
Design and Interim Monitoring of Flexible Clinical Trials Featuring the East Software

Cyrus R. Mehta

President, Cytel Software Corporation

and

Adjunct Professor, Harvard University

email: mehta@cytel.com – web: www.cytel.com – tel: 617-661-2011

Topics Covered in this Talk

1. What are flexible clinical trials
2. Two examples of flexible trials
 - Normal endpoint cholesterol reduction study – illustrates spending functions, stopping boundaries and gaining early insights from the trial
 - Binomial endpoint psoriasis study – illustrates sample size re-adjustment
3. Open discussion of our software, training, design and interim monitoring services

What are Flexible Clinical Trials

Traditional Design: fix the sample size in advance and only perform **one** efficacy analysis after all subjects have been enrolled and evaluated.

Flexible Design: monitor the accruing efficacy data at administratively convenient intervals and make important decisions concerning the future course of the study along the way.

Advantages of Flexible Clinical Trials

- **Early intimation of efficacy**; option to either terminate or prepare for an early regulatory submission.
- **Early intimation of inefficacy**; option to either terminate for futility, drop the ineffective arm, or divert key resources to more promising studies.
- **Verify design assumptions** (variance, effect size, covariates, etc) from accumulating data; option to revise the sample size to avoid an underpowered study.

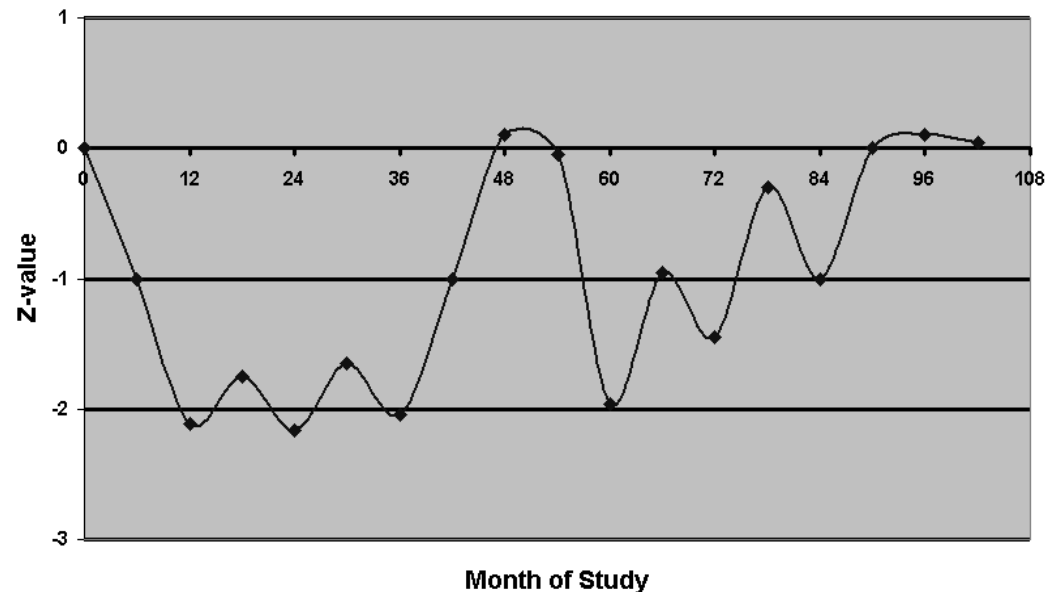
What are the Logistical problems?

- Access to the interim results by trial investigators could **bias** the future conduct of the study
- For **pivotal** trials, the potential for bias is controlled through:
 - An external Data Monitoring Committee
 - An external statistical center preparing the interim analysis reports

What are the Statistical Problems?

- If you take multiple looks you are much more likely to see spurious effects due to chance fluctuations in the data.

Coronary Drug Project (1966-1974)



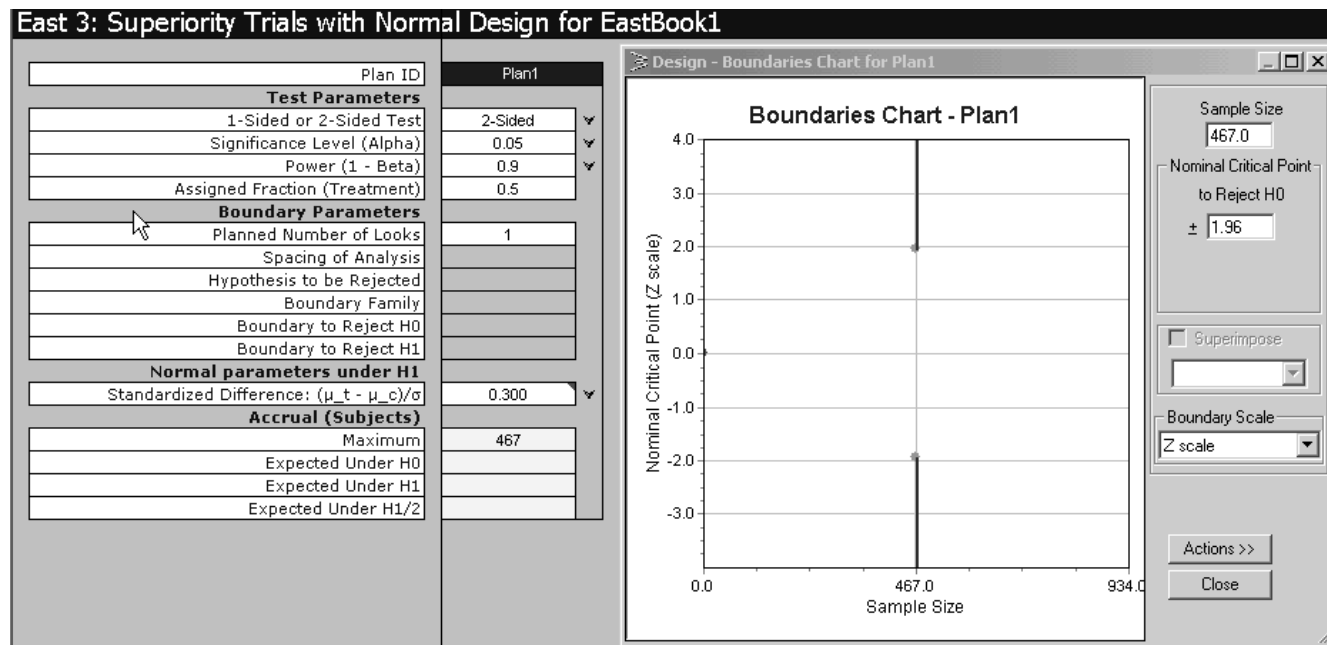
Appropriate Statistical Methods are Necessary

