Sequential Monitoring

Choice of Stopping Rules

Where am I going?

• A wide variety of stopping rules have been proposed for different RCT settings.

• Families of designs have been described on a variety of statistical scales.
### Statistical Criteria

- Extreme estimates of treatment effect
- Statistical significance (Frequentist)
  - At final analysis: Curtailment
  - Based on experimentwise error
    - Group sequential rule
    - Error spending function
- Statistical credibility (Bayesian)
- Probability of achieving statistical significance / credibility at final analysis
  - Condition on current data and presumed treatment effect

### Working Example

- Fixed sample two-sided tests
  - Test of a two-sided alternative \((\theta_1 > \theta_0 > \theta_2)\)
    - Upper Alternative: \(H_+ : \theta \geq \theta_+\) (superiority)
    - Null: \(H_0 : \theta = \theta_0\) (equivalence)
    - Lower Alternative: \(H_- : \theta \leq \theta_-\) (inferiority)
  - Decisions:
    - Reject \(H_0, H_-\) (for \(H_+\)) \(\iff T \geq c_U\)
    - Reject \(H_+, H_-\) (for \(H_0\)) \(\iff c_L \leq T \leq c_U\)
    - Reject \(H_+\), \(H_0\) (for \(H_-\)) \(\iff T \leq c_L\)

### Sampling Plan: General Approach

- Perform analyses when sample sizes \(N_1 \ldots N_J\)
  - Can be randomly determined
- At each analysis choose stopping boundaries
  - \(a_j < b_j < c_j < d_j\)
- Compute test statistic \(T_j = T(X_1 \ldots X_{N_j})\)
  - Stop if \(T_j < a_j\) (extremely low)
  - Stop if \(b_j < T_j < c_j\) (approximate equivalence)
  - Stop if \(T_j > d_j\) (extremely high)
- Otherwise continue (maybe adaptive modification of analysis schedule, sample size, etc.)
  - Boundaries for modification of sampling plan

### Choice of Stopping Rule

- The choice of stopping rule will vary according to the exact scientific and clinical setting for a clinical trial
  - Each clinical trial poses special problems
  - Wide variety of stopping rules needed to address the different situations
  - (One size does not fit all)
Unified Family: MLE Scale

- Down columns: Early stopping vs no early stopping
- Across rows: One-sided vs two-sided decisions

Impact on Sampling Density

- When using a stopping rule, the sampling density depends on exact stopping rule
  - This is obvious from what we have already seen.
  - A fixed sample test is merely a particular stopping rule:
    - Gather all N subjects’ data and then stop

Compared to Fixed Sample

- The magnitude of the effect of the stopping rule on trial design operating characteristics and statistical inference can vary substantially
  - Rule of thumb:
    - The more conservative the stopping rule at interim analyses, the less impact on the operating characteristics and statistical inference when compared to fixed sample designs.
Reasons for Early Stopping

• Efficacy, Futility, Harm

• Ethical
  – Individual
    • Protect patients on study
    • Protect patients who might be accrued to study
  – Group
    • Promote rapid discovery of new treatments

• Economic
  – Avoid unnecessary costs of RCT
  – Facilitate earlier marketing

Role of Futility Boundaries

• When clinically relevant improvement has been convincingly ruled out and no further useful information to be gained
  – (Is further study of subgroups or other endpoints still in keeping with informed consent?)

• Futility boundaries usually do not indicate harm

• Because most RCT do not reject the null hypothesis, the major savings in early stopping are through a futility boundary
  – Also, not as much need for early conservatism

Potential Issue

• Compared to a stopping rule with no futility boundary
  – The critical value at the final analysis can be lower
    • Some of the trials stopped early for futility might have otherwise been type I errors at the final analysis
    • Depends on the early conservatism of the futility boundary

Nonbinding Futility

• Some clinical trialists believe that FDA requires that the futility rule be ignored when making inference
  – Such builds in conservatism
  – True type I error is smaller than nominal
  – True power is smaller than normal

• This is purposely using the wrong sampling density
  – Not good statistics—game theory must be motivation
### Correct Inference

- The statistically correct, efficient approach is to base inference on the real futility boundary
  - Demands correct pre-specification of the futility boundary
  - Demands a clear paper trail of analyses performed

