
Sample Size Re-estimation in Confirmatory Clinical Trials

PhRMA Adaptive Working Group
KOL Series, April 9, 2010

Cyrus R. Mehta

President, Cytel Inc.

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

Reasons for Sample Size Re-estimation

1. Due to uncertainty about variability in the data (σ^2)
 - Usually handled in-house without unblinding the data
2. Due to uncertainty about primary effect effect size (δ)
 - Must unblind the interim data
 - Requires an independent interim analysis committee

This talk deals with unblinded sample size re-estimation. We shall discuss two case studies, some statistical methodology and some regulatory issues

Case I: Negative Symptoms Schizophrenia

- New Drug versus Control for negative symptoms schizophrenia trial
- Primary endpoint is the change in negative symptoms assessment (NSA) **at week 26** relative to the baseline
- Smallest clinically meaningful effect size is 0.2
- Sponsor prefers to power for effect size of 0.27

Why Sponsor Chose $\delta = 0.27$

Normal Superiority Trials: Two-Sample Test - Difference of Means		
Plan ID	Plan1	Plan2
Test Parameters		
1-Sided or 2-Sided Test	2-Sided	2-Sided
Significance Level (α)	0.05	0.05
Power (1 - β)	0.9	0.9
Assigned Fraction (Treatment)	0.5	0.5
Boundary Parameters		
Planned Number of Looks	1	1
Spacing of Looks		
Hypothesis to be Rejected		
Boundary Family		
Boundary to Reject H0		
Boundary to Reject H1		
Normal Parameters		
Difference of Means (δ_{11})	0.27	0.2
Standard Deviation (σ)	1.0	1.0
Accrual (Subjects)		
Maximum	577	1051
Expected Under H0		
Expected Under H1		
Expected Under H1/2		

As is typical, the sample size decision is heavily influenced by sponsor's resource constraints. Sponsor can free up resources for at most 600 subjects **up-front**. Working backwards, study is powered at $\delta = 0.27$

What's the Right Sample Size?