- Unless appropriate statistical methods are used, multiple looks, premature termination, or boosting the sample size can affect the type-1 and type-2 errors
- **The Good News:** Appropriate statistical methods for interim monitoring do exist, are endorsed in the FDA Guidance Document ICH-E9, and validated software for implementing these methods is now available

Cholesterol Reduction Example

- Placebo controlled efficacy trial with primary endpoint being reduction in total serum cholesterol over a 4-week period (Facey, 1992).
- A subject receives either Control drug (C) or Experimental drug (E).
- Design for 90% power to detect an average benefit of 60 mg/dl for drug E compared to drug C. Two-sided test with 5% significance level.
- Patient to patient variability assumed to be $\sigma^2 = 200^2$.

Static Single-Look Design



How interim monitoring can help

The static design makes many assumptions that cannot be corrected in mid-stream

- Maybe $\delta \ll 60$. If so, we'd like to know early and cut our losses
- Maybe $\delta \gg 60$. If so, we'd like to know early and possibly terminate for efficacy
- Maybe $\delta \approx 60$ but $\sigma > 200$. If so, we'd like to increase the sample size

More complex designs

- Similarly, in **survival studies** or **longitudinal studies** we make assumptions about event rates, accrual rates, drop-out rates, missing values, covariate effects, etc.
- Sample size calculations based on these assumptions may lead to underpowered studies

In Summary: Interim monitoring makes it possible to make early termination decisions and/or to implement mid-course corrections to the study design

Frequently Expressed Concerns

1. Will type-1 error be preserved?
2. Is there a penalty for taking interim looks?
 - (a) stricter p-value requirements?
 - (b) larger sample size requirement?
3. Will the sponsor be locked into a rigid schedule with respect to the number and timing of the interim looks?

How to take interim looks and also protect the type-1 error

- Choose an α -spending function to specify how the type-1 error will be spent over the course of the trial
- The spending function should be appropriate to the needs of the trial

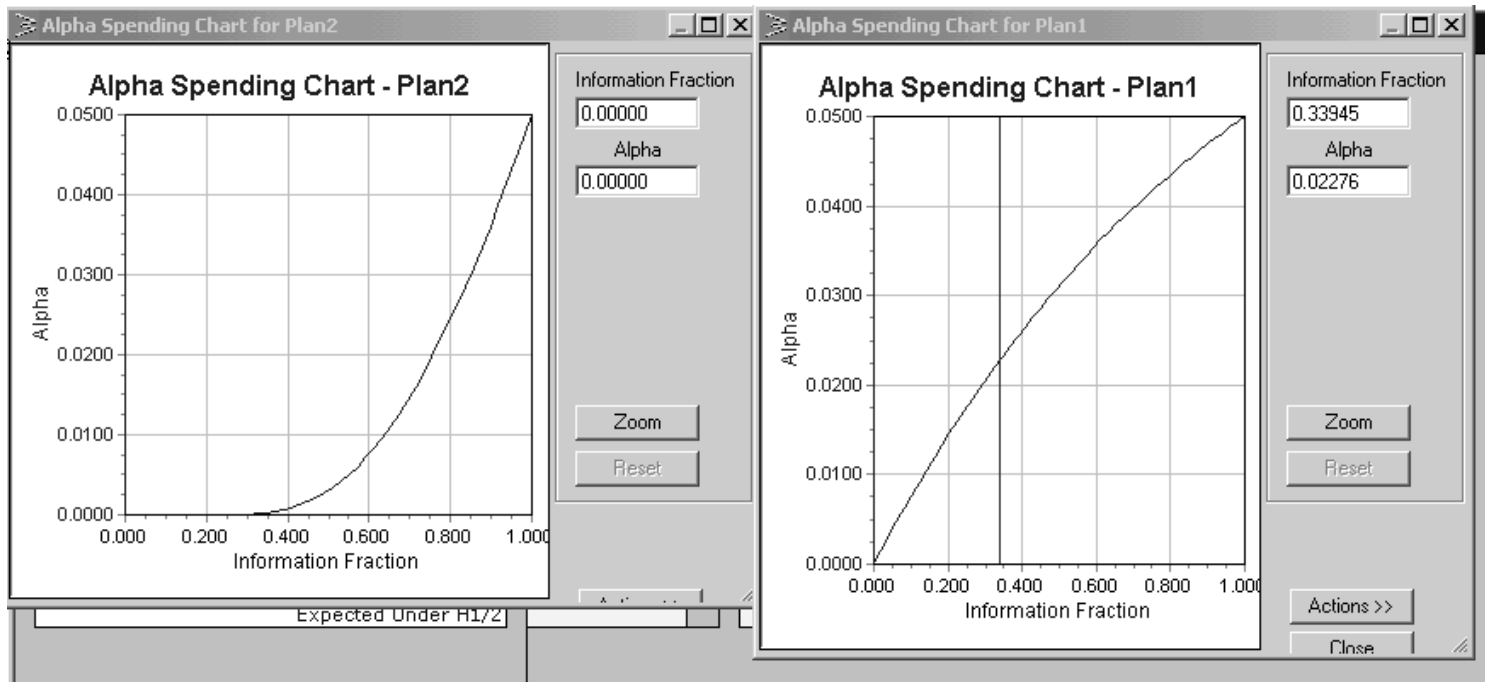
Conservative Not intended for early efficacy stopping.

