Session 4

Analyses Adjusted for Stopping Rules
- Reporting Results
- Adjustment for Stopping Rules

Choice of Inferential Methods
- Methods of Point Estimation
- Orderings of the Outcome Space
- Relative Advantages

Sensitivity to Poorly Specified Stopping Rules
- Case Study I: Absence of Formal Stopping Rule
- Case Study II: Unexpected Toxicities

Documentation of Design, Monitoring and Analysis

Reporting Results

Reporting Results

At the end of the study analyze the data to provide

Estimate of the treatment effect
- Single best estimate
- Range of reasonable estimates

Decision of efficacy, equivalence, harm, or futility
- Binary decision
- Quantification of strength of evidence

Reporting Results

Methods of point estimation

Frequentist methods
- Find estimates which minimize bias
- Find estimates with minimal variance
- Find estimates which minimize mean squared error

Bayesian methods
- Use mean, median, or mode of posterior distribution of θ based on some prespecified prior

Reporting Results

Methods of point estimation (cont.)

Method of moments

- Use a function of sample moments to estimate a function of moments of the sampling distribution
- For example
  - if θ is the mean of the sampling distribution, use sample mean as an estimate of θ
  - if θ is the variance of the sampling distribution, use sample variance as an estimate of θ

Reporting Results

Methods of point estimation (cont.)

Maximum likelihood estimation

- Find the value of θ such that the sampling density evaluated at the observed data is maximized
- E.g., in one sample inference about a normal mean maximize density when θ equals the sample mean

$$\prod_{i=1}^{n} \frac{1}{\sigma} \varphi\left(\frac{X_i - \vartheta}{\sigma} \right)$$

Reporting Results

Methods of point estimation (cont.)

Median unbiased estimation

- Assume that the observed statistic is the median of its sampling distribution
- E.g., if observed T=t, then find θ such that

$$\Pr\left(T \le t \mid \vartheta\right) \ge 0.5$$
$$\Pr\left(T \ge t \mid \vartheta\right) \ge 0.5$$

Reporting Results

Methods of point estimation (cont.)

Bias adjusted

- Assume that the observed statistic is the mean of its sampling distribution
- E.g., if observed T=t, then find θ such that

$$E\left(T \mid \vartheta\right) = t$$

Reporting Results

Methods of point estimation (cont.)

Variance improvement for unbiased estimators

- Use Rao-Blackwell improvement theorem to find expectation of unbiased estimate conditioned on sufficient statistic
- E.g., for S unbiased and T sufficient

$$E(S \mid T = t)$$

Reporting Results

Methods of interval estimation

Confidence interval

- 100(1-α)% confidence interval for θ is (θ_L , θ_U) where

$$\Pr(T \le t \mid \vartheta_L) = 1 - \frac{\alpha}{2}$$

$$\Pr(T \le t \mid \vartheta_U) = \frac{\alpha}{2}$$

Reporting Results

Methods of interval estimation (cont.)

Bayesian methods

- Use central 100(1-α)% of posterior distribution of θ based on some prespecified prior

Reporting Results

Criteria for decisions

Hypothesis tests

- Reject hypothesis that $\theta = \theta_0$ with a level α test if $T > c_\alpha$ where

$$\Pr(T \ge c_\alpha \mid \vartheta_0) = \alpha$$

Reporting Results

Criteria for decisions (cont.)

Bayesian Methods

- Reject hypothesis that $\theta = \theta_0$ based on posterior distribution, e.g.,

$$\Pr\left(\vartheta \geq \vartheta_0 \mid \vec{X}\right) \geq \beta$$

Reporting Results

Quantification of Evidence for Decisions

Hypothesis testing

- P value $\Pr\left(T \geq t \mid \vartheta_0\right)$

Bayesian Methods

- Posterior probability $\Pr\left(\vartheta \geq \vartheta_0 \mid \vec{X}\right)$

Adjustment for Stopping Rules

Adjustment for Stopping Rules

Fixed sample methods for testing and estimation are well developed

Many methods of point estimation yield same estimate (including Bayesian with noninformative prior)

Confidence intervals easily computed

Testing well developed

Adjustment for Stopping Rules

Stopping rule greatly affects sampling distribution for estimates of treatment effect

