td>4.383</td>
<td>3.894</td>
<td>(0.731, 6.669)</td>
<td>0.009</td>
</tr>
<tr>
<td>2 ($N = 0.587$)</td>
<td>3.099</td>
<td>2.770</td>
<td>(0.264, 4.940)</td>
<td>0.016</td>
</tr>
<tr>
<td>3 ($N = 0.880$)</td>
<td>2.530</td>
<td>2.355</td>
<td>(0.094, 4.210)</td>
<td>0.021</td>
</tr>
<tr>
<td>4 ($N = 1.173$)</td>
<td>2.191</td>
<td>2.119</td>
<td>(0.021, 3.956)</td>
<td>0.024</td>
</tr>
<tr>
<td>5 ($N = 1.466$)</td>
<td>1.960</td>
<td>1.960</td>
<td>(0.000, 3.920)</td>
<td>0.025</td>
</tr>
</tbody>
</table>
4.4 Frequentist evaluation (inference at the boundary)

Example (Pocock-type decision criteria)

- Characteristics:
  - Maximal sample size is 47% larger in order to maintain power.
  - At maximal sample size, bias-adjusted inference at final analysis is the same as a fixed-sample design:
    - Final p-value = 0.025.
    - Critical value = 1.96.
    - CI discriminates between null hypothesis ($\delta \leq 0$) and 97.5% alternative ($\delta \geq 3.92$).
  - At interim analyses:
    - Bias-adjustment shrinks estimates toward 1.96.
    - Superiority is decided when $\delta \leq 0$ (lower end of bias-adjusted CI $\approx 0$).
    - Non-superiority is decided when $\delta \geq 3.92$ (upper end of bias-adjusted CI $\approx 3.92$).
4.4 Frequentist evaluation (inference at the boundary)

Example (O’Brien-Fleming-type decision criteria)

- **Lower boundary:**

<table>
<thead>
<tr>
<th>IA (N) (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 ($N = 0.211$)</td>
<td>-5.880</td>
<td>-5.451</td>
<td>(-8.597, -1.576)</td>
<td>0.997</td>
</tr>
<tr>
<td>2 ($N = 0.422$)</td>
<td>-0.980</td>
<td>-0.674</td>
<td>(-3.235, 2.103)</td>
<td>0.682</td>
</tr>
<tr>
<td>3 ($N = 0.633$)</td>
<td>0.653</td>
<td>0.891</td>
<td>(-1.303, 3.203)</td>
<td>0.212</td>
</tr>
<tr>
<td>4 ($N = 0.844$)</td>
<td>1.470</td>
<td>1.571</td>
<td>(-0.310, 3.709)</td>
<td>0.049</td>
</tr>
<tr>
<td>5 ($N = 1.055$)</td>
<td>1.960</td>
<td>1.960</td>
<td>(0.000, 3.920)</td>
<td>0.025</td>
</tr>
</tbody>
</table>

- **Upper boundary:**

<table>
<thead>
<tr>
<th>IA (N) (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 ($N = 0.211$)</td>
<td>9.800</td>
<td>9.371</td>
<td>(5.496, 12.517)</td>
<td>0.000</td>
</tr>
<tr>
<td>2 ($N = 0.422$)</td>
<td>4.900</td>
<td>4.594</td>
<td>(1.817, 7.155)</td>
<td>0.001</td>
</tr>
<tr>
<td>3 ($N = 0.633$)</td>
<td>3.267</td>
<td>3.029</td>
<td>(0.717, 5.223)</td>
<td>0.005</td>
</tr>
<tr>
<td>4 ($N = 0.844$)</td>
<td>2.450</td>
<td>2.349</td>
<td>(0.211, 4.230)</td>
<td>0.015</td>
</tr>
<tr>
<td>5 ($N = 1.055$)</td>
<td>1.960</td>
<td>1.960</td>
<td>(0.000, 3.920)</td>
<td>0.025</td>
</tr>
</tbody>
</table>
4.4 Frequentist evaluation (inference at the boundary)

Example (O’Brien-Fleming-type decision criteria)

- Characteristics:
  - Maximal sample size is 6% larger in order to maintain power.
  - At maximal sample size, bias-adjusted inference at final analysis is the same as a fixed-sample design:
    - Final $p$-value $= 0.025$.
    - Critical value $= 1.96$.
    - CI discriminates between null hypothesis ($\delta \leq 0$) and 97.5% alternative ($\delta \geq 3.92$).
  - At interim analyses:
    - Bias-adjustment shrinks estimates toward 1.96.
    - The OBF design is extremely conservative at early analyses:
      - Futility corresponds to significant harm at first analysis (reject $\delta \geq -1.58$)
      - Superiority decision reached when reject $\delta \leq 5.50$
      - Require $p < 0.000004$ for superiority decision
4.4 Frequentist evaluation (example: sepsis trial)