### Boundary Scales

- Stopping rule for one test statistic is easily transformed to a rule for another statistic
  - “Group sequential stopping rules”
    - Sum of observations
    - Point estimate of treatment effect
    - Normalized (Z) statistic
    - Fixed sample P value
    - Error spending function
  - Bayesian posterior probability
  - Stochastic Curtailment
    - Conditional probability
    - Predictive probability

### Correspondence Among Scales

- Choices for test statistic $T_j$
  - All of those choices for test statistics can be shown to be transformations of each other
  - Hence, a stopping rule for one test statistic is easily transformed to a stopping rule for a different test statistic
  - We regard these statistics as representing different scales for expressing the boundaries

### Relative Advantages

- Which is the best scale to view a stopping rule?
  - Maximum likelihood estimate
  - Z score / fixed sample P value
  - Error spending scale
  - Stochastic curtailment
    - Conditional power
    - Predictive power
Statistics Used In Science

- "Scientific scales"
  - Summary measures of the effect
    - Means, medians, geometric means, proportions...
  - Interval estimates for those summary measures
    - (Probabilities merely used to characterize the definition of the interval)
- "Statistical scales"
  - The precision with which you know the true effect
    - Power, P values, posterior probabilities
  - Predictions of the sample you will obtain
    - Conditional power, predictive power

Example

- Pre-hospital emergency setting
  - Severe trauma
- Waiver of informed consent
  - Effectiveness studies
  - Impact on prisoners, minors, DOD
  - Notification of participants
- Treatment in field
  - Hospital care according to current local standards
  - Largely passive collection of hospital data

Hypertonic Resuscitation

- Hypertonic saline +/- dextran vs normal saline
  - Osmotic pressure to restore blood volume
  - Modulation of immune response during reperfusion
- Hypovolemic shock
  - SBP ≤ 70 OR SBP < 90 and HR > 108
  - Proportion alive at 28 days
    - 4.8% absolute improvement (69.4% vs 64.6%)

Sample Size

- Multiple comparison issue
  - HSD vs NS
  - HS vs NS
- Bonferroni adjustment
  - One-sided level 0.0125 tests
- Experimentwise power: 80%
  - Each comparison has 62.6% power
- Sample size: 3,726
  - 1 HSD : 1 HS : 1.414 NS
Noninferiority

- Department of Defense
  - 250 cc HS weighs less than 2,000 cc NS
  - Even if no benefit from HS, may want to use if not inferior to NS
- Proving noninferior
  - Define margin of "unacceptably inferior"
    - Absolute decrease of 3%
  - CI at end of trial must exclude the margin
    - 80% confidence interval

Okay, so far?

- 4.8% improvement in 28 day survival
  - 28 day survival clinically relevant?
  - 4.8% improvement clinically important?
  - Realistic based on prior knowledge?
- Experimentwise errors
  - HS and HSD clinically equivalent?
  - 0.025 type I error, 80% power statistically credible?

Okay, so far?

- Noninferiority
  - 3% decrease justified? In civilians?
  - 80% confidence interval reasonable standard?
  - Are we answering the DoD’s questions?
    - (Additional fluids not restricted)
  - Sample size of 3,726 without consent?

Statistical Sampling Plan

- Ethical and efficiency concerns are addressed through sequential sampling
  - During the conduct of the study, data are analyzed at periodic intervals and reviewed by the DMC
  - Using interim estimates of treatment effect
    - Decide whether to continue the trial
    - If continuing, decide on any modifications to
      - scientific / statistical hypotheses and/or
      - sampling scheme
Protocol Stopping Rule

<table>
<thead>
<tr>
<th></th>
<th>N Accrue</th>
<th>Z</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>621</td>
<td>-4.000</td>
<td>6.000</td>
</tr>
<tr>
<td>Second</td>
<td>1,242</td>
<td>-2.800</td>
<td>4.170</td>
</tr>
<tr>
<td>Third</td>
<td>1,863</td>
<td>-1.800</td>
<td>3.350</td>
</tr>
<tr>
<td>Fourth</td>
<td>2,484</td>
<td>-1.200</td>
<td>2.860</td>
</tr>
<tr>
<td>Fifth</td>
<td>3,105</td>
<td>-0.700</td>
<td>2.540</td>
</tr>
<tr>
<td>Sixth</td>
<td>3,726</td>
<td>-0.290</td>
<td>2.290</td>
</tr>
</tbody>
</table>

