Session 3

Issues in Implementing Stopping Rules
Schedule of Analyses
Estimation of Statistical Information

Constraining Boundaries at Prior Analyses
Flexible Determination of Boundaries
Boundary scales
Measuring study time

Monitoring Secondary Endpoints

Issues in Implementing Stopping Rules: Schedule of Analyses

Schedule of Analyses

Design of clinical trial

Selection of stopping rule to provide desired operating characteristics
  - Type I error
  - Statistical power to detect design alternative
  - Efficiency
  - Bayesian properties
  - Futility considerations

Schedule of Analyses

At time of study design

Sample size (power, alternative) calculations based on
  - Specifying a maximum of J analyses
  - Specifying sample sizes at which analyses will be performed

Schedule of Analyses

During conduct of study

   Timing of analyses may be different
   - Monitoring scheduled by calendar time
   - Slow (or fast) accrual
   - Estimation of available information at time of locking database
   - External causes
   - (should not be influenced by study results)

Schedule of Analyses

Example: Stopping rule chosen at design
   Test of normal mean:
   - Null: $\mu \leq 0.0$
   - Alternative: $\mu \geq 0.5$

   One-sided symmetric test
   - Size .025, Power .975
   - Four equally spaced analyses
   - Pocock (1977) boundary relationships

Schedule of Analyses

Example: Stopping rule chosen at design (cont.)
```
One-sided test of a greater alternative:
Null : Theta <= 0      (size  = 0.025)
Alt  : Theta >= 0.5    (power = 0.975)
(Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                        a       d
   Time 1 (N=  86.31) 0.0000 0.5000
   Time 2 (N= 172.62) 0.1464 0.3536
   Time 3 (N= 258.92) 0.2113 0.2887
   Time 4 (N= 345.23) 0.2500 0.2500
```

Schedule of Analyses

Example: Analyses after 40%, 60%, 80%, 100% (maintain power)
```
Null: Theta <= 0      (size  = 0.025)
Alt : Theta >= 0.5    (power = 0.975)
(Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                        a       d
   Time 1 (N= 131.97) 0.1047 0.3953
   Time 2 (N= 197.95) 0.1773 0.3227
   Time 3 (N= 263.93) 0.2205 0.2795
   Time 4 (N= 329.91) 0.2500 0.2500
```

Schedule of Analyses

**Example: Analyses after 40%, 60%, 80%, 100%
(maintain maximal sample size)**

```
Null: Theta <= 0        (size  = 0.025)
Alt : Theta >= 0.4888   (power = 0.975)
(Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                           a       d
   Time 1 (N= 138.09) 0.1024 0.3864
   Time 2 (N= 207.14) 0.1733 0.3155
   Time 3 (N= 276.19) 0.2155 0.2732
   Time 4 (N= 345.23) 0.2444 0.2444
```

Schedule of Analyses

During conduct of study

Number of analyses may be different
- Monitoring scheduled by calendar time
- Slow (or fast) accrual
- External causes
- (should not be influenced by study results)

Schedule of Analyses

Example: Stopping rule chosen at design (cont.)

```
One-sided test of a greater alternative:
Null : Theta <= 0    (size  = 0.025)
Alt  : Theta >= 0.5  (power = 0.975)
(Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                         a       d
   Time 1 (N=  86.31) 0.0000 0.5000
   Time 2 (N= 172.62) 0.1464 0.3536
   Time 3 (N= 258.92) 0.2113 0.2887
   Time 4 (N= 345.23) 0.2500 0.2500
```

Schedule of Analyses

**Example: Analyses after 20%, 40%, 60%, 80%, 100%
(maintain power)**

```
Null: Theta <= 0    (size  = 0.025)
Alt : Theta >= 0.5  (power = 0.975)
(Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                          a       d
   Time 1 (N=  72.10) -0.0590 0.5590
   Time 2 (N= 144.20)  0.1047 0.3953
   Time 3 (N= 216.31)  0.1773 0.3227
   Time 4 (N= 288.41)  0.2205 0.2795
   Time 5 (N= 360.51)  0.2500 0.2500
```

Schedule of Analyses

Example: Analyses after 20%, 40%, 60%, 80%, 100% (maintain maximal sample size)

```
Null: Theta <= 0          (size  = 0.025)
Alt : Theta >= 0.5109     (power = 0.975)
(Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                               a        d
   Time 1 (N=  69.05) -0.0603  0.5713
   Time 2 (N= 138.09)  0.1070  0.4039
   Time 3 (N= 207.14)  0.1811  0.3298
   Time 4 (N= 276.19)  0.2253  0.2856
   Time 5 (N= 345.23)  0.2555  0.2555
```