- **Setting**
  - Sepsis is a severe system infection that is often fatal.
  - A new antibody to an endotoxin that is produced by gram-negative bacteria was developed to treat sepsis.
  - Randomized double-blind placebo-controlled trial was planned to evaluate if the antibody improves survival.
Statistical Design

- Primary outcome:
  - 28-day mortality
  - \( X_{ki} \sim B(1, \rho_k) \) for \( ith \) patient in treatment group \( k = 0, 1 \)

- Within-group summary measure:
  - Proportion dying before 28 days
  - \( \hat{\rho}_k = \frac{\sum X_{ki}}{N} \)

- Between-group contrast:
  - Difference in proportions
  - \( \theta = \rho_1 - \rho_0 \)

- Design hypotheses (1-sided superiority test):
  
  Null: \( \rho_1 - \rho_0 \geq 0 \)
  
  Alternative: \( \rho_1 - \rho_0 \leq -0.07 \)

- Sample size: 1700 patients (850 per group) gives power 0.907 for \( \theta = -0.07 \) if \( \rho_0 = 0.3 \).
Possible conclusions upon trial completion

- Clinically Important Benefit
- No Difference
- Clinically Important Harm
Possible conclusions upon trial completion

- Clinically Important Benefit
- No Difference
- Clinically Important Harm
- Superior
- Inferior
- Important Superiority
- Important Inferiority
Possible conclusions upon trial completion

- E, F ⇒ Use new antibody
- A, B, C, D ⇒ Do not use new antibody

Diagram:

- Clinically Important Benefit
- No Difference
- Clinically Important Harm
- Superior
- Inferior
- Important Superiority
- Important Inferiority

Potential CI upon trial completion:

- A
- B
- C
- D
- E
- F

Use new antibody if E, F
Do not use new antibody if A, B, C, D
Possible conclusions at interim analysis

- F ⇒ Stop: use new antibody
- D, E ⇒ Continue trial
- A, B, C ⇒ Stop: do not use new antibody
Example: Sepsis Trial

Recall setting:

- ICU patient with gram-negative sepsis
- Randomized to receive a single injection of antibody to bacterial endotoxin
- Outcome: 28-day mortality
- Design hypotheses:

  \[ \begin{align*}
  \text{Null:} & \quad \rho_1 - \rho_0 \geq 0 \\
  \text{Alternative:} & \quad \rho_1 - \rho_0 \leq -0.07
  \end{align*} \]

- Sample size: 1700 patients (850 per group) gives power 0.907 for \( \theta = -0.07 \) if \( \rho_0 = 0.3 \).
- Test based on 2-sided level \( \alpha = 0.045 \) test because of pre-planned subgroup analysis.
Example: Sepsis Trial

1. FDA requires an interim safety analysis half way through trial. Sponsor design:
   - One-sided level 0.05 test for a significant *increase* in mortality due to the antibody injection.
   - (The efficacy decision criteria are still based on a 2-sided $\alpha = 0.045$; i.e., $|Z| > 2.0047$.)
   - Decision rules (on Z-scale):

<table>
<thead>
<tr>
<th>Interim Analysis</th>
<th>Decide Benefit</th>
<th>Decide Lack of Benefit</th>
<th>Decide Harm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ($N = 850$)</td>
<td>$Z &gt; 1.645$</td>
<td>$-2.005 \leq Z \leq 2.005$</td>
<td>$Z &gt; 2.005$</td>
</tr>
<tr>
<td>2 ($N = 1700$)</td>
<td>$Z &lt; -2.005$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Trial Monitoring ()  4.4 Frequentist Evaluation  11 April 2011  15 / 38
Sponsor safety design: ASN

Average Sample Size

75th percentile

Sample Size

Difference in Proportions

Sample Size

Difference in Proportions

4. Trial Monitoring ()  4.4 Frequentist Evaluation  11 April 2011  16 / 38
- Sponsor safety design: Power
Sponsor safety design: Power relative to fixed-sample design
Notes on sponsor safety boundary

- Notice increased power for detecting harm. Type I error for concluding significant harm (rejecting $\theta_1 - \theta_0 \leq 0$) is larger than 0.0225.
- Notice that expected size of the trial (ASN) is much smaller for harmful effects.
2. DSMB safety design:
   - DSMB decides it is not ethical to continue the trial to establish harm if benefit can be ruled out (i.e., a one-sided superiority test is more appropriate than a 2-sided test).
   - DSMB considers a one-sided level 0.0225 test for lack of benefit.
   - DSMB considers a single-boundary design with 4 interim analyses and OBF boundary shape (sample mean scale):