Efficacy Boundary

<table>
<thead>
<tr>
<th></th>
<th>N Accrue</th>
<th>Z</th>
<th>Crude Diff</th>
<th>Est (95% CI; One-sided P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>621</td>
<td>6.000</td>
<td>0.272</td>
<td>0.263 (0.183, 0.329); P &lt; 0.0001</td>
</tr>
<tr>
<td>Second</td>
<td>1,242</td>
<td>4.170</td>
<td>0.134</td>
<td>0.129 (0.070, 0.181); P &lt; 0.0001</td>
</tr>
<tr>
<td>Third</td>
<td>1,863</td>
<td>3.350</td>
<td>0.088</td>
<td>0.082 (0.035, 0.129); P &lt; 0.0004</td>
</tr>
<tr>
<td>Fourth</td>
<td>2,484</td>
<td>2.860</td>
<td>0.065</td>
<td>0.060 (0.019, 0.102); P &lt; 0.0025</td>
</tr>
<tr>
<td>Fifth</td>
<td>3,105</td>
<td>2.540</td>
<td>0.052</td>
<td>0.048 (0.010, 0.085); P &lt; 0.0070</td>
</tr>
<tr>
<td>Sixth</td>
<td>3,726</td>
<td>2.290</td>
<td>0.042</td>
<td>0.040 (0.005, 0.078); P &lt; 0.0130</td>
</tr>
</tbody>
</table>

Futility Boundary

<table>
<thead>
<tr>
<th></th>
<th>N Accrue</th>
<th>Z</th>
<th>Crude Diff</th>
<th>Est (95% CI; One-sided P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>621</td>
<td>-4.000</td>
<td>-0.181</td>
<td>-0.172 (-0.238, -0.092); P &gt; 0.9999</td>
</tr>
<tr>
<td>Second</td>
<td>1,242</td>
<td>-2.800</td>
<td>-0.090</td>
<td>-0.084 (-0.137, -0.026); P = 0.9973</td>
</tr>
<tr>
<td>Third</td>
<td>1,863</td>
<td>-1.800</td>
<td>-0.047</td>
<td>-0.041 (-0.088, 0.006); P = 0.9581</td>
</tr>
<tr>
<td>Fourth</td>
<td>2,484</td>
<td>-1.200</td>
<td>-0.027</td>
<td>-0.022 (-0.064, 0.019); P = 0.8590</td>
</tr>
<tr>
<td>Fifth</td>
<td>3,105</td>
<td>-0.700</td>
<td>-0.014</td>
<td>-0.010 (-0.048, 0.028); P = 0.7090</td>
</tr>
<tr>
<td>Sixth</td>
<td>3,726</td>
<td>-0.290</td>
<td>-0.005</td>
<td>-0.003 (-0.041, 0.032); P = 0.5975</td>
</tr>
</tbody>
</table>

Sequential Monitoring

Evaluation of Designs

Where am I going?

- RCT design is most often an iterative process that involves
  - Defining an initial design,
  - Evaluating its operating characteristics, and
  - Modifying the design to better address constraints.
Evaluation of Designs

- Process of choosing a trial design
  - Define candidate design
    - Usually constrain two operating characteristics
      - Type I error, power at design alternative
      - Type I error, maximal sample size
    - Evaluate other operating characteristics
      - Different criteria of interest to different investigators
      - Modify design
      - Iterate

Which Operating Characteristics

- The same regardless of the type of stopping rule
  - Frequentist power curve
    - Type I error (null) and power (design alternative)
  - Sample size requirements
    - Maximum, average, median, other quantiles
    - Stopping probabilities
  - Inference at study termination (at each boundary)
    - Frequentist or Bayesian (under spectrum of priors)
  - (Futility measures
    - Conditional power, predictive power)