Aproprate for administrative looks, futility stopping or sample size adjustment.

Aggressive Primarily for early efficacy stopping.

Using East for the Design

East 3: Superiority Trials with Normal Design for EastBook1					
Plan ID	Plan1	Plan2	Plan3	Plan4	
Test Parameters					
1-Sided or 2-Sided Test	2-Sided	2-Sided	2-Sided	2-Sided	
Significance Level (Alpha)	0.05	0.05	0.05	0.05	
Power (1 - Beta)	0.9	0.9	0.9	0.9	
Assigned Fraction (Treatment)	0.5	0.5	0.5	0.5	
Boundary Parameters					
Planned Number of Looks	1	3	3	3	
Spacing of Analysis		Equal	Equal	Equal	
Hypothesis to be Rejected		H0 Only	H0 Only	H0 Only	
Boundary Family		SpF(Pub)	SpF(Pub)	SpF(Pub)	
Boundary to Reject H0		LD(OF)	Gm(-2)	Gm(1)	
Boundary to Reject H1					
Normal parameters under H1					
Standardized Difference: $(\mu_t - \mu_c)/\sigma$	0.300	0.300	0.300	0.300	
Accrual (Subjects)					
Maximum	467	473	487	540	
Expected Under H0		471	482	529	
Expected Under H1		379	348	337	
Expected Under H1/2		453	454	481	



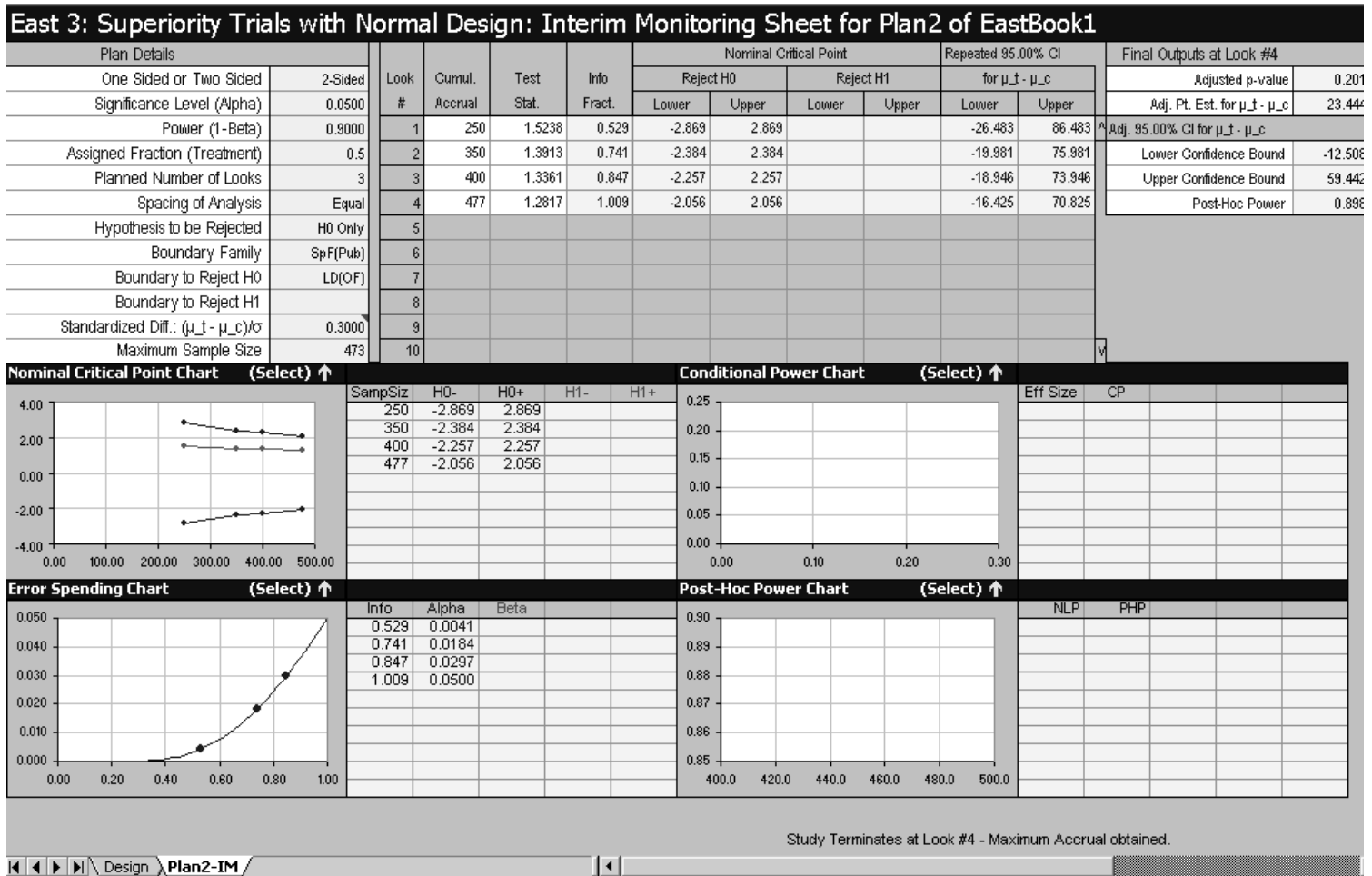
- There are many different types of spending functions, not just the two proposed by Lan and DeMets
- Industry statisticians are often at odds with FDA statisticians about the so called “administrative look”. They want to take an interim look but pay no penalty for it on the grounds that they have no intention of stopping the trial based on that look. But under ICH-E9 burden of proof is on sponsor to show that type-1 error was preserved. FDA is always concerned that if a large difference emerges trial will be terminated regardless of prior assurances to the contrary. Therefore they will insist that a penalty, in terms of a stricter final p-value must be paid. Sponsors then propose the Haybittle-Peto family of stopping boundaries. Conservative spending functions are an alternative, more flexible way to take these interim looks.
- You pay just the right amount of penalty for taking multiple looks. And it is a small penalty, but formally endorsed.