Schedule of Analyses

Summary for Pocock boundary relationships

			Final
Analysis Times	Alt	Max N	Bound
========================	====	======	=====
.25, .50, .75, 1.00	.500	345.23	.2500
.40, .60, .80, 1.00	.500	329.91	.2500
.40, .60, .80, 1.00	.489	345.23	.2444
.20, .40, .60, .80, 1.00	.500	360.51	.2500
.20, .40, .60, .80, 1.00	.511	345.23	.2555

Schedule of Analyses

Summary for O'Brien-Fleming boundary relationships

			Final
Analysis Times	Alt	Max N	Bound
========================	====	======	=====
.25, .50, .75, 1.00	.500	256.83	.2500
.40, .60, .80, 1.00	.500	259.44	.2500
.40, .60, .80, 1.00	.503	256.83	.2513
.20, .40, .60, .80, 1.00	.500	259.45	.2500
.20, .40, .60, .80, 1.00	.503	256.83	.2513

Schedule of Analyses

Need methods that allow flexibility in determining number and timing of analyses

Should maintain some (but not, in general, all) desired operating characteristics, e.g.:
- Type I error
- Type II error
- Maximal sample size
- Futility properties
- Bayesian properties

February, 2003

Schedule of Analyses

Validity of flexible determination of analysis times
> Inference conditional on actual schedule of analyses

> Can disregard rule for scheduling analyses if it is independent of measures of treatment effect

> If all possible adaptations maintain particular operating characteristics, then so will adaptive rule

> (May affect other operating characteristics of design)

Issues in Implementing Stopping Rules: Estimation of Statistical Information

---

Estimation of Statistical Information

At time of study design

> Sample size (power, alternative) calculations based on
> - Specifying statistical information available from each sampling unit

---

Estimation of Statistical Information

During conduct of study

> Statistical information from a sampling unit may be different than originally estimated
> - Variance of measurements
> - Baseline event rates
> - (Altered sampling distribution for treatment levels)

Estimation of Statistical Information

Sample size formulas used in group sequential test design

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

- n is the maximal number of sampling units
- δ_1 is the alternative for which a standardized form of a level α test has power β
- 1 / V is the statistical information contributed by each sampling unit

Estimation of Statistical Information

Parallels with fixed sample test design

Sample size formulas used in group sequential test design are completely analogous to those used in fixed sample studies

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

In fixed sample two arm tests of a normal mean

$$\delta_1 = z_{1-\alpha/2} + z_\beta \qquad V = 2\sigma^2$$

Estimation of Statistical Information

Effect of using incorrect estimates of statistical information at the design stage

Using the specified sample size, the design alternative will not be detected with the desired power

Using the specified sample size, the alternative detected with the desired power will not be the design alternative

In order to detect the design alternative with the desired power, a different sample size is needed

Estimation of Statistical Information

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

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

$$n = \frac{\delta_1^2}{\left(\dfrac{(\Delta_1 - \Delta_0)^2}{V} \right)}$$

Estimation of Statistical Information

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

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

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

Estimation of Statistical Information

Validity of flexible determination of sample size
   Inference conditional on actual sample size

   Can disregard rule for determining sample size if it is independent of measures of treatment effect

   If all possible adaptations maintain particular operating characteristics, then so will adaptive rule

   (May affect other operating characteristics of design)

Flexible Determination of Boundaries

Flexible Determination of Boundaries

Previously described methods for implementing stopping rules
   - (Adhere exactly to monitoring plan)
   - (Approximations based on design parameters: Emerson and Fleming, 1989)
   - Christmas tree approximation for triangular tests: Whitehead and Stratton, 1983
   - Error spending functions: Lan and DeMets, 1983; Pampallona, Tsiatis, and Kim, 1995
   - Constrained boundaries in unified design family: Emerson, 2000

Flexible Determination of Boundaries

Common features

Stopping rule specified at design parameterizes the boundary for some statistic (boundary scale)

At the first interim analysis, parametric form is used to compute the boundary for actual time on study

At successive analyses, the boundaries are recomputed accounting for the exact boundaries used at previously conducted analyses

Maximal sample size estimates may be updated

Flexible Determination of Boundaries

Specification of implementation strategy

Boundary scale used to modify boundaries

How analysis times will be determined (maintain blind)