Collaboration of Disciplines

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Collaborators</th>
<th>Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scientific</td>
<td>Epidemiologists</td>
<td>Hypothesis generation</td>
</tr>
<tr>
<td></td>
<td>Basic Scientists</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clinical Scientists</td>
<td></td>
</tr>
<tr>
<td>Clinical</td>
<td>Experts in disease / treatment</td>
<td>Efficacy of treatment</td>
</tr>
<tr>
<td></td>
<td>Experts in complications</td>
<td>Adverse experiences</td>
</tr>
<tr>
<td>Ethical</td>
<td>Ethicists</td>
<td>Individual ethics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Group ethics</td>
</tr>
<tr>
<td>Economic</td>
<td>Health services</td>
<td>Cost effectiveness</td>
</tr>
<tr>
<td></td>
<td>Sponsor management</td>
<td>Cost of trial / Profitability</td>
</tr>
<tr>
<td></td>
<td>Sponsor marketers</td>
<td>Marketing appeal</td>
</tr>
<tr>
<td>Governmental</td>
<td>Regulators</td>
<td>Safety</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Efficacy</td>
</tr>
<tr>
<td>Statistical</td>
<td>Biostatisticians</td>
<td>Estimates of treatment effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Precision of estimates</td>
</tr>
<tr>
<td>Operational</td>
<td>Study coordinators</td>
<td>Collection of data</td>
</tr>
<tr>
<td></td>
<td>Data management</td>
<td>Study burden</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data integrity</td>
</tr>
</tbody>
</table>

At Design Stage

- In particular, at design stage we can know
  - Conditions under which trial will continue at each analysis
    - Estimates
      - (Range of estimates leading to continuation)
    - Inference
      - (Credibility of results if trial is stopped)
    - Conditional and predictive power
  - Tradeoffs between early stopping and loss in unconditional power
Operating Characteristics

- For any stopping rule, however, we can compute the correct sampling distribution with specialized software
  - From the computed sampling distributions we then compute
    - Bias adjusted estimates
    - Correct (adjusted) confidence intervals
    - Correct (adjusted) P values
  - Candidate designs are then compared with respect to their operating characteristics

Case Study: Clinical Trial In Gm- Sepsis

- Randomized, placebo controlled Phase III study of antibody to endotoxin
  - Intervention: Single administration
  - Endpoint: Difference in 28 day mortality rates
    - Placebo arm: estimate 30% mortality
    - Treatment arm: hope for 23% mortality
  - Analysis: Large sample test of binomial proportions
    - Frequentist based inference
    - Type I error: one-sided 0.025
    - Power: 90% to detect $\theta < -0.07$
    - Point estimate with low bias, MSE; 95% CI

Evaluation: Sample Size

- Number of subjects is a random variable
  - Quantify summary measures of sample size distribution as a function of treatment effect
    - maximum (feasibility of accrual) (Sponsor)
    - mean (Average Sample N- ASN) (Sponsor, DMC)
    - median, quartiles
  - Stopping probabilities (Sponsor)
    - Probability of stopping at each analysis as a function of treatment effect
    - Probability of each decision at each analysis

Sample Size

- What is the maximal sample size required?
  - Planning for trial costs
  - Regulatory requirements for minimal N treated
- What is the average sample size required?
  - Hopefully low when treatment does not work or is harmful
  - Acceptable to be high when uncertainty of benefit remains
  - Hopefully low when treatment is markedly effective
    - (But must consider burden of proof)
ASN Curve

- Expected sample size as function of true effect

Evaluation: Power Curve

- Probability of rejecting null for arbitrary alternatives
  - Level of significance (power under null)
  - Power for specified alternative
  - Alternative rejected by design
    - Alternative for which study has high power
      - Interpretation of negative studies

Evaluation: Boundaries

- Decision boundary at each analysis: Value of test statistic leading to early stopping
  - On the scale of estimated treatment effect
    - Inform DMC of precision
    - Assess ethics
      - May have prior belief of unacceptable levels
    - Assess clinical importance
  - On the Z or fixed sample P value scales

Evaluation: Inference

- Inference on the boundary at each analysis
  - Frequentist
    - Adjusted point estimates
    - Adjusted confidence intervals
    - Adjusted P values
  - Bayesian
    - Posterior mean of parameter distribution
    - Credible intervals
    - Posterior probability of hypotheses
    - Sensitivity to prior distributions
Frequentist Inference