Deviations from Planned Interim Analyses

Look (j)	N_j	t_j	$\hat{\delta}(t_j)$	$se[\hat{\delta}(t_j)]$
1	250	0.529	30	19.688
2	350	0.741	28	20.125
3	400	0.847	27.5	20.583
4	477	1.009	27.2	21.221

- It is commonly believed that, under ICH E9, the number and timing of the interim looks must be fixed in advance and rigidly adhered to. This is not the case. The ICH E9 allows flexibility to deviate from these design parameters if the protocol specifies that a group sequential design is being used. “The protocol should describe the schedule of interim analyses or at least the considerations that will govern its generation, for example if flexible alpha spending function approaches are to be employed.”
- The important quantities to fix in the design and rigidly adhere to in the interim monitoring are the choice of spending function and the maximum sample size (or maximum information, in the case of an information-based design).

Interim Monitoring with East



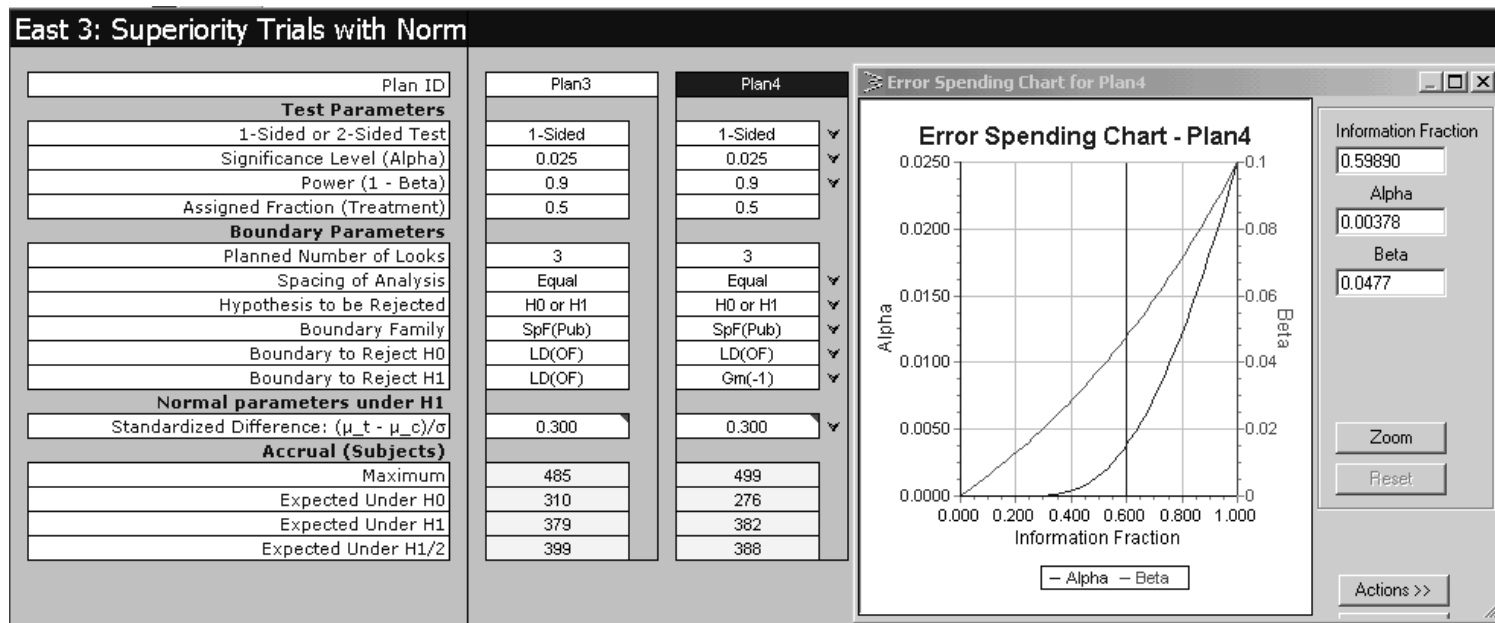
- Conditional power is flexible but ad-hoc. $CP(\text{designed delta})$, $CP(\text{estimated delta})$, somewhere in between
- How low is low? 20%? 10%?
- Formal Futility Boundaries are an alternative approach

Formal Futility Boundaries

- We could have terminated the trial for futility at look 3.
(CP = 7.5%)
- Could we have terminated even earlier?
 - At look 1, CP = 56.7%
 - At look 2, CP = 22.3%
- CP is an arbitrary criterion. Formal futility boundaries provide a more efficient and objective early stopping criterion.