How study time will be measured

Operating characteristics which will be maintained

Boundary Scales

Boundary Scales

Families of group sequential stopping rules can be defined on a number of scales

Parametric family relates stopping boundaries at successive analyses

- Π_j = proportion of maximal information available at j-th analysis
- d_j = stopping boundary at j-th analysis for some statistic
- $d_j = f(\Pi_j)$ is parametric boundary function

February, 2003

Boundary Scales

Unified family of group sequential designs (Kittelson and Emerson, 1999)

Defined for estimate of treatment effect (sample mean scale)

Includes Pocock (1977), O'Brien and Fleming (1979), Whitehead and Stratton (1983), Wang and Tsiatis (1987), Emerson and Fleming (1989), Pampallona and Tsiatis (1994), Xiong (1995)

Boundary Scales

Error spending family (Kim and DeMets, 1987; Jennison and Turnbull, 1989)

Power family for error spending function

Pampallona, Tsiatis, and Kim (1995) describe a family by interpolating the error spending function for tests defined on the sample mean scale

Boundary Scales

Extensions to those parametric families in S+SeqTrial: Constrained boundaries

Motivation: Extreme conservatism of the O'Brien-Fleming design

Specify a design that has stopping boundaries that are the least extreme of an O'Brien-Fleming boundary relationship or a fixed sample P value of .001

Boundary Scales

Example: O'Brien-Fleming boundaries on fixed sample P value scale

```
Null: Theta <= 0      (size  = 0.025)
Alt : Theta >= 0.5    (power = 0.975)
(Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Fixed Sample P-value scale
                        a       d
   Time 1 (N=  64.21) 0.9774 0.0000
   Time 2 (N= 128.41) 0.5000 0.0023
   Time 3 (N= 192.62) 0.1237 0.0104
   Time 4 (N= 256.83) 0.0226 0.0226
```

Boundary Scales

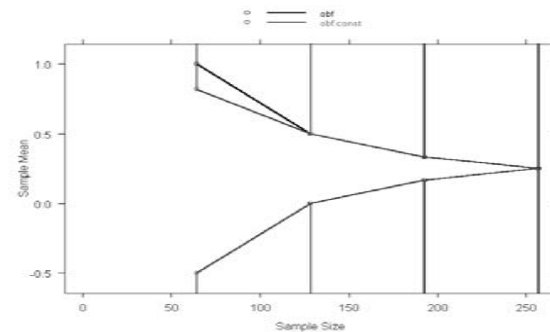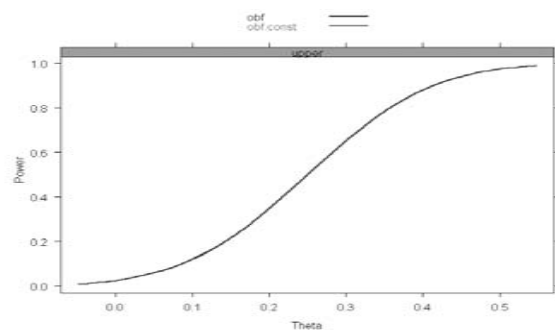**Example: Constrained O'Brien-Fleming boundaries on fixed sample P value scale**

```
Null: Theta <= 0      (size  = 0.025)
Alt : Theta >= 0.5    (power = 0.975)
(Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Fixed Sample P-value scale
                               a        d
    Time 1 (N=  64.31) 0.9773 0.0005
    Time 2 (N= 128.61) 0.4989 0.0023
    Time 3 (N= 192.92) 0.1231 0.0102
    Time 4 (N= 257.23) 0.0224 0.0224
```

Schedule of Analyses

Example: Display of boundaries

Schedule of Analyses

Example: Display of power curves

Schedule of Analyses

Example: Display of ASN curves

Boundary Scales

Constrained boundaries also defined for error spending family

Allows arbitrary departures from the parametric families

Boundary Scales

Use of constrained families in flexible implementation of stopping rules

At the first analysis, compute stopping boundary from parametric family

At successive analyses, use parametric family with constraints (on some scale) for the previously conducted interim analyses

When the error spending scale is used, this is just the error spending approach of Lan & DeMets or Pampallona, Tsiatis, & Kim