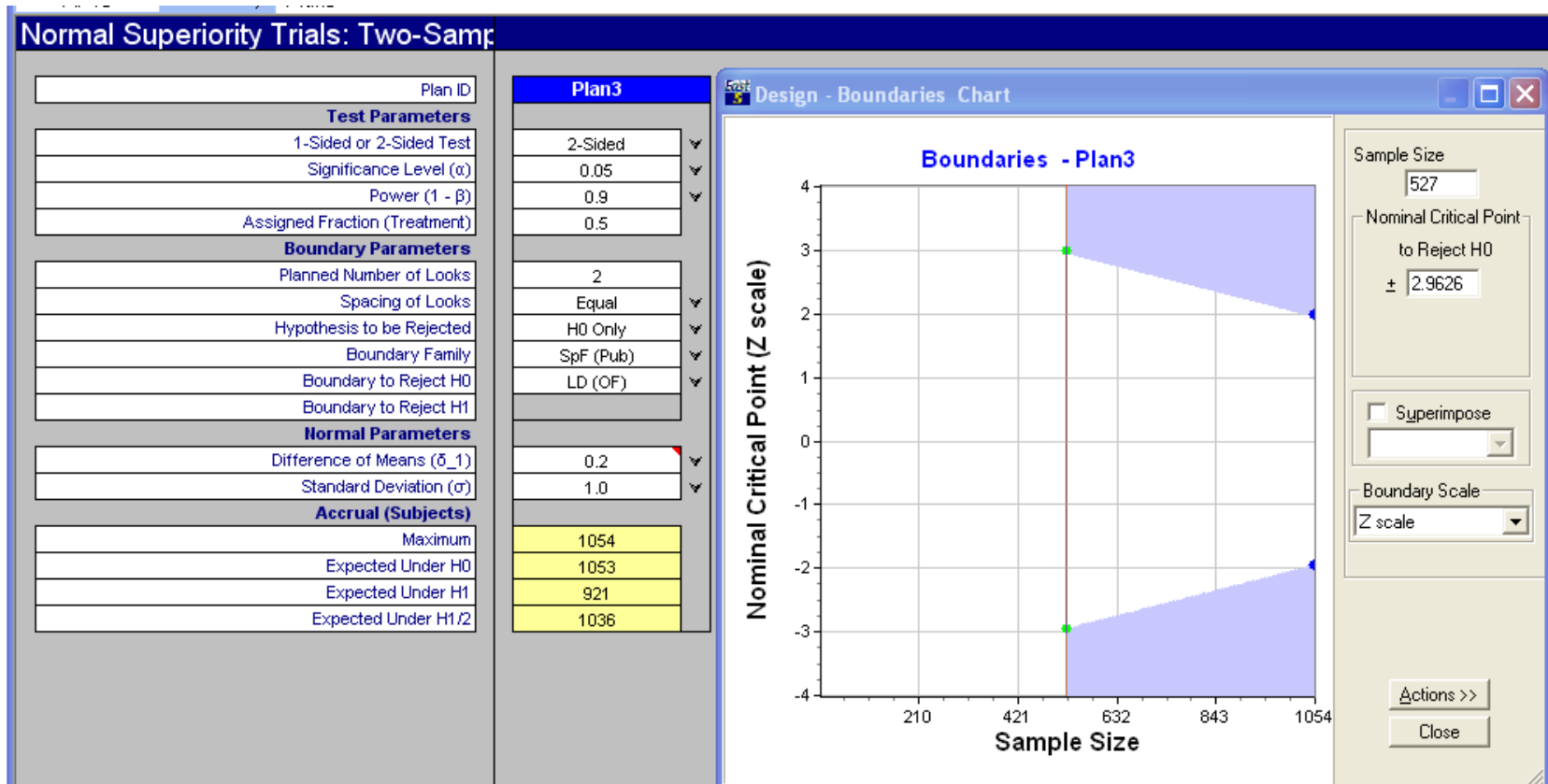
Table 1: Operating Characteristics of Plan1 and Plan2

δ	Plan1		Plan2	
	Sample Size	Power	Sample Size	Power
0.27	577	90%	1051	99%
0.25	577	85%	1051	98%
0.23	577	79%	1051	96%
0.21	577	71%	1051	93%
0.20	577	67%	1051	90%

- Plan1 is adequately powered if $\delta = 0.27$ but underpowered if $\delta = 0.2$
- Plan2 is adequately powered if $\delta = 0.2$ but overpowered if $\delta = 0.27$

Try a Group Sequential Design

Design for $\delta = 0.2$ possible early stopping if interim results are compelling, as they will be if in truth $\delta = 0.27$