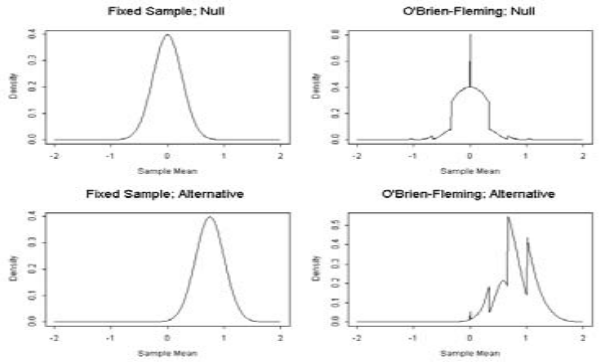
Data which lead to normally distributed sampling distributions under fixed sample testing lead to skewed, multimodal densities with jump discontinuities under sequential testing

Treatment effect is no longer a shift parameter

Exact shape of sampling distribution therefore depends upon stopping rule and alternative

Adjustment for Stopping Rules

Sampling Densities for Estimate of Treatment Effect



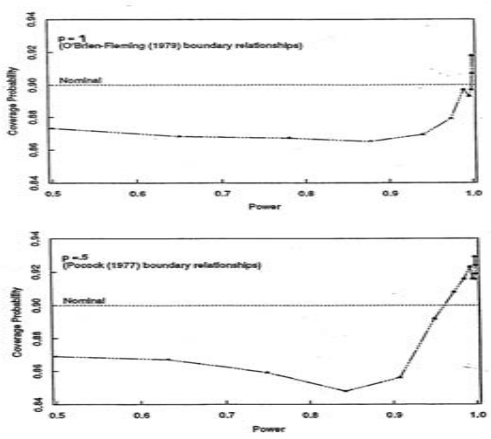
Adjustment for Stopping Rules

Failure to adjust estimates and P values for stopping rule is tantamount to repeated significance testing

- P values will tend to be wrong
- Estimates will tend to be biased toward extreme
- Confidence intervals will have the wrong coverage probabilities

(No effect on Bayesian analysis)

Coverage probability of unadjusted CI



Adjustment for Stopping Rules

Frequentist inferential techniques can still be used, providing we can compute the sampling density for the test statistic under arbitrary choices for θ

In these techniques, the stopping rule is just viewed as a sampling distribution

cf: binomial versus geometric sampling

Adjustment for Stopping Rules

P values adjusted for stopping rule

Probability of observing more extreme results under the null hypothesis

Compute sampling distribution of test statistic under the null

Requires a definition of "extreme" across analysis times

- Ordering of the outcome space

Adjustment for Stopping Rules

Point estimates adjusted for stopping rule

Maximum likelihood estimate is unadjusted estimate

- Generally biased
- Tends to have large mean squared error

Find estimates that decrease the bias and mean squared error

Adjustment for Stopping Rules

Confidence interval adjusted for stopping rule

Based on duality of testing and CI

Exact coverage probability under normal probability model

Requires definition of an ordering of the outcome space

Methods of Point Estimation

Methods of Point Estimation

Point estimates adjusted for stopping rule

Maximum likelihood estimate is unadjusted estimate

- Generally biased
- Large mean squared error

Methods of Point Estimation

Point estimates adjusted for stopping rule

Bias adjusted mean (Whitehead, 1986)

- Assume observed outcome is mean of true distribution
- Requires knowing number and timing of future analyses
- Generally still biased
- Often least mean squared error

Methods of Point Estimation

Point estimates adjusted for stopping rule

Median unbiased estimate (Whitehead, 1984)

- Assume observed outcome is median of true distribution
- Requires an ordering of the outcome space
- Some orderings require knowledge of number and timing of future analyses
- Generally still biased for mean

Methods of Point Estimation

Point estimates adjusted for stopping rule

UMVUE-like estimate

- Uses Rao-Blackwell improvement theorem
- Unbiased for normal probability model
- Does not require knowledge of number and timing of future analyses

Orderings of the Outcome Space

Orderings of the Outcome Space

Ordering of the outcome space

Orderings of outcomes within an analysis time intuitive

- Based on the value of T_j at that analysis

Need to define ordering between outcomes at successive analyses

- How does sample mean of 3.5 at second analysis compare to sample mean of 3 at first analysis (when estimate more variable)?