Schedule of Analyses

Example: Stopping rule chosen at first analysis (with estimates for later analyses)

```
One-sided test of a greater alternative:
Null : Theta <= 0      (size  = 0.025)
Alt  : Theta >= 0.5    (power = 0.975)
(Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                           a       d
   Time 1 (N=  77.08) -0.0743 0.5744
   Time 2 (N= 154.16)  0.1211 0.3789
   Time 3 (N= 231.23)  0.2029 0.2971
   Time 4 (N= 308.31)  0.2500 0.2500
```

Schedule of Analyses

Example: Stopping rule based on updated schedule (when first analysis boundary unconstrained)

```
One-sided test of a greater alternative:
Null : Theta <= 0      (size  = 0.025)
Alt  : Theta >= 0.4973 (power = 0.975)
(Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                           a       d
   Time 1 (N=  77.08)  -0.0740 0.5713
   Time 2 (N= 100.00)   0.0087 0.4887
   Time 3 (N= 231.23)   0.2018 0.2955
   Time 4 (N= 308.31)   0.2487 0.2487
```

Schedule of Analyses

Use of constrained families is necessary because critical values are dependent upon exact schedule

In Unified Family, boundary at first analysis is affected by timing of later analyses

Compare boundary at first analysis when timing of second analysis differs:
- `a' boundary: -0.0743 versus -0.0740
- 'd' boundary: 0.5744 versus 0.5713

Must constrain first boundaries at the levels actually used, and then use parametric form for future analyses

---

Measuring Study Time

Measuring Study Time

Flexible methods compute boundaries at an interim analysis according to study time at that analysis

Study time can be measured by
- Proportion of planned number of subjects accrued (maintains maximal sample size)
- Proportion of planned statistical information accrued (maintains statistical power)
- (Calendar time-- not really advised)

Measuring Study Time

In either case, we must decide how we will deal with estimates of statistical information at each analysis when constraining boundaries

Statistical information in clinical trials typically has two parts
- V = variability associated with a single sampling unit
- The distribution of sampled levels of treatment

In many clinical trials, the dependence on the distribution of treatment levels across analyses is only on the sample size N

Measuring Study Time

Possible approaches

At each analysis estimate the statistical information available, and use that estimate at all future analyses

- Theoretically, this can result in estimates of negative information gained between analyses

At each analysis use the sample size with the current best estimate of V

- The 1:1 correspondence between boundary scales is thus broken at previously conducted analyses

Measuring Study Time

Possible approaches (cont.)

In S+SeqTrial, all probability models have statistical information directly proportional to sample size for block randomized experiments, thus we chose to update V at all analyses using the current best estimate

Other statistical packages (PEST, ?EaSt) constrain boundaries using the estimate of statistical information available at the previous analyses.

There is no clear best approach

Measuring Study Time

Example

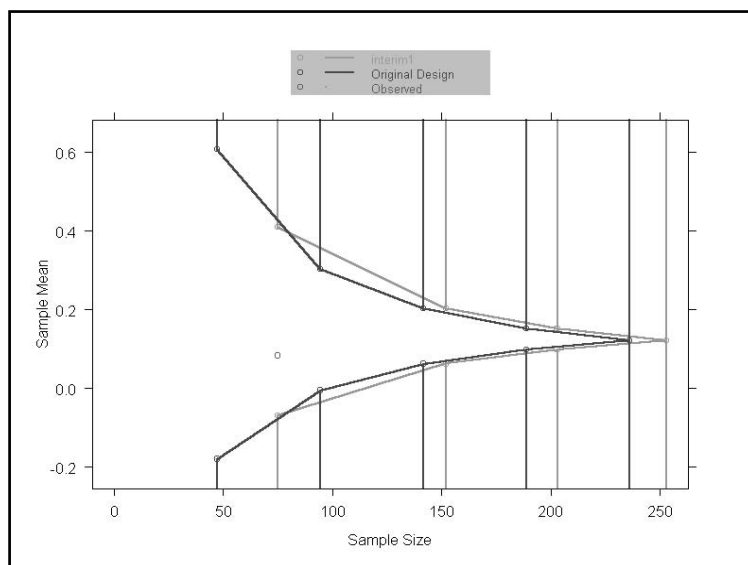A clinical trial of a binary endpoint is designed using a unified family design

- One-sided test for an increased event probability
- Designed with 5 analyses
- O'Brien-Fleming efficacy boundary
- Futility boundary intermediate to O'Brien-Fleming and Pocock

Measuring Study Time

Example: At first interim analysis using unified family approach

The use of the parametric form for the boundary function will result in a boundary on the same curve as the original design