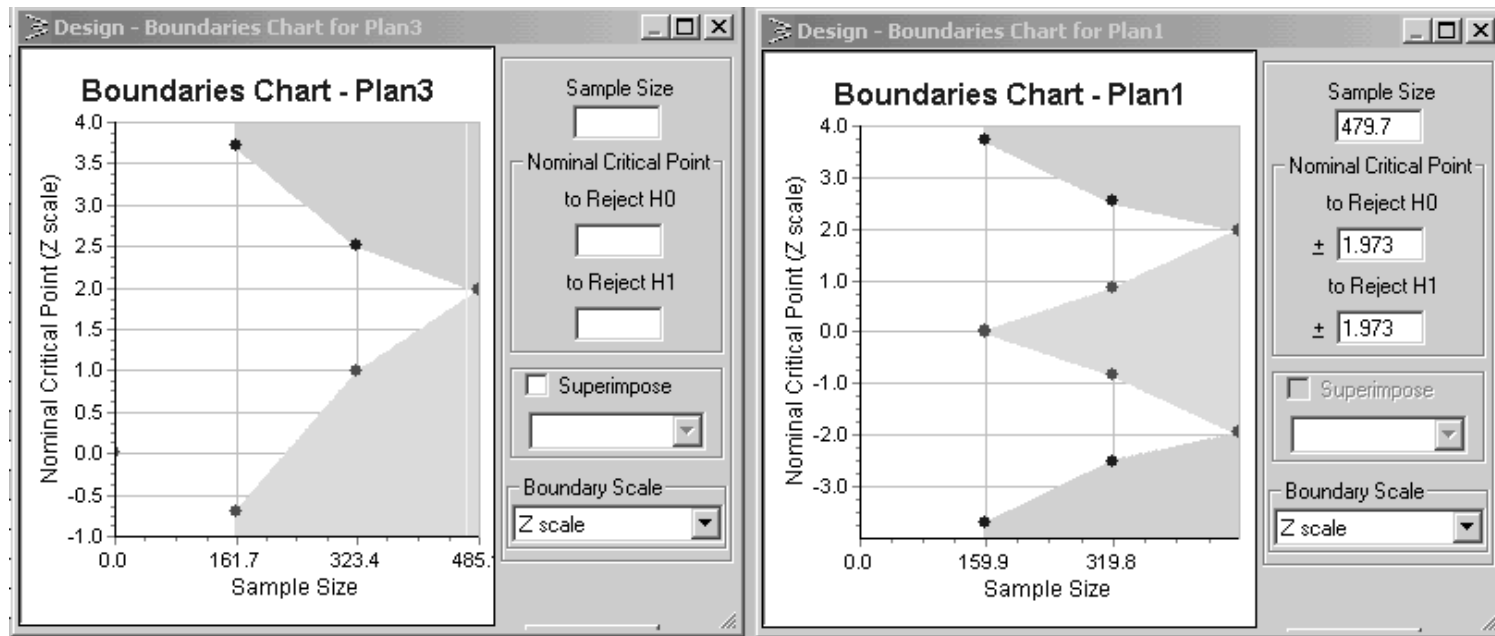
The β -Spending Function

We can generate formal futility boundaries by specifying how we wish to spend β , the type-2 error. In the current example, $\beta = 1 - 0.9 = 0.1$.



Corresponding Stopping Boundaries

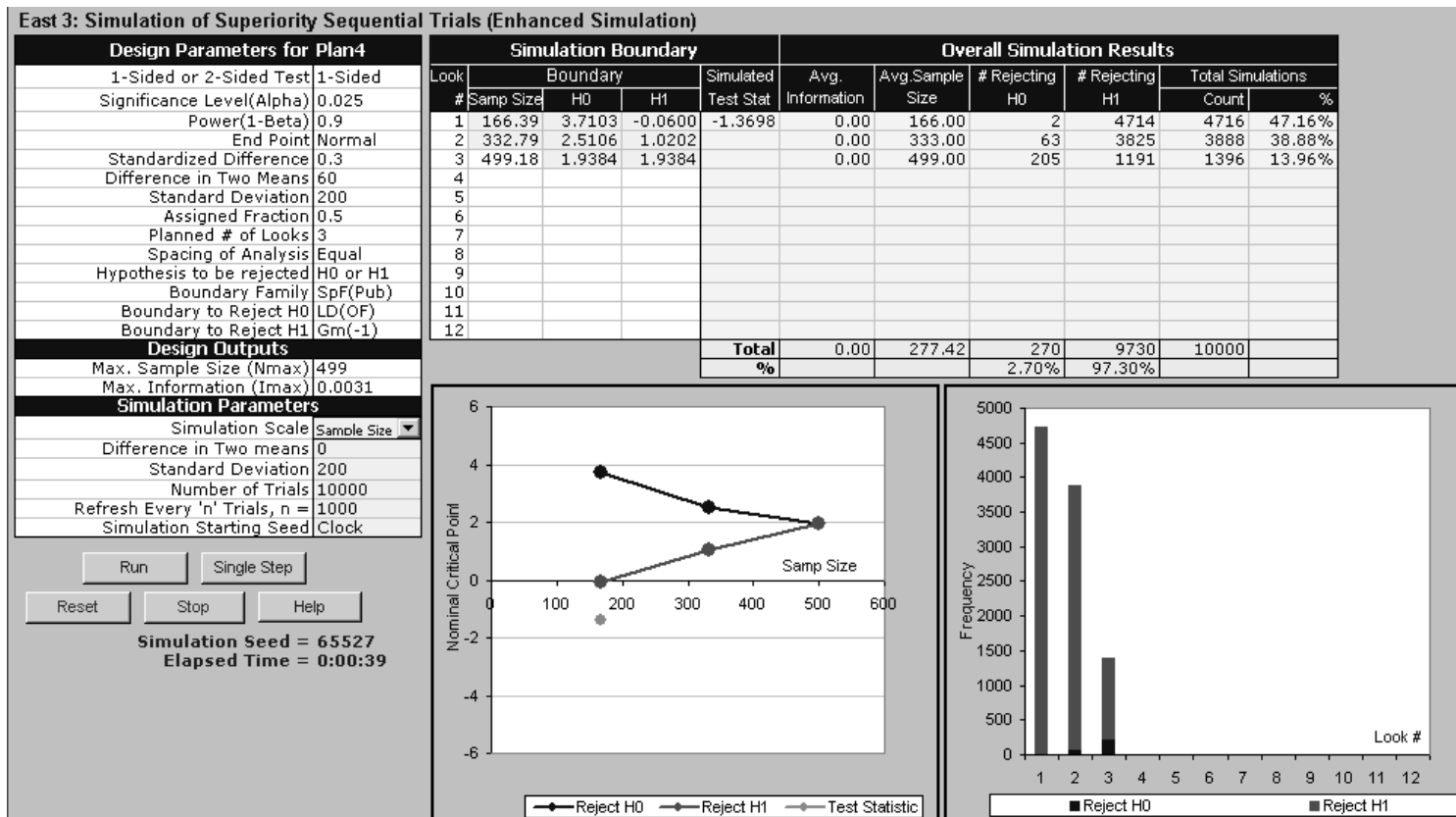
The corresponding stopping boundaries are generalizations of Whitehead's (1992) triangular boundaries.