Orderings of the Outcome Space

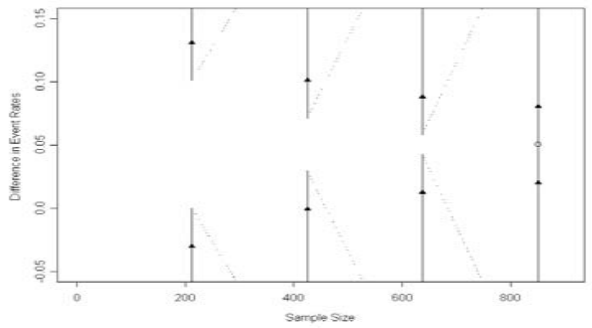
Analysis time ordering (Jennison and Turnbull, 1983; Tsiatis, Rosner, and Mehta, 1984)

Results leading to earlier stopping are more extreme

- Linearizes the outcome space
- Does not require knowledge of future analysis times
- Not defined for two-sided tests with early stopping for both null and alternative

Orderings of the Outcome Space

Analysis time ordering



Orderings of the Outcome Space

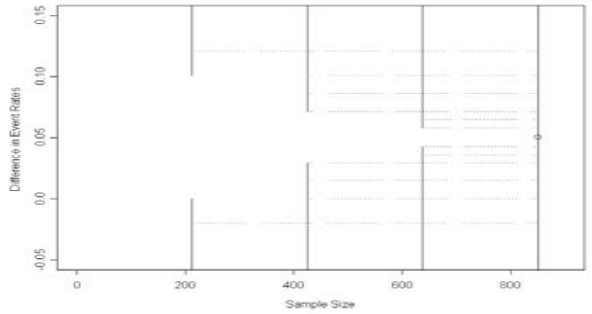
Sample mean ordering (Duffy and Santner, 1987; Emerson and Fleming, 1990)

Consider only magnitude of sample mean

- Requires knowledge of future analysis times
- Tends to result in narrower CI and less biased median unbiased estimates

Orderings of the Outcome Space

Sample mean ordering contours



Relative Advantages

Relative Advantages

Properties of methods for inference

Point estimates differ in bias reduction, mean squared error

Confidence intervals differ in

- average width of CI
- inclusion of various point estimates
- need for knowledge about future analyses

(ref: Emerson and Fleming, 1990)

Relative Advantages

Choice of methods for inference

Fixed sample tests

- All frequentist methods described here agree with each other

Group sequential tests

- No method is uniformly better
- Usually fairly good agreement between various methods
- Failure to agree can be informative regarding time trends in data

Relative Advantages

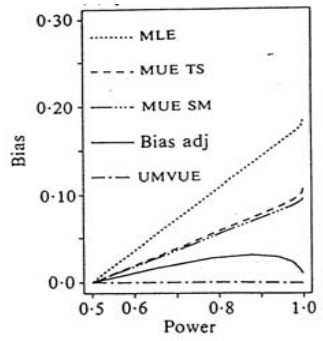
Point estimates: General tendencies for bias from least to most

(best)
UMVUE-like (in normal model)
Bias adjusted mean
Median unbiased with sample mean ordering
Median unbiased with analysis time ordering
Maximum likelihood estimate
(worst)

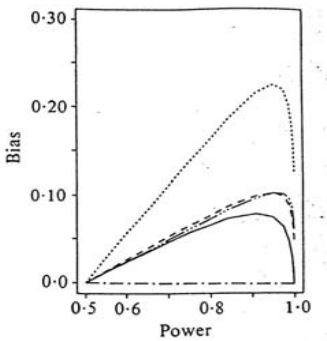
Relative Advantages

Point estimates: General tendencies for bias

O'Brien-Fleming



Pocock



Relative Advantages

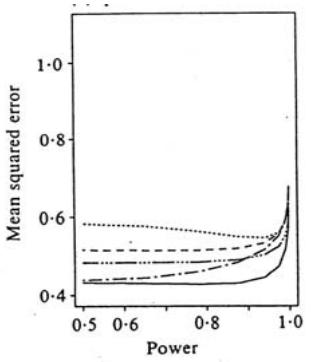
Point estimates: General tendencies for mean squared error (MSE) from least to most

(best)
Bias adjusted mean
Median unbiased with sample mean ordering
UMVUE-like
Median unbiased with analysis time ordering
Maximum likelihood estimate
(worst)