<table>
<thead>
<tr>
<th></th>
<th>O'Brien-Fleming</th>
<th>Pocock</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>MLE</td>
</tr>
<tr>
<td><strong>Efficacy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>425</td>
<td>0.171</td>
<td>-0.163</td>
</tr>
<tr>
<td>850</td>
<td>-0.086</td>
<td>0.000</td>
</tr>
<tr>
<td>1275</td>
<td>-0.037</td>
<td>-0.054</td>
</tr>
<tr>
<td>1700</td>
<td>-0.043</td>
<td>-0.043</td>
</tr>
<tr>
<td><strong>Futility</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>425</td>
<td>0.086</td>
<td>0.077</td>
</tr>
<tr>
<td>850</td>
<td>0.000</td>
<td>-0.006</td>
</tr>
<tr>
<td>1275</td>
<td>-0.029</td>
<td>-0.031</td>
</tr>
<tr>
<td>1700</td>
<td>-0.043</td>
<td>-0.043</td>
</tr>
</tbody>
</table>

At Design Stage: Example

- With O’Brien-Fleming boundaries having 90% power to detect a 7% absolute decrease in mortality
  - Maximum sample size of 1700
  - Continue past 1275 if crude difference in 28 day mortality is between -2.9% and -5.7%
  - If we just barely stop for efficacy after 425 patients we will report
    - Estimated difference in mortality: -16.3%
    - 95% confidence interval: -8.7% to -22.4%
    - One-sided lower P < 0.0001

Z scale, Fixed Sample P value

- Error Spending Functions
  - My view: Poorly understood even by the researchers who advocate them
    - There is no such thing as THE Pocock or O’Brien-Fleming error spending function
      - Depends on type I or type II error
      - Depends on number of analyses
      - Depends on spacing of analyses
OBF, Pocock Error Spending

- Error spending functions depend on the alternative used to compute them
  - The same design has many error spending functions
- JSM 2009: Session on early stopping for harm in a noninferiority trial
  - Attempts to use error spending function approach
  - How to calibrate with functions used for lack of benefit?

Error Spent by Alternative

- Consider the probability that a different decision would result if trial continued
  - Compare unconditional power to fixed sample test with same sample size
- Conditional power
  - Assume specific hypotheses
  - Assume current best estimate
- Predictive power
  - Assume Bayesian prior distribution

Evaluation: Futility
Efficiency / Unconditional Power

- Tradeoffs between early stopping and loss of power
  - Boundaries
  - Loss of Power
  - Avg Sample Size

But What If?

- It is common for people to ask about the possibility of a reversed decision
  - But suppose we did not stop for futility. What would be the probability of getting a significant result if we continued to the maximal sample size
- This is easily computed conditional on the observed results.
  - Conditional power: Assume a particular effect
  - Predictive power: Use a Bayesian prior distribution

Stochastic Curtailment

- Stopping boundaries chosen based on predicting future data
- Probability of crossing final boundary
  - Frequentist: Conditional Power
    - A Bayesian prior with all mass on a single hypothesis
  - Bayesian: Predictive Power

Stochastic Curtailment

- Boundaries transformed to conditional or predictive power
  - Key issue: Computations are based on assumptions about the true treatment effect
    - Conditional power
      - "Design": based on hypotheses
      - "Estimate": based on current estimates
    - Predictive power
      - "Prior assumptions"
So What?

• Why not use stochastic curtailment?
  – What treatment effect should we presume?
    • Hypothesis rejected; current estimate?
  – What threshold should be used for a “low” probability
    • Choice of thresholds poorly understood
    – 10%, 20%, 50%, 80%?
    – How should it depend on sample size and treatment effect
  – Inefficient designs result
    • Conditional and predictive power do not correspond directly to unconditional power

Assumed Effect and Threshold

• Probability threshold should take into account the timing of the analysis and the presumed treatment effect
  – It is not uncommon for naïve users to condition on a treatment effect that has already been excluded

Predictive Power: Example 1

• Sepsis trial to detect difference in 28 day survival: Null 0.00 vs Alt -0.07 (90% power)
• Futility boundary at first of 4 analyses
  – Futile if observed diff > 0.0473 (so wrong direction)
  – Inference at boundary
    • Bias adjusted: 0.038 (95% CI -0.037 to 0.101)