I had the sample size re-estimated to allow for errors in guessing the baseline rate at the design phase
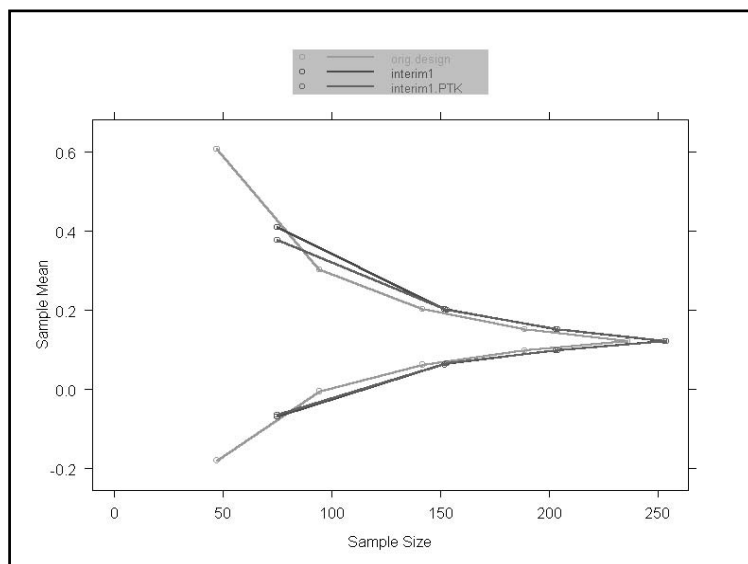
Measuring Study Time

Example: Comparison with error spending approach using interpolated error spending function

The stopping boundary based on the error spending function will not agree exactly with the curve for the original design, because the error spending function is not linear for this design.

Had the monitoring occurred at the prespecified time, the two curves would agree.
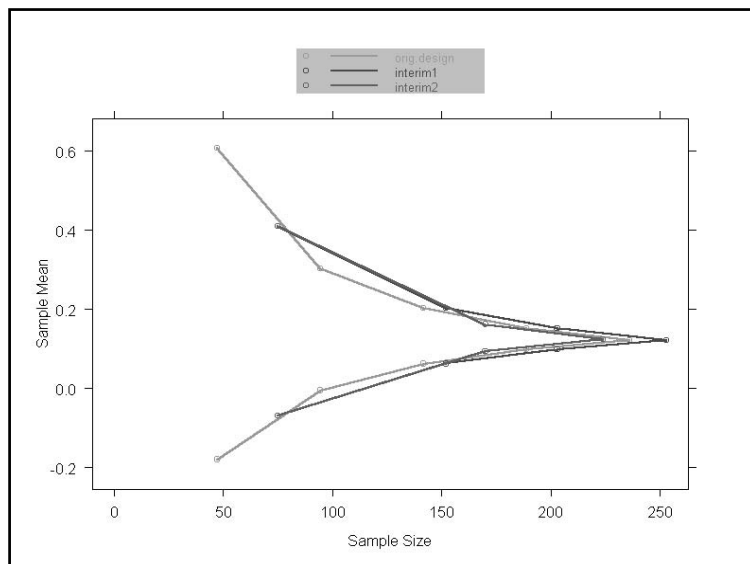
Measuring Study Time

Example: Superposed stopping rules from first and second interim analyses using unified family

Because the monitoring bounds were constrained on the sample mean scale, the stopping boundaries computed for the first analysis agree at both analyses when plotted on the sample mean scale

If the boundaries were plotted on some other scale, they would not agree
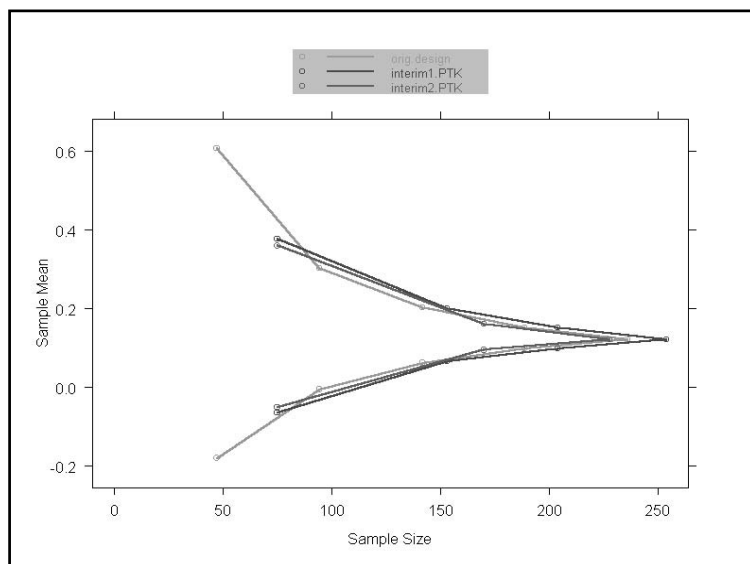
Measuring Study Time