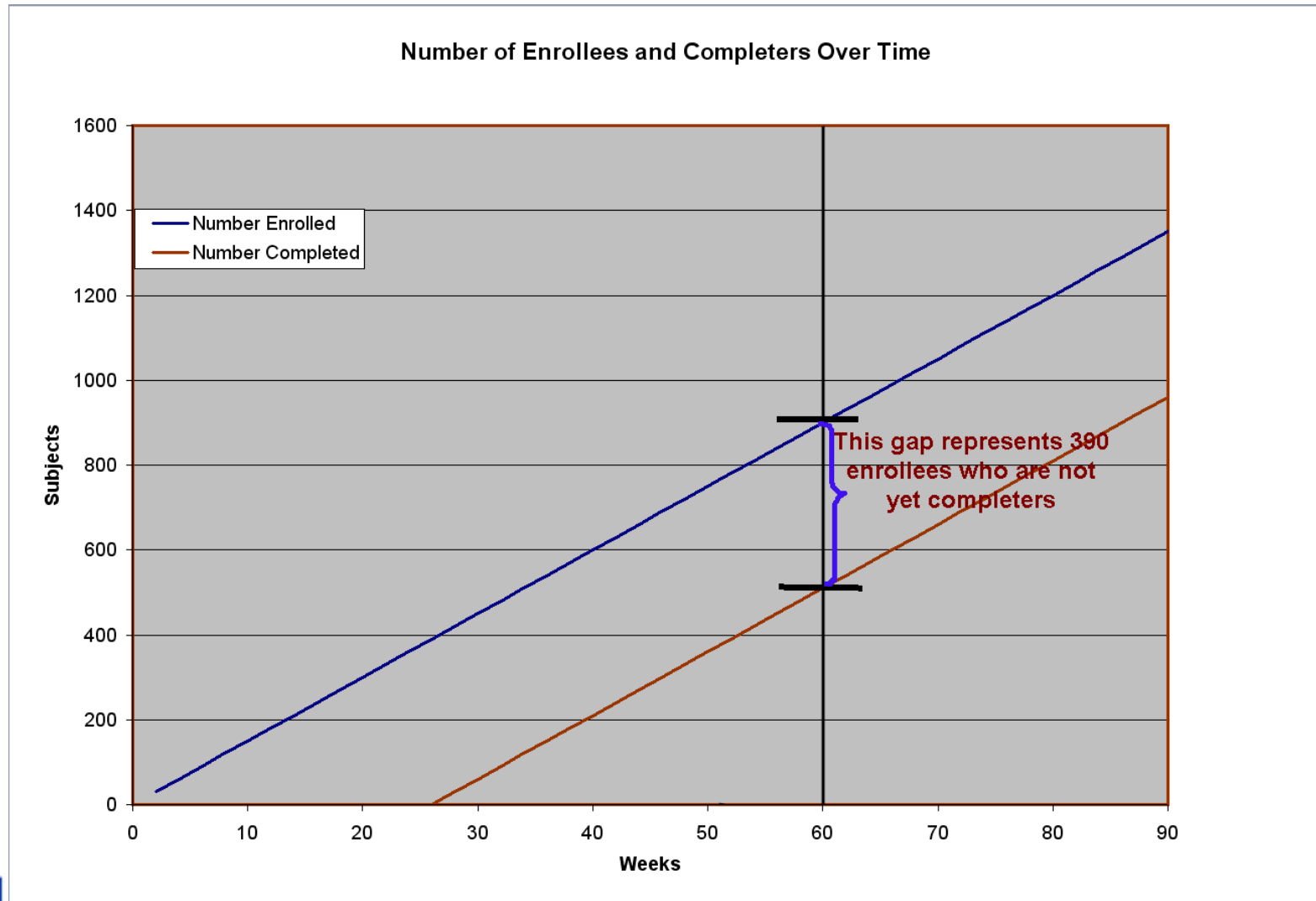


Practical Problems with Group Sequential Design

- **OVERRUNS CAN OFFSET THE SAMPLE SIZE SAVINGS**
- **TOUGH BOUNDARIES LOWER THE CHANCE OF EARLY STOPPING**

The Problem of Overruns

26-week endpoint; 15/week enrollment; overrun is $15 \times 26 = 390$



Why Tough Boundaries?

- **General reluctance to spend much alpha at the interim**
- **Interim results must be very compelling in order to alter medical practice if trial is terminated prematurely**
- **Hence chances of early efficacy stopping are low. But up-front commitment is high**

Adaptive Design

1. Commit 577 subjects up-front (90% power at $\delta = 0.27$)
2. Perform an interim analysis at week 36 with $15 \times 36 = 540$ enrolled
 - 150 completers and 390 still in follow-up
3. Perform interim analysis only on the 150 completers
4. Increase the sample size, and hence the power, if the interim results fall in a “promising” zone