Examining the Design Analytically

East 3: Superiority Trials with Norm						
Plan ID	Plan3	Plan4	Cumulative Accrual	Stopping Probabilities		
Test Parameters						
1-Sided or 2-Sided Test	1-Sided	1-Sided	166	0.476	0.061	0.155
Significance Level (Alpha)	0.025	0.025	333	0.386	0.583	0.361
Power (1 - Beta)	0.9	0.9	499	0.138	0.356	0.484
Assigned Fraction (Treatment)	0.5	0.5				
Boundary Parameters						
Planned Number of Looks	3	3				
Spacing of Analysis	Equal	Equal				
Hypothesis to be Rejected	H0 or H1	H0 or H1				
Boundary Family	SpF(Pub)	SpF(Pub)				
Boundary to Reject H0	LD(OF)	LD(OF)				
Boundary to Reject H1	LD(OF)	Gm(-1)				
Normal parameters under H1						
Standardized Difference: $(\mu_t - \mu_c)/\sigma$	0.300	0.300				
Accrual (Subjects)						
Maximum	485	499				
Expected Under H0	310	276				
Expected Under H1	379	382				
Expected Under H1/2	399	388				

Examining the Design by Simulation



Summary

- **It is legal to take multiple looks at accruing data**
- **the type-1 error is controlled through a spending function**
- **The penalty for taking the interim looks is controlled by the choice of spending function**
- **Early insights about treatment efficacy can be obtained through conditional power and formal efficacy or futility boundaries**

Binomial Example: Sample Size Re-estimation

- Treatment versus placebo for high-need **psoriasis** patients. Primary endpoint is attainment of PASI-75 by week 16
- Design for 95% power to detect a 10% improvement with new treatment relative to placebo
- 2-sided test at 5% significance level
- Current best guess is that placebo rate is 7.5%. But it could be anywhere in the range 5% to 15%

Sensitivity of Sample Size to Placebo Response Rate

All these plans specify an improvement of $\delta = 0.1$ for the treatment arm. But the sample sizes are all different. They depend on π_c , the placebo response rate.

East 3: Superiority Trials with Binomial Design for EastBook3				
Plan ID	Plan1	Plan2	Plan3	Plan4
Test Parameters				
1-Sided or 2-Sided Test	2-Sided	2-Sided	2-Sided	2-Sided
Significance Level (Alpha)	0.05	0.05	0.05	0.05
Power (1 - Beta)	0.95	0.95	0.95	0.95
Assigned Fraction (Treatment)	0.5	0.5	0.5	0.5
Boundary Parameters				
Planned Number of Looks	3	3	3	3
Spacing of Analysis	Equal	Equal	Equal	Equal
Hypothesis to be Rejected	H0 Only	H0 Only	H0 Only	H0 Only
Boundary Family	SpF(Pub)	SpF(Pub)	SpF(Pub)	SpF(Pub)
Boundary to Reject H0	LD(OF)	LD(OF)	LD(OF)	LD(OF)
Boundary to Reject H1				
Binomial parameters under H1				
Proportion Response (Control: n_c)	0.075	0.100	0.150	0.050
Proportion Response (Treatment: n_t)	0.175	0.200	0.250	0.150
Accrual (Subjects)				
Maximum	562	657	828	460
Expected Under H0	559	654	824	458
Expected Under H1	426	498	627	349
Expected Under H1/2	533	623	785	436

This is a situation in which the company wants to differentiate its product from its competitor. Suppose that the product label for the competitor shows an 8% improvement. Thus the company wants to power the study to show at least a 10% improvement. The 10% is a given. It is not something to be estimated.

On the other hand, the placebo response rate is not a given and should be taken into account for sample size calculations

Maximum Sample Size Study

- The maximum **sample size** for a K -look, 2-sided, level- α , binomial endpoint study, having power $1 - \beta$ to detect a difference in response rates of δ , is given by:

$$N_{\max} = 2[\pi_c(1-\pi_c) + (\pi_c + \delta)(1-\pi_c - \delta)] \times \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta} \right]^2 \times \mathbf{IF}$$

where IF is an inflation factor that depends on K , α , β and the spending function.

- Notice that N_{\max} depends on π_c , the response rate of the control arm

Maximum Information Study

- The maximum **information** for a K -look, 2-sided, level- α , binomial endpoint study, having power $1 - \beta$ to detect a difference in response rates of δ , is given by:

$$I_{\max} = \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta} \right]^2 \times \mathbf{IF}$$

where information (or Fisher information) is estimated by the square inverse standard error of the estimate of δ .

- Notice that I_{\max} does not depend on π_c .

Fix Information – Be Flexible with Sample Size

- The information about δ available at an interim look is estimated by

$$I = \frac{1}{[\text{se}(\hat{\delta})]^2}$$

- Each time we monitor the data we can compute I . As soon as I equals or exceeds

$$I_{\max} = \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta} \right]^2 \times \text{IF}$$

full power is attained and the study is terminated.

- Information is **fixed** whereas sample size **floats**

Information Based Design of Psoriasis Study

East 3: Superiority Trials with Info-Based Design	
Plan ID	Plan1
Test Parameters	
1-Sided or 2-Sided Test	2-Sided
Significance Level (Alpha)	0.05
Power (1 - Beta)	0.95
Assigned Fraction (Treatment)	0.5
Boundary Parameters	
Planned Number of Looks	3
Timing of Analysis	Equal
Hypothesis to be Rejected	H0 Only
Boundary Family	SpF(Pub)
Boundary to Reject H0	LD(OF)
Boundary to Reject H1	
Info-Based parameters under H1	
Effect Size	0.100
Information	
Maximum	1313.9831
Expected Under H0	1308.5942
Expected Under H1	995.9137
Expected Under H1/2	1245.9583

Keep the study open until either a boundary is crossed or $I = 1314$ is reached

Relating Information to Sample Size

Since N_{\max} is allowed to float until $I_{\max} = 1314$, tabulate range of plausible values of N_{\max} in the protocol.