Example: Superposed stopping rules from first and second interim analyses using error spending

When plotted on the sample mean scale, the monitoring bounds from the first and second analyses will not agree if the boundary at the first analysis is constrained on the error spending scale

This is due to the need to estimate the statistical information

Final Comments

I think it makes more sense to use the best estimate of the variance of an observation when estimating a sampling distribution. This avoids the possibility of negative information, but allows the conflicting results described above.

In the absence of a need to estimate the statistical information, monitoring on the sample mean or error spending scales would agree exactly (modulo interpolation to obtain the error spending function).

Final Comments (cont.)

When estimating the statistical information, all approaches merely approximate the sampling distribution of the test statistic. At this point there is no clear "best" approach

On purely esthetic grounds, I prefer that the monitoring bounds match across analyses on the sample mean scale

---

Monitoring Secondary Endpoints

Monitoring Secondary Endpoints

So far, we have stressed the monitoring of the primary endpoint

Of course, far more time in a DSMB meeting is devoted to monitoring the secondary endpoints related to patient safety than is devoted to examining the primary endpoint

---

Monitoring Secondary Endpoints

Role of DSMB: Maintain validity of informed consent

Evaluate the safety of the trial in light of information made available since the start of study
- Data from current trial
- Data from related trials
- Changing clinical environment

February, 2003

Monitoring Secondary Endpoints

Safety issues to be addressed

Is there evidence that individual patients might be being harmed?

- Serious adverse experiences
- Individual abnormal lab values

Is there evidence of trends toward harm in the population of treated patients

- Proportion with adverse experiences
- Average (median) lab values

Monitoring Secondary Endpoints

Statistical issues due to rare events

Invariably, there is very little statistical precision to establish increased rates of Serious Adverse Experiences (SAEs) or increased rates of individual toxicities

- As a general rule, the DSMB therefore must act based on their prior knowledge and principles of conservatism
  - E.g., decisions to modify entry criteria by age due to statistically nonsignificant trends in the data

Monitoring Secondary Endpoints

Statistical issues due to rare events (cont.)

The increased error rate of acting on such trends is a necessary evil

- Of some solace is the fact that most new treatments do not prove beneficial, so such conservatism is probably not <u>too</u> harmful in the quest for new treatments
- In essence, we decide to only look at the most safe treatments (and the trials that tended to result in the safest profile)

Monitoring Secondary Endpoints

Statistical issues when using aggregate statistics to examine the safety profile

When examining the safety profile statistically, must consider multiple comparison problems

- over multiple adverse experience categories
  - (the DSMB is largely on their own)
- over multiple analyses of the accruing data
  - (group sequential methods can be used as a guideline)

Monitoring Secondary Endpoints

Statistical issues when using aggregate statistics to examine the safety profile (cont.)

Group sequential methods for monitoring safety profiles

- Bayesian approaches
    - But how do you ever detect unexpected toxicities-- where is the burden of proof?
- Frequentist approaches
    - Using group sequential stopping rules to compute
        - Repeated confidence intervals
        - Ersatz P values

Monitoring Secondary Endpoints

Presentation of results to the DSMB

Generally avoid providing any P values or RCI for specific analyses to avoid their difficult interpretation

- Have to account for multiple comparisons across endpoints
- Have to consider tradeoffs between efficacy and toxicity
- Statistical significance may be secondary to safety concerns-- may need to act before statistical significance is attained

Monitoring Secondary Endpoints

Presentation of results to the DSMB (cont.)

If an issue arises where stopping a trial for safety reasons is potentially indicated, it is useful to have some sort of guideline available for reference

Monitoring Secondary Endpoints

Selection of stopping rules for use with safety endpoints

Need to consider whether harm should be proven
- existing treatments

I think that the general philosophy of clinical testing dictates that such a stopping rule should not be as conservative as those typically used for efficacy endpoints

- An O'Brien-Fleming guideline is probably too conservative for safety