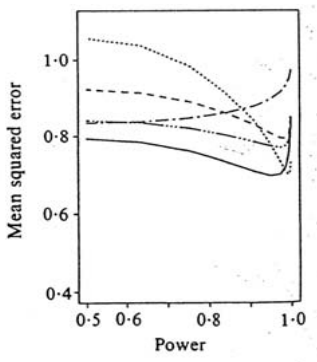
Relative Advantages

Point estimates: General tendencies for (MSE)

O'Brien-Fleming



Pocock



Relative Advantages

Point estimates: Dependence on timing of future analyses

(None)
UMVUE-like
Median unbiased with analysis time ordering
Maximum likelihood estimate

(Some)
Bias adjusted mean
Median unbiased with sample mean ordering

Relative Advantages

Point estimates: Spectrum of group sequential designs for which defined

(All)
Bias adjusted mean
Median unbiased with sample mean ordering
UMVUE-like
Maximum likelihood estimate

(Not two-sided tests with stopping under both hypotheses)
Median unbiased with analysis time ordering

Relative Advantages

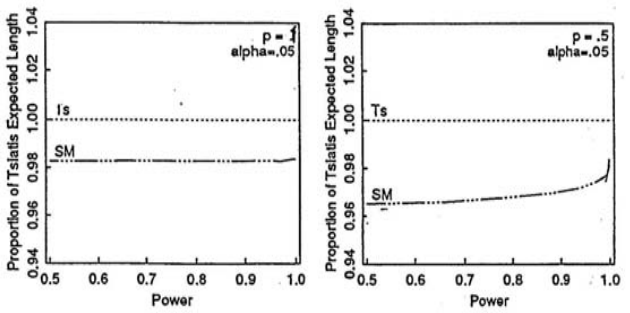
Interval estimates: General tendencies toward narrower confidence intervals

(Narrowest)
Sample mean ordering based
Analysis time ordering based
(Widest)

Relative Advantages

Interval estimates: Average length of confidence intervals

O'Brien-Fleming Pocock



Relative Advantages

Interval estimates: Dependence on timing of future analyses

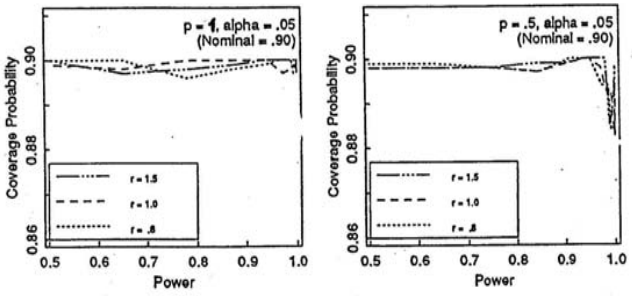
(None)
Analysis time ordering based

(Some)
Sample mean ordering based

Relative Advantages

Interval estimates: Coverage probability for CI using estimated schedule of analyses

O'Brien-Fleming Pocock



Relative Advantages

Interval estimates: Spectrum of group sequential designs for which defined

(All)
Sample mean ordering based

(Not two-sided tests with stopping under both hypotheses)
Analysis time ordering based

Relative Advantages

Interval estimates: Possible exclusion of point estimates

(Tends to occur with less than 0.5% probability)

Analysis time ordering might not include
- Bias adjusted mean
- Sample mean ordering based MUE
- Maximum likelihood estimate

Sample mean ordering might not include
- UMVUE-like
- Analysis time ordering based MUE

Relative Advantages

P values

Tend to agree for the sample mean and analysis time orderings for making typical decisions regarding statistical significance

Relative Advantages

P values: Spectrum of group sequential designs for which defined

(All)
Sample mean ordering based

(Not two-sided tests with stopping under both hypotheses)
Analysis time ordering based

Sensitivity to Poorly Specified Stopping Rules

Poorly Specified Stopping Rule Approach

Based on statistical inference

Consider class of stopping rules parameterized by
- level of significance
- boundary shape functions
- number and timing of analyses

Adjust estimates, P values for stopping rules

Evaluate sensitivity of conclusions to choice of stopping rules within that class

Poorly Specified Stopping Rule Approach

Determining class of stopping rules to consider

Consider interim results of study at potential analysis times that did not result in stopping
- True stopping rule must have been more extreme

Consider interim results of study at analysis times that did result in stopping
- True stopping rule must have been less extreme

Case Study 1

Case Study 1: Idarubicin in AML

Idarubicin in Acute Myelogenous Leukemia

Patients randomized to receive Idarubicin (Ida) or Daunorubicin (Dnr) in equal numbers