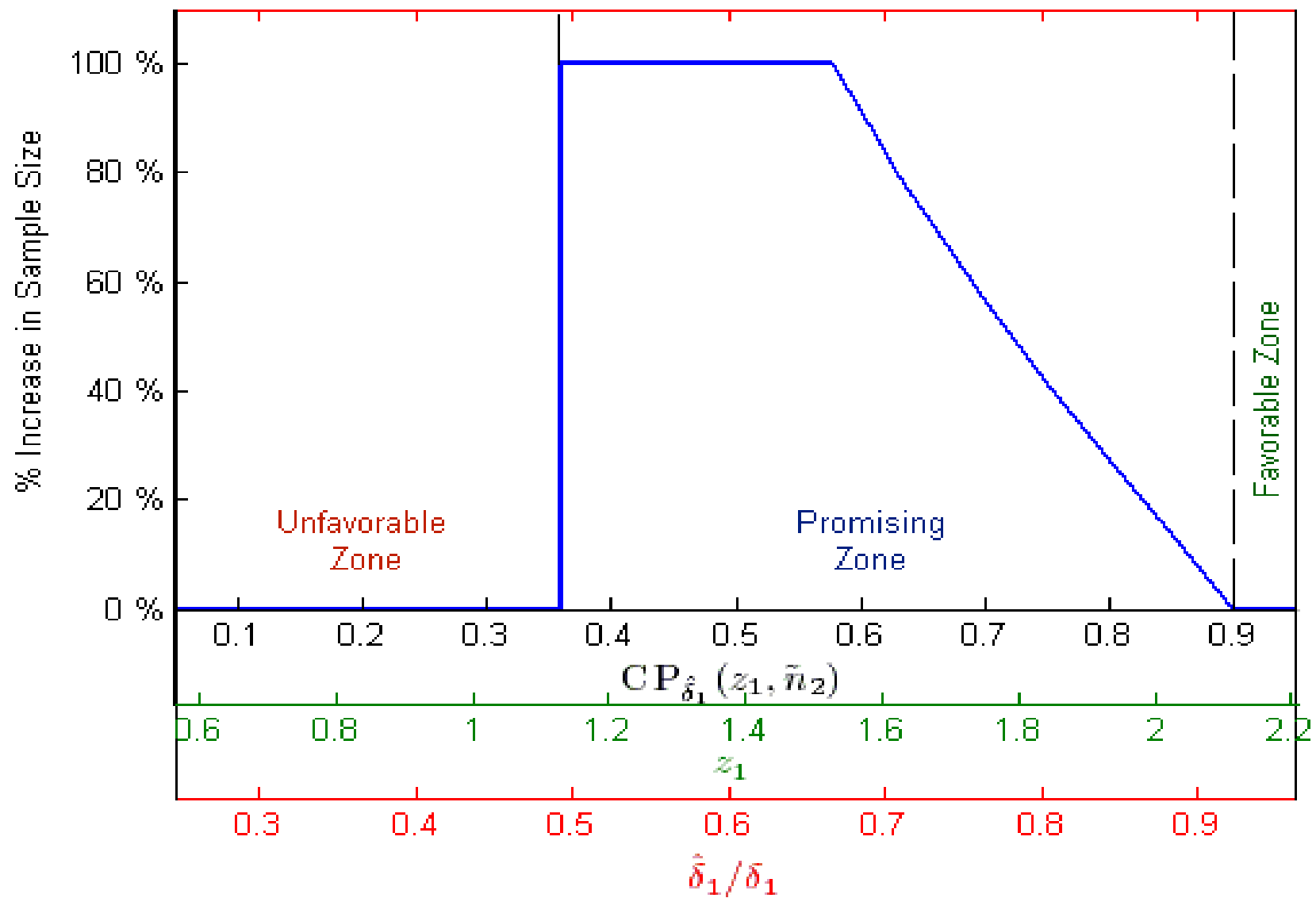
Sample Size Re-estimation Rule

- Partition interim outcome into unfavorable, promising and favorable zones, based on **conditional power**
- Only increase sample size if interim outcome is **promising**
- Specify a **target conditional power** and calculate sample size needed to achieve the target
- Set minimum and maximum limits on sample size change

Specifying the Promising Zone

Criterion for Adapting: Min. CP	0.3600
Max. CP	0.9000
Min. Usable Sample Size	577
Max. Usable Sample Size	1154
Desired Conditional Power (CP)	0.9000