Predictive Power: Example 1

• MLE: 0.0473  Bias Adj: 0.038 (CI: -0.037, 0.101)

<table>
<thead>
<tr>
<th>Presumed Effect</th>
<th>Predictive Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Effect</td>
<td>Power</td>
</tr>
<tr>
<td>-0.086</td>
<td>71.9%</td>
</tr>
<tr>
<td>-0.070</td>
<td>43.2%</td>
</tr>
<tr>
<td>-0.037</td>
<td>10.3%</td>
</tr>
<tr>
<td>Spons prior</td>
<td>2.8%</td>
</tr>
<tr>
<td>Flat prior</td>
<td>0.8%</td>
</tr>
<tr>
<td>0.047</td>
<td>&lt;0.005%</td>
</tr>
</tbody>
</table>
**Predictive Power: Ex 2 (OBF)**

- Sepsis trial to detect difference in 28 day survival: Null 0.00 vs Alt -0.07 (90% power)
- Futility boundary at first of 4 analyses
  - Futile if observed diff > 0.0855 (so wrong direction)
  - Inference at boundary
- Bias adjusted: 0.077 (95% CI 0.000 to 0.139)

**Predictive Power: Ex 2 (OBF)**

- MLE: 0.0855  Bias Adj: -0.077 (CI: 0.000, 0.139)

<table>
<thead>
<tr>
<th>Presumed</th>
<th>Predictive</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Effect</td>
<td>Power</td>
</tr>
<tr>
<td>-0.086</td>
<td>50.0%</td>
</tr>
<tr>
<td>-0.070</td>
<td>26.5%</td>
</tr>
<tr>
<td>0.000</td>
<td>0.03%</td>
</tr>
<tr>
<td>Spon prior</td>
<td>0.3%</td>
</tr>
<tr>
<td>Flat prior</td>
<td>0.03%</td>
</tr>
<tr>
<td>0.085</td>
<td>&lt;0.005%</td>
</tr>
</tbody>
</table>

**Key Issues**

- Very different probabilities based on assumptions about the true treatment effect
  - Extremely conservative O’Brien-Fleming boundaries correspond to conditional power of 50% (!) under alternative rejected by the boundary
  - Resolution of apparent paradox: if the alternative were true, there is less than .003 probability of stopping for futility at the first analysis

**Stopping Probs for \( \theta = -0.07 \)**

<table>
<thead>
<tr>
<th>Group Sequential test</th>
<th>Efficacy</th>
<th>Futility</th>
</tr>
</thead>
<tbody>
<tr>
<td>N= 425</td>
<td>0.009</td>
<td>&lt; -0.170</td>
</tr>
<tr>
<td>N= 850</td>
<td>0.298</td>
<td>&lt; -0.085</td>
</tr>
<tr>
<td>N= 1275</td>
<td>0.401</td>
<td>&lt; -0.057</td>
</tr>
<tr>
<td>N= 1700</td>
<td>0.179</td>
<td>&lt; -0.042</td>
</tr>
<tr>
<td>Total</td>
<td>0.888</td>
<td>0.112</td>
</tr>
</tbody>
</table>
Apples with Apples

- Can compare a group sequential rule to a fixed sample test providing
  - Same maximal sample size (N= 1700)
  - Same (worst case) average sample size (N= 1336)
  - Same power under the alternative (N= 1598)
- Consider probability of “discordant decisions”
  - Conditional probability (conditional power)
  - Unconditional probability (power)

Cond/Uncond Comparison

- Probability of achieving the opposite result at the final analysis
  - Conditional probability
    - Probability among all studies that would stop at that analysis
  - Unconditional probability
    - Change in power of the test due to early stopping

<table>
<thead>
<tr>
<th>Group Sequential test</th>
<th>Efficacy</th>
<th>Futility</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cond</td>
<td>Uncond</td>
</tr>
<tr>
<td>N= 425</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>N= 850</td>
<td>0.003</td>
<td>0.001</td>
</tr>
<tr>
<td>N= 1275</td>
<td>0.009</td>
<td>0.004</td>
</tr>
<tr>
<td>N= 1700</td>
<td>0.094</td>
<td>0.017</td>
</tr>
<tr>
<td>Total</td>
<td>0.024</td>
<td>0.022</td>
</tr>
</tbody>
</table>