Primary response: Induction of complete remission

Secondary response: Survival

Case Study 1: Idarubicin in AML

Initial design

Fixed sample study

Two-sided level 0.05 hypothesis test

80% power to detect absolute difference in response rates of 0.20

90 patients per treatment arm

Case Study 1: Idarubicin in AML

Chronology

Several informal analyses of the data

Formal analysis of the data when N=45 per arm

- CR rate - Ida: 35/45 (78%); Dnr: 25/45 (56%)
- Retrospective adoption of O'Brien-Fleming design
- Trial continued

Formal analysis of the data when N=65 per arm

- CR rate - Ida: 51/65 (78%); Dnr: 38/65 (58%)
- Trial stopped

Case Study 1: Idarubicin in AML

FDA Questions

Was the O'Brien-Fleming design truly the one used?

- Number and timing of analyses
- Level of test
- Boundary shape function

(Can we trust retrospective imposition of the stopping rule?) (Case Study 2)

(Interpretation of secondary endpoint of survival?)

Case Study 1: Idarubicin in AML

Selection of Class of Stopping Rules for Sensitivity Analysis

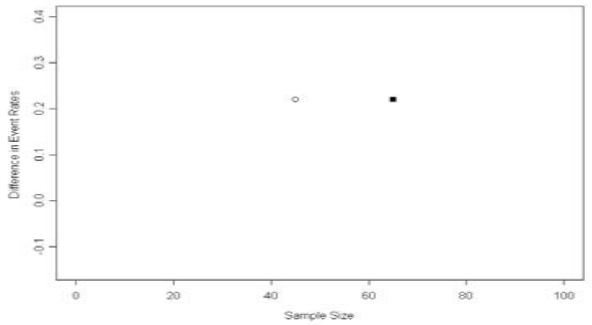
Study did not stop with treatment difference of 0.22 when N= 45 / arm

Study did stop with treatment difference of 0.20 when N= 65 / arm

Consider stopping boundaries that are between those two points

Case Study 1: Idarubicin in AML

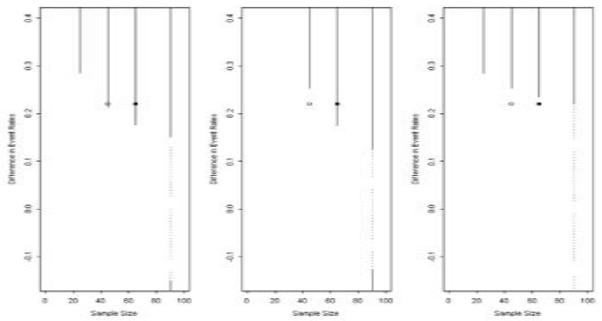
Observed results



Case Study 1: Idarubicin in AML

Stopping rules

Impossible Possible Impossible



Case Study 1: Idarubicin in AML

Parameterization

Number and timing of analyses

Boundary shape function

Level of significance

- Worst case: just barely continued at N= 45
- Best case: just barely stopped at N= 65

Case Study 1: Idarubicin in AML

Sensitivity analysis

(45, 65, 90)	Best Case	Worst Case
• Level	.958	.868
• P value	.008	.015
• Estimate	.184	.181
• 95% CI	(.034,.325)	(.018,.348)

Case Study 1: Idarubicin in AML

Sensitivity analysis

(25, 45, 65, 90)	Best Case	Worst Case
• Level	.958	.868
• P value	.008	.016
• Estimate	.184	.175
• 95% CI	(.034,.325)	(.017,.348)

Case Study 1: Idarubicin in AML

Sensitivity analysis

(12, 25, 35, 45, 65, 90)	Best Case	Worst Case
• Level	.958	.866
• P value	.008	.017
• Estimate	.182	.171
• 95% CI	(.034,.325)	(.015,.347)

Case Study 2

Case Study 2: Unexpected Toxicities

Background

Clinical trial of G-CSF to reduce a certain type of toxicity in cancer chemotherapy

Early in trial, high rates of another toxicity noted

Ed Korn at NCI consulted re early stopping

Much later, Ed Korn invites panel to address this problem as an unknown at the JSM

Clinical Setting

Clinical trial of Granulocyte Colony Stimulating Factor (G-CSF)

Oral mucositis toxicity with 5-FU/LV chemotherapy

Observation of decreased incidence when G-CSF was given for other indications

Hence clinical trial planned to address role in reducing oral mucositis

Clinical Setting

Clinical trial design

Fixed sample design

35 patients to receive G-CSF in first chemo cycle; nothing in second

Primary endpoint: difference in oral mucositis between cycles

Clinical Setting

Chronology

3 of 4 first patients experience life threatening leukopenia

A fifth patient currently under treatment

Question: When should we be concerned enough to stop the trial?