π_c	0.05	0.075	0.10	0.15
N_{\max}	460	562	657	828
I_{\max}	1314	1314	1314	1314

ICH E9 Position on Sample Size Re-estimation

“An interim check conducted on blinded data may reveal that overall response variances, event rates or survival experience are not as anticipated. A revised sample size may then be calculated using suitably modified assumptions, and should be justified and documented in a protocol amendment and in the clinical study report.”

- Blinded estimates of current information are possible. But it would be more accurate to have an independent organization compute current information with unblinded data
- Alternatively if EDC technology is being used, one could integrate appropriate statistical procedures into the EDC system so that current information, current accrual, estimated time to complete the trial are automatically, and other statistics for managing the trial are automatically generated on an interim monitoring report.
- Information based designs are routinely used in survival analysis where number of events, not sample size captures the information.
- We presented the information based approach to the FDA in a workshop and it was well received. There are several publications on it by Mehta, Tsiatis and Scharfstein. It is built into East. The climate at FDA has changed. They actually encourage innovated methods that improve trial efficiency provided issues of bias and type-1 error inflation can be guarded against.

Summary

- Many parameters that affect the sample size of a trial are only guesses at the design stage. (e.g., π_c , σ^2 , baseline hazard)
- Sample size calculations based on these guesses could lead to underpowered studies
- But interim monitoring permits us to estimate the current information about δ from the actual data of the trial, re-estimate sample size, and preserve the power of the study

- **Information based designs can be applied even more generally to test the primary endpoint while controlling for covariates in regression models**
- **Osteoporosis Example:**

$$\log(\text{odds of response}) = \beta_0 + \beta_1 \times \text{treatment} + \beta_2 \times \text{prior fract}$$

The information based approach provides a smooth way to obtain just the right sample size

- The analysis plan usually specifies that the primary efficacy endpoint will be evaluated in a regression model so as to control for the effects of key covariates
- For example in an Osteoporosis study, prior bone fracture, age, weight and race might be very important covariates and would be included along with treatment as part of a logistic regression model for the final analysis.
- In a maximum sample size design it would be almost impossible to compute the sample size that takes these covariate effects into account. These effects would not be known at the design stage.
- In a maximum information design one would simply fit the regression model to the interim data and estimate the standard error of the treatment coefficient. The square inverse standard error is the estimate of current information
- The trial stops when full information is reached

Adaptive Designs

- In some cases there is uncertainty about the value of δ at which to provide adequate power for the study; i.e., what value of δ is clinically meaningful?
- For example, at the design stage it was believed that $\delta \geq 10\%$ is clinically meaningful. On that basis suppose we have designed the study to detect an improvement from 5% on the control arm to 15% on the treatment arm
- Could we change our mind about the clinically meaningful effect size after the study has been activated?

Why would you change your mind

- **External evidence from other trials might necessitate change**
- **Competitor's product might be withdrawn from the market**
- **Not much experience with current product. Hence would like the flexibility to perform a mid-course correction**

Psoriasis Example, continued

- The majority of patients belonged to a high-need subgroup for which the effect size was not well established
- An interim look was taken half way through the trial and it demonstrated only a 5% improvement relative to placebo
- However, even 5% improvement is acceptable, because there is no other treatment for high need patients
- Unless the sample size is increased, this trial will fail.

Could this trial be saved?

Yes! An adaptive approach might save this trial

- **Basic idea of adaptive trials:**
 - increase the sample size if the observed data suggest loss of power
 - at the same time make appropriate adjustments to preserve the type-1 error
- **Extensive Literature: Bauer and Kohne (1994), Proschan and Hunsberger (1995), Cui, Hung and Wang (1999), Shen and Fisher (1999), Lemacher and Wassmer (1999), Liu and Chi (2001), Lan and Trost (1997), Tsiatis and Mehta (2003).**

The CHW adaptive method

- Two-stage trial is planned for n_1 observations/arm at stage 1 and n_{\max} observations/arm overall.
- At stage 1 compute increase sample size so as to recover:

Unconditional Power

$$n_{\max}^* = \left[\frac{\delta}{\hat{\delta}} \right]^2 n_{\max}$$

Conditional Power

$$1 - \Phi \left[\frac{u_2 \sqrt{n_{\max}^*} - z_1 \sqrt{n_1} - \sqrt{(n_{\max} - n_1)(n_{\max}^* - n_1)} \hat{\delta} / \hat{\sigma}}{\sqrt{n_{\max}^* - n_1}} \right]$$

where $\hat{\sigma} = \sqrt{2\hat{\pi}_t(1 - \hat{\pi}_t) + 2\hat{\pi}_t(1 - \hat{\pi}_t)}$

Test Statistic at Final Look

- At the final stage the usual test statistic is $Z = \hat{\delta}/\text{se}(\hat{\delta})$. If we do not adapt we can re-write Z as

$$Z = \sqrt{\left(\frac{n_1}{n_{\max}}\right)} \sum_{i=1}^{n_1} \frac{(X_{ti} - X_{ci})}{\hat{\sigma}\sqrt{n_1}} + \sqrt{\left(\frac{n_{\max} - n_1}{n_{\max}}\right)} \sum_{i=n_1+1}^{n_{\max}} \frac{(X_{ti} - X_{ci})}{\hat{\sigma}\sqrt{n_{\max} - n_1}}$$