<table>
<thead>
<tr>
<th>Interim Analysis</th>
<th>Decide Benefit</th>
<th>Decide Lack of Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ((N = 425))</td>
<td>0.0900</td>
<td></td>
</tr>
<tr>
<td>2 ((N = 850))</td>
<td>0.0018</td>
<td></td>
</tr>
<tr>
<td>3 ((N = 1275))</td>
<td>-0.0276</td>
<td>-0.0423</td>
</tr>
<tr>
<td>4 ((N = 1700))</td>
<td>-0.0423</td>
<td>-0.0423</td>
</tr>
</tbody>
</table>
Sepsis DSMB safety design: stopping boundaries
Sepsis DSMB safety design: power
Sepsis DSMB safety design: ASN

Average Sample Size

75th percentile
Notes on DSMB safety boundary

- Notice increased power for detecting harm:
  - Type I error for concluding significant harm (rejecting $\theta_1 - \theta_0 \leq 0$) is 0.9775.
  - Power when $\theta_1 - \theta_0 \leq -0.07$ is 0.9 (probability of type II error = 0.1).

- Notice that expected size of the trial (ASN) is much smaller for harmful effects than the sponsor safety design.
Example: Sepsis Trial

3. DSMB efficiency design:
   ▶ DSMB decides to evaluate stopping for benefit in addition to lack of benefit.
   ▶ DSMB evaluates adding a lower (efficacy) boundary with an OBF boundary shape:

<table>
<thead>
<tr>
<th>Interim Analysis</th>
<th>Decide Benefit</th>
<th>Decide Lack of Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ((N = 425))</td>
<td>-0.1745</td>
<td>0.0872</td>
</tr>
<tr>
<td>2 ((N = 850))</td>
<td>-0.0872</td>
<td>0.0000</td>
</tr>
<tr>
<td>3 ((N = 1275))</td>
<td>-0.0582</td>
<td>-0.0291</td>
</tr>
<tr>
<td>4 ((N = 1700))</td>
<td>-0.0436</td>
<td>-0.0436</td>
</tr>
</tbody>
</table>
Sepsis DSMB efficiency design: stopping boundaries
Sepsis DSMB efficiency design: Power
- Sepsis DSMB efficiency design: power relative to DSMB safety boundary
Sepsis DSMB efficiency design: ASN

**Average Sample Size**

- dsmb2
- dsmb
- fixed

**75th percentile**

- dsmb2
- dsmb
- fixed

4. Trial Monitoring
4.4 Frequentist Evaluation

11 April 2011 29 / 38
Notes on DSMB efficiency boundary

- Negligible effects on power when compared with DSMB safety boundary.
- Now the expected sample size is also reduced for beneficial effects.
Example: Sepsis Trial

4. Sponsor futility boundaries
   - When the sponsor saw the efficiency gains using a futility boundary, they wanted to evaluate being less conservative when making the futility decision.
   - They evaluated 3 different boundary shapes corresponding to $P = 0.9$, $P = 0.75$, and $P = 0.5$. (Larger values of $P$ give more early conservatism.)
   - The following are boundaries for $P = 0.9$:

     | Interim Analysis | Decide Benefit | Decide Lack of Benefit |
     |------------------|----------------|------------------------|
     | 1 ($N = 425$)    | -0.1739        | 0.0662                 |
     | 2 ($N = 850$)    | -0.0869        | -0.0052                |
     | 3 ($N = 1275$)   | -0.0580        | -0.0304                |
     | 4 ($N = 1700$)   | -0.0435        | -0.0435                |
Sponsor futility designs: stopping boundaries
Sponsor futility designs: ASN

Average Sample Size

75th percentile

4. Trial Monitoring ()  4.4 Frequentist Evaluation  11 April 2011
- Sponsor futility designs: power

![Graph showing power vs. theta for different futility designs.]

- Trial Monitoring

- Frequentist Evaluation

- April 2011
Sponsor futility designs: power relative to efficiency design
Inference at the boundary: DSMB efficiency design (OBF upper)

**Inference corresponding to futility boundary**

- Difference in Proportions
  - Sample Size
  - 8.72e−02
  - 1.86e−09
  - 2.91e−02
  - 4.36e−02

**Inference corresponding to efficacy boundary**

- Difference in Proportions
  - Sample Size
  - −0.1745
  - −0.0872
  - −0.0582
  - −0.0436
Inference at the boundary: Sponsor futility design (Pocock upper)

Inference corresponding to futility boundary

Inference corresponding to efficacy boundary

Sample Size

Difference in Proportions

Observed

Adjusted
Notes on sponsor futility boundary

- Decreasing early conservatism gave smaller ASN for unimportant benefits.
- Decreasing early conservatism also reduces power for efficacy.