Ordering of the Outcome Space

- Choosing a threshold based on conditional power can lead to nonsensical orderings based on unconditional power
  - Decisions based on 35% conditional power may be more conservative than decisions based on 18% conditional power
  - Can result in substantial inefficiency (loss of power)

Further Comments

- Neither conditional power nor predictive power have good foundational motivation
  - Frequentists should use Neyman-Pearson paradigm and consider optimal unconditional power across alternatives
    - And conditional/predictive power is not a good indicator in loss of unconditional power
  - Bayesians should use posterior distributions for decisions
Evaluation: Marketable Results

- Probability of obtaining estimates of treatment effect with clinical or marketing appeal
  - Modified power curve
    - Unconditional
    - Conditional at each analysis
  - Predictive probabilities at each analysis

Sequential Monitoring

Adaptive Designs

Where am I going?

- There has been much recent interest in the ability to modify an RCT design in the middle of the trial
  - Many of the "adaptive designs" are strikingly ill-advised on scientific grounds as well as being statistically inefficient

Sequential Sampling Strategies

- Two broad categories of sequential sampling
  - Prespecified stopping guidelines
  - Adaptive procedures

Adaptive Sampling Plans

- At each interim analysis, possibly modify
  - Scientific and statistical hypotheses of interest
  - Statistical criteria for credible evidence
  - Maximal statistical information
  - Randomization ratios
  - Schedule of analyses
  - Conditions for early stopping
Adaptive Sampling: Examples

- Prespecified on the scale of statistical information
  - E.g., Modify sample size to account for estimated information (variance or baseline rates)
- No effect on type I error IF
  - Estimated information independent of estimate of treatment effect
    » Proportional hazards
    » Normal data, and/or
    » Carefully phrased alternatives
  - And willing to use conditional inference
    » Carefully phrased alternatives

Estimate Alternative

- If maximal sample size is maintained, the study discriminates between null hypothesis and an alternative measured in units of statistical information

\[
n = \frac{\delta_1^2 V}{(\Delta_1 - \Delta_0)^2}
\]

Estimate Sample Size

- If statistical power is maintained, the study sample size is measured in units of statistical information

\[
n = \frac{\delta_1^2 V}{(\Delta_1 - \Delta_0)^2}
\]

Adaptive Sampling: Examples

- E.g., Proschan & Hunsberger (1995)
  - Modify ultimate sample size based on conditional power
    - Computed under current best estimate (if high enough)
  - Make adjustment to inference to maintain Type I error
Incremental Statistics

- Statistic at the $j$-th analysis a weighted average of data accrued between analyses

\[ \hat{\theta}_j = \frac{\sum_{k=1}^{j} N_k^* \hat{\theta}_k^*}{N_j}, \quad Z_j = \frac{\sum_{k=1}^{j} \sqrt{N_k^*} Z_k^*}{\sqrt{N_j}}. \]

Conditional Distribution

- \[ \hat{\theta}_j \mid N_j^* \sim N\left(\theta, \frac{V}{N_j^*}\right) \]
- \[ Z_j^* \mid N_j^* \sim N\left(\frac{\theta - \theta_0}{\sqrt{V/N_j^*}}, 1\right) \]
- \[ H_0 \]
- \[ P_j^* \mid N_j^* \sim U(0, 1). \]

Unconditional Distribution

- A mixture of normals, rather than a normal distribution

\[ \Pr(Z_j^* \leq z) = \sum_{n=0}^{\infty} \Pr(Z_j^* \leq z \mid N_j^*) \Pr(N_j^* = n). \]

Two Stage Design

- Proschan & Hunsberger consider worst case
  - At first stage, choose sample size of second stage
    - \[ N_2 = N_2(Z_1) \] to maximize type I error
  - At second stage, reject if \[ Z_2 > a_2 \]
- Worst case type I error of two stage design
  \[ \alpha_{\text{worse}} = 1 - \Phi\left(a_2^{(Z_1)}\right) + \frac{\exp\left(-\left(a_2^{(Z_1)}\right)^2/2\right)}{4}, \]
- Can be more than two times the nominal
  - \[ a_2 = 1.96 \] gives type I error of 0.0616
  - (Compare to Bonferroni results)
Better Approaches