- If CP at the interim is between 0.36 and 0.9, **the outcome is considered promising** and sample size is increased
 - CP = 0.36 corresponds to $\hat{\delta} = 50\%$ of design δ
 - CP = 0.9 corresponds to $\hat{\delta} = 95\%$ of design δ
- New sample size is set so as to boost the CP to 0.9
- Range of new sample size is 577 (original) to 1154 (double)



Results of 100,000 Simulated Trials: Unconditional and by Zone

Overall Simulation Results					
Avg. Info.	Avg. Sample Size	# Rejecting H0	# Unable to Reject H0	Total Simulations	
				Count	%
173.94	695.76	73284	26716	100000	100.00%
173.94	695.76	73284	26716	100000	
		73.28%	26.72%		
Simulation Results for Adapted Trial Only					
240.85	963.40	27208	3526	30734	100.00%
240.85	963.40	27208	3526	30734	
		88.53%	11.47%		

Simulation Results by Zone							
Zone	Avg. Sample Size	Simulations Rejecting H0		Simulations not Rejecting H0		Total Simulations	
		Count	%	Count	%	Count	%
Unfavorable: CP < 0.300	577.00	14503	44.87%	17819	55.13%	32322	32.32%
Promising: 0.300 <= CP < 0.900	963.40	27208	88.53%	3526	11.47%	30734	30.73%
Favorable: CP >= 0.900	577.00	31573	85.46%	5371	14.54%	36944	36.94%

Table 2: Operating Characteristics of Fixed Sample and Adaptive Designs

Value of δ	Fixed Sample		Adaptive	
	Power	N	Power	E(N)
0.27	90%	577	93%	677
0.23	79%	577	83%	689
0.20	67%	577	73%	695

- Modest power gain is offset by corresponding sample size increase
- But adaptive design reduces the sponsor's risk:
 - it only increases sample size if interim result is promising
 - if that happens, the payoff to sponsor is huge
 - if not, sponsor is no worse off than before

Table 3: Operating Characteristics Conditional on Interim Outcome

δ	Interim Outcome	Probability of Interim Outcome	Power Conditional on Interim Outcome		Expected Sample Size	
			Fixed	Adaptive	Fixed	Adaptive
0.27	Unfavorable	15%	74%	74%	577	577
	Promising	31%	88%	98%	577	944
	Favorable	53%	97%	97%	577	577
0.23	Unfavorable	22%	57%	57%	577	577
	Promising	34%	77%	94%	577	956
	Favorable	44%	92%	92%	577	577
0.2	Unfavorable	28%	45%	45%	577	577
	Promising	36%	67%	88%	577	963
	Favorable	37%	86%	86%	577	577

The Value Proposition

- With fixed 577 subjects, trial only has 67% power
- With fixed 1051 subjects, trial has 90% power; but up front commitment too large
- With adaptive trial sponsor's risk is reduced:
 - Only commits resources for 577 subjects initially
 - If interim result is promising, then commits up to 577 additional subjects
 - Happy to make the additional commitment since it raises the power substantially

Case II: Acute Myeloid Myeloma

- Primary endpoint is overall survival
- Design for 90% power; 5% significance level
- Plan for 24 month enrollment; 30 month trial
- Optimistic Scenario
 - Assume 5/7 month median on Ctrl/Trtm (HR=0.71)
 - Require 375 events and 450 subjects @ 19/month
- Pessimistic Scenario
 - Assume 5/6.5 month median on Ctrl/Trtm (HR=0.77)
 - Require 616 events and 732 subjects @ 31/month
 - Not a feasible option for sponsor

Sponsor is Resource and Time Constrained

- Unable to invest up-front to protect power in case of pessimistic scenario
- But willing to invest additional resources if interim results are promising