Clinical Setting

Biological issues

G-CSF stimulates division of leukocytes; chemotherapy kills rapidly dividing cells

Leukopenia was a secondary endpoint

Current trial included patients with prior chemotherapy unlike previous trials

- 2 of 3 toxicities were with prior chemotx
- 2 of 2 patients with prior chemotx had toxicities

Clinical Setting

Acceptable levels of toxicity

Only 12 / 176 (6.8%) of patients on 5-FU / LV in previous study experienced leukopenia

Clinical researcher: Maybe 50% toxicity rate would be acceptable

Approach to Problem

Outline

Selection of a group sequential stopping rule

Analysis of results

Sensitivity of analysis to data driven selection of stopping rule

Bayesian analysis

Selection of Stopping Rule

Selection of hypotheses

Power to detect toxicity rate greater than 50%

Null hypothesis: toxicity rate less than 20%

- arbitrary choice
- allows for prior chemotherapy

Selection of Stopping Rule

Schedule of analyses

First analysis at N = 5

Additional analyses every 5 patients to maximum of 35

Selection of Stopping Rule

Structure of stopping rule

Early stopping only for excess toxicity

Boundaries defined for number of toxicities

Consider boundary shape functions of

- O'Brien-Fleming
- Pocock

Binary Endpoint

Issues in small studies with binary endpoint

Size, power not attained exactly

Large sample approximations not appropriate

Implementation of boundary relationships approximate

- rounding vs truncation of boundaries

Sampling Density

Group Sequential Test Statistic

Observations	$X_i \sim B(1,\pi)$
Analysis times	$N_1, N_2, N_3, \ldots, N_J$
Continuation sets	(a_j, b_j)

Statistics $S_j = \sum_{i=1}^{N_j} X_i$

$$M = \min\left\{j : S_j \notin (a_j, b_j)\right\}$$

$$S = S_M$$

Sampling Density

After Armitage, McPherson, and Rowe (1969)

$$\Pr[M = m, S = s; \pi] = \begin{cases} f(m,s)\pi^s(1-\pi)^{N_j - s} & if\ s \notin (a_m, b_m) \\ 0 & else \end{cases}$$

$$f(j,s) = \begin{cases} \binom{N_1}{s} & j = 1 \\ \sum_{i=a_{j-1}+1}^{b_{j-1}-1} f(j-1,i)\binom{N_j - N_{j-1}}{s-i} & j = 2, \ldots, J \end{cases}$$

Candidate Designs

Threshold for rejecting the null hypothesis

		OBF	Poc
Boundaries	$N_1 = 5$	6	4
	$N_2 = 10$	7	6
	$N_3 = 15$	8	8
	$N_4 = 20$	9	10
	$N_5 = 25$	10	11
	$N_6 = 30$	11	13
	$N_7 = 35$	12	14

Candidate Designs

Operating characteristics

		OBF	Poc
Hypotheses	Null	.183	.209
	Alternative	.488	.542
ASN	$\pi = 0.2$	34.7	34.6
	$\pi = 0.5$	18.6	17.5

Candidate Designs

Inference at the boundaries

Earliest possible stopping time

	OBF	Poc
S_M / N_M	7 / 10	4 / 5
P val ($\pi = 0.2$) (SM)	.0009	.0067
Estimate (BAM)	.675	.753
95% CI (SM)	.347, .859	.283, .915

Candidate Designs

Inference at the boundaries

Smallest rejection of Null

	OBF	Poc
S_M / N_M	12 / 35	14 / 35
P val (π = 0.2) (SM)	.0447	.0196
Estimate (BAM)	.321	.354
95% CI (SM)	.183, .488	.209, .542

Candidate Designs

Inference at the boundaries

Largest nonrejection of Null

	OBF	Poc
S_M / N_M	11 / 35	13 / 35
P val (π = 0.2) (SM)	.0774	.0259
Estimate (BAM)	.297	.333
95% CI (SM)	.167, .462	.199, .518