Proschan and Hunsberger describe adaptations using restricted procedures to maintain experimentwise type I error.
- Must prespecify a conditional error function which would maintain type I error
  - Then find appropriate $a_2$ for second stage based on $N_2$ which can be chosen arbitrarily
  - But still have loss of power

Other Approaches

Bauer and Kohne:
- Use R.A. Fisher’s method for combining independent P values

L. Fisher:
- Variance spending function using prespecified weights at each stage

Muller and Schafer:
- Maintain conditional power function from some prespecified fixed sample test

Disadvantage Common to All

- Nonintuitive weighting of information from the different stages
  - Stages are not necessarily assigned weights proportional to the statistical information (sample size)

- Violation of the sufficiency principle
  - Inference depends on more information than is available in the minimal sufficient statistic

Motivation for Adaptive Designs

- Scientific and statistical hypotheses of interest
  - Modify target population, intervention, measurement of outcome, alternative hypotheses of interest
  - Possible justification
    - Changing conditions in medical environment
      - Approval/withdrawal of competing/ancillary treatments
      - Diagnostic procedures
    - New knowledge from other trials about similar treatments
    - Evidence from ongoing trial
      - Toxicity profile (therapeutic index)
      - Subgroup effects
Motivation for Adaptive Designs

- Modification of other design parameters may have great impact on the hypotheses considered
  - Statistical criteria for credible evidence
  - Maximal statistical information
  - Randomization ratios
  - Schedule of analyses
  - Conditions for early stopping

Cost of Planning Not to Plan

- Major issues with use of adaptive designs
  - What do we truly gain?
    - Can proper evaluation of trial designs obviate need?
  - What can we lose?
    - Efficiency? (and how should it be measured?)
    - Scientific inference?
      - Science vs Statistics vs Game theory
      - Definition of scientific/statistical hypotheses
      - Quantifying precision of inference

Prespecified Modification Rules

- Adaptive sampling plans exact a price in statistical efficiency
  - Tsiatis & Mehta (2002)
    - A classic prespecified group sequential stopping rule can be found that is more efficient than a given adaptive design
  - Shi & Emerson (2003)
    - Fisher's test statistic in the self-designing trial provides markedly less precise inference than that based on the MLE
      - To compute the sampling distribution of the latter, the sampling plan must be known

Conditional/Predictive Power

- Additional issues with maintaining conditional or predictive power
  - Modification of sample size may allow precise knowledge of interim treatment effect
    - Interim estimates may cause change in study population
      - Time trends due to investigators gaining or losing enthusiasm
    - In extreme cases, potential for unblinding of individual patients
      - Effect of outliers on test statistics
Final Comments

- Adaptive designs versus prespecified stopping rules
  - Adaptive designs come at a price of efficiency and (sometimes) scientific interpretation
  - With adequate tools for careful evaluation of designs, there is little need for adaptive designs

Sequential Monitoring

Documentation of Designs

Where am I going?

- Prespecification of the RCT design, monitoring plan, and analysis plan is of utmost importance

Specify Stopping Rule

- Null, design alternative hypotheses
- One-sided, two-sided hypotheses
- Type I error, Power to detect design alternative
- For each boundary
  - Hypothesis rejected
  - Error
  - Boundary scale
  - Boundary shape function parameters
- Constraints (minimum, maximum, exact)

Documentation of Rule

- Specification of stopping rule
- Estimation of sample size requirements
- Example of stopping boundaries under estimated schedule of analyses
  - sample mean scale, others?
- Inference at the boundaries
- Power under specific alternatives
- Behavior under possible scenarios
  - Alternative baseline rates, variability
### Implementation

- Method for determining analysis times
- Operating characteristics to be maintained
  - Power (up to some maximum $N$?)
  - Maximal sample size
- Method for measuring study time
- Boundary scale for making decisions
- Boundary scale for constraining boundaries at previously conducted analyses
- (Conditions stopping rule might be modified)

### Analysis Plan

- Stopping rule for inference
  - Nonbinding futility?
- Method for determining $P$ values
- Method for point estimation
- Method for confidence intervals
- Handling additional data that accrues after decision to stop