- If we increase the sample size from n_{\max} to n_{\max}^* , the statistic becomes

$$Z^* = \sqrt{\left(\frac{n_1}{n_{\max}^*}\right)} \sum_{i=1}^{n_1} \frac{(X_{ti} - X_{ci})}{\hat{\sigma}\sqrt{n_1}} + \sqrt{\left(\frac{n_{\max}^* - n_1}{n_{\max}^*}\right)} \sum_{i=n_1+1}^{n_{\max}^*} \frac{(X_{ti} - X_{ci})}{\hat{\sigma}\sqrt{n_{\max}^* - n_1}}$$

Downweight the final test statistic

Cui, Hung and Wang (1999) propose that instead of using

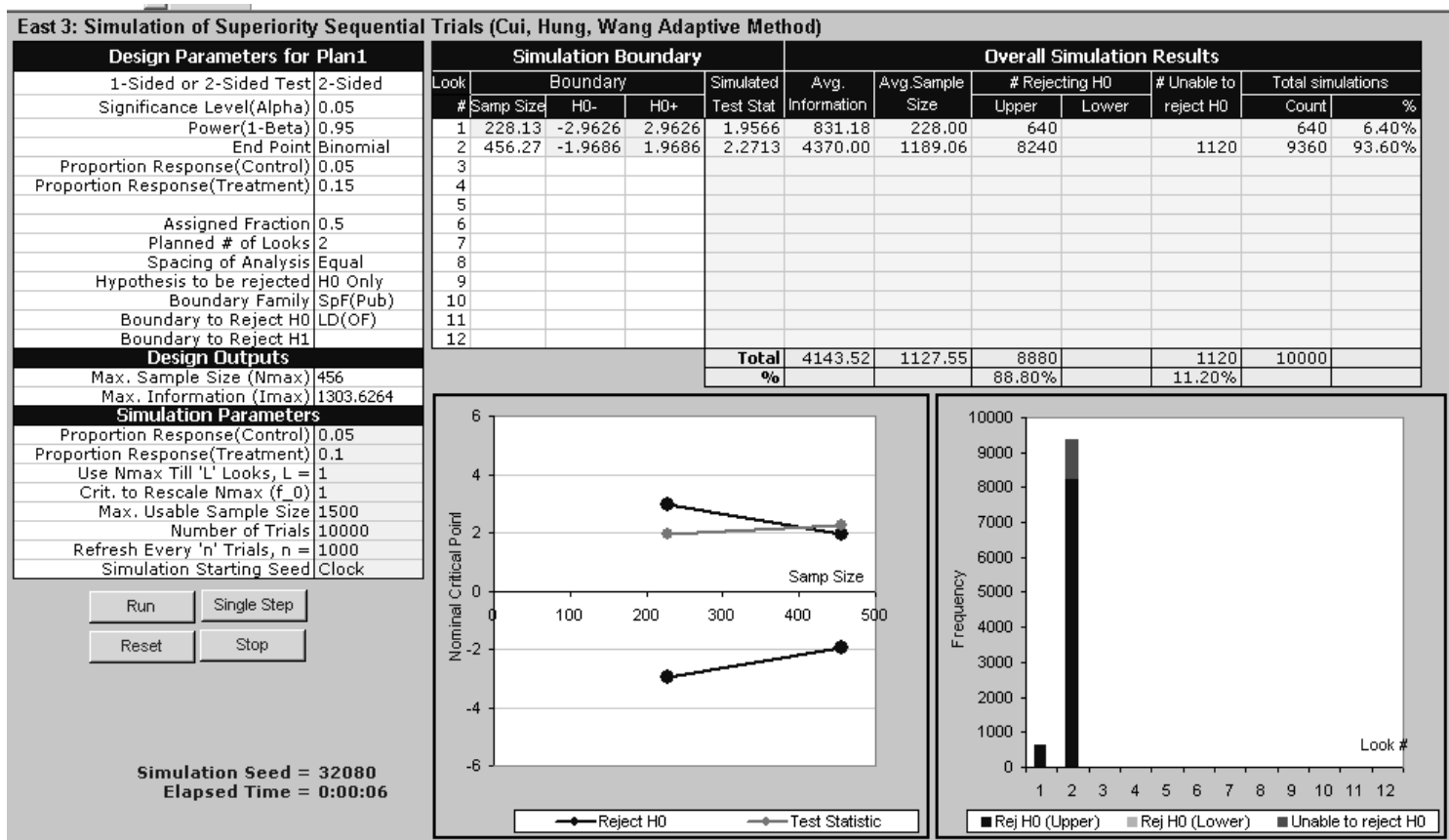
$$Z^* = \sqrt{\left(\frac{n_1}{n_{\max}^*}\right)} \sum_{i=1}^{n_1} \frac{(X_{ti} - X_{ci})}{\hat{\sigma} \sqrt{n_1}} + \sqrt{\left(\frac{n_{\max}^* - n_1}{n_{\max}^*}\right)} \sum_{i=n_1+1}^{n_{\max}^*} \frac{(X_{ti} - X_{ci})}{\hat{\sigma} \sqrt{n_{\max}^* - n_1}}$$

we should use

$$Z^{**} = \sqrt{\frac{n_1}{n_{\max}}} \sum_{i=1}^{n_1} \frac{(X_{it} - X_{ic})}{\sqrt{\hat{\sigma}} \sqrt{n_1}} + \sqrt{\frac{n_{\max} - n_1}{n_{\max}}} \sum_{i=n_1+1}^{n_{\max}^*} \frac{(X_{it} - X_{ic})}{\sqrt{\hat{\sigma}} \sqrt{(n_{\max}^* - n_1)}}$$

- Under H_0 , the distribution of Z^{**} is identical to that of Z . So type-1 error is preserved
- But the contribution of the stage 1 patients counts more than the contribution of the stage 2 patients.

Simulation of Adaptive Design



Cytel Offerings

- **Training on statistical methodology and software use**
- **Identification of studies that are suitable for interim monitoring**
- **Exploring the benefits of information based and adaptive designs**
- **Involving Cytel in these studies early on, at the protocol development stage**