Stopping Rule

Pocock bounds

Conservatism of O'Brien-Fleming less desirable for new therapy

Fifth patient (no prior chemotherapy) had toxicity

Trial stopped (modified)

Clinical Trial Results

Toxicities	4 / 5
P values	
♦ Sample Mean	.00674
♦ Analysis Time	.00672
Point Estimates	
♦ Bias adjusted mean	.753
♦ UMVUE	.800
♦ MUE (Sample Mean)	.784
♦ MUE (Analysis Time)	.767
♦ MLE	.800
Confidence Intervals	
♦ Sample mean	.283, .915
♦ Analysis time	.284, .947

Data Driven Selection of Stopping Rule

Model hybrid test

Y is number of toxicities in first four patients

If Y < c, stay with fixed sample design

If Y > c, switch to group sequential test

$$\Pr[rej\ H_0;\pi] = \sum_{y=0}^{4} \Pr[rej\ H_0 \mid Y = y;\pi]\Pr[Y = y;\pi]$$

First term computed under FST or (shifted) GST according to value of c

Sensitivity Analysis Results

Size, Power of Hybrid Tests

Threshold for switch to GST	Size (π = 0.2)	Power (π = 0.5)
0 / 4 (GST)	.0196	.9292
1 / 4	.0205	.9343
2 / 4	.0218	.9466
3 / 4	.0208	.9551
4 / 4	.0155	.9557
5 / 4 (FST)	.0142	.9552

Selected Bayesian Analysis Results

Uniform prior

Obs Tox	E(π|S)	Pr(π<0.2|S)	Pr(π>0.5|S)
0 / 1	.333	.360	.250
1 / 1	.667	.040	.750
2 / 2	.500	.104	.500
2 / 2	.750	.008	.875
2 / 3	.600	.027	.688
3 / 3	.800	.002	.938
3 / 4	.667	.007	.812
3 / 5	.571	.017	.656
4 / 5	.714	.002	.891

Selected Bayesian Analysis Results

***Ad hoc* prior (uniform mass on null: 0.5)**

Obs Tox	E(π|S)	Pr(π<0.2|S)	Pr(π>0.5|S)
0 / 1	.209	.628	.091
1 / 1	.471	.176	.460
2 / 2	.365	.289	.271
2 / 2	.590	.050	.671
2 / 3	.483	.104	.476
3 / 3	.664	.013	.808
3 / 4	.565	.032	.646
3 / 5	.490	.061	.485
4 / 5	.623	.009	.770

Selected Bayesian Analysis Results

***Ad hoc* prior (uniform mass on null: 0.8)**

Obs Tox	E(π|S)	Pr(π<0.2|S)	Pr(π>0.5|S)
0 / 1	.135	.871	.031
1 / 1	.354	.462	.300
2 / 2	.256	.619	.145
2 / 2	.532	.174	.584
2 / 3	.404	.316	.363
3 / 3	.645	.049	.778
3 / 4	.530	.116	.590
3 / 5	.438	.208	.410
4 / 5	.611	.035	.750

Final comments

Hybrid rule could have been more complicated to account for later decisions to switch

Sensitivity analysis suggests appropriate inference in this case (could use as a criterion for GST)

Adjusted inference possible, but more complex

Bayesian analysis of some interest, but it is questionable that a proper prior could ever be selected to detect unexpected toxicities

Documentation of Design, Monitoring, and Analysis Plans

Documentation of Design

Specification of stopping rule

- Null, design alternative hypotheses
- Type I error (alpha, beta parameters)
- Power to detect design alternative
- One-sided, two-sided hypotheses (epsilon parameters)
- Boundary scale for design family
- Boundary shape function parameters (P, R, A) for each boundary
- Constraints (minimum, maximum, exact)

Documentation of Design

Documentation of stopping rule

- Specification of stopping rule
- Estimation of sample size requirements
- Example of stopping boundaries under estimated schedule of analyses
  - sample mean scale
  - other scales?
- Inference at the boundaries
- Futility, Bayesian properties?

Documentation of Implementation

Specification of implementation methods

- 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)

Documentation of Analysis

Specification of analysis methods

- Method for determining P values
- Method for point estimation
- Method for confidence intervals
- (Handling additional data that accrues after decision to stop)