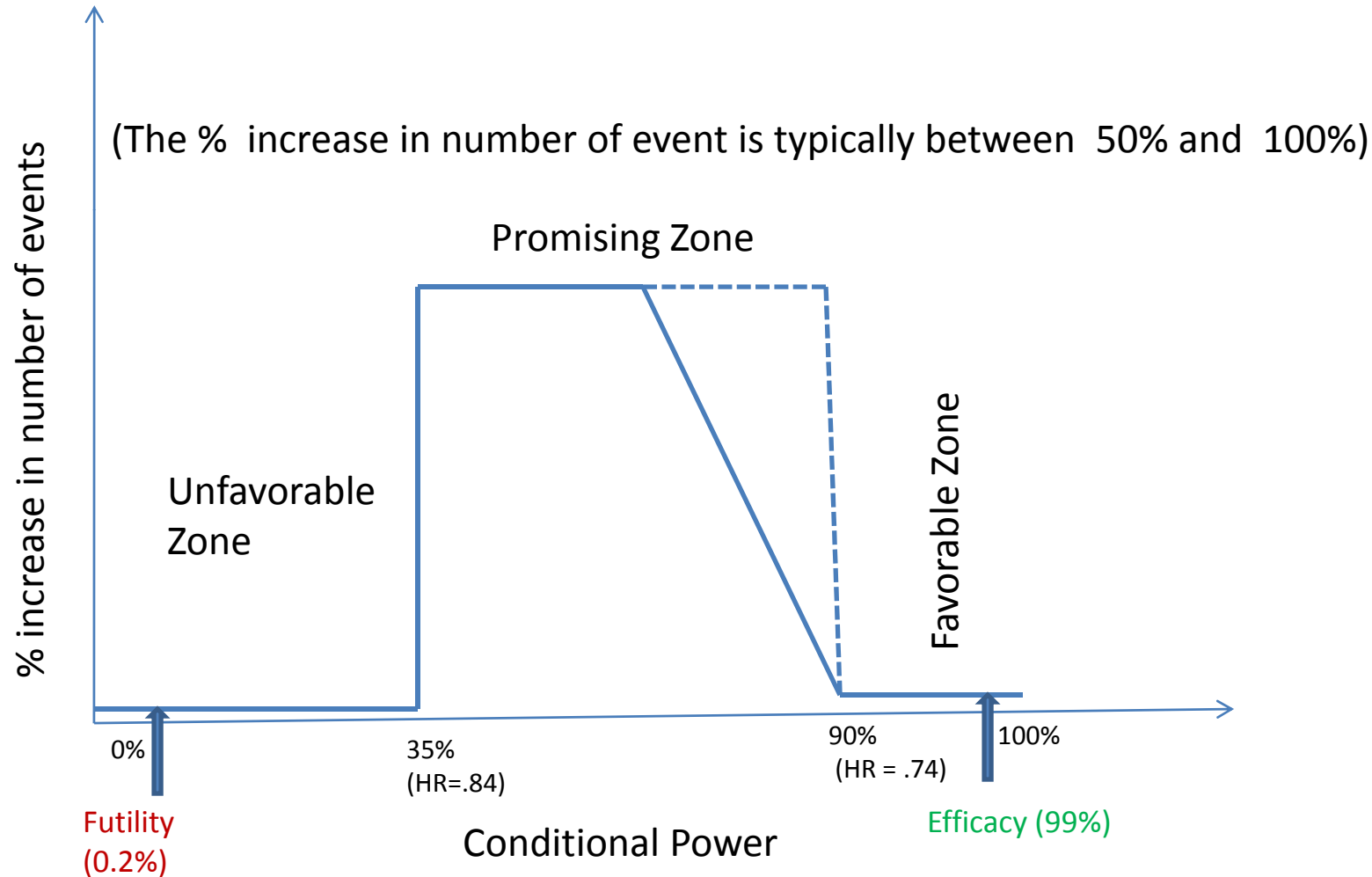
True HR	Power of Optimistic Design	Power of Pessimistic Design
0.71	91%	99%
0.74	83%	97%
0.77	71%	90%

Sponsor Adopts an Adaptive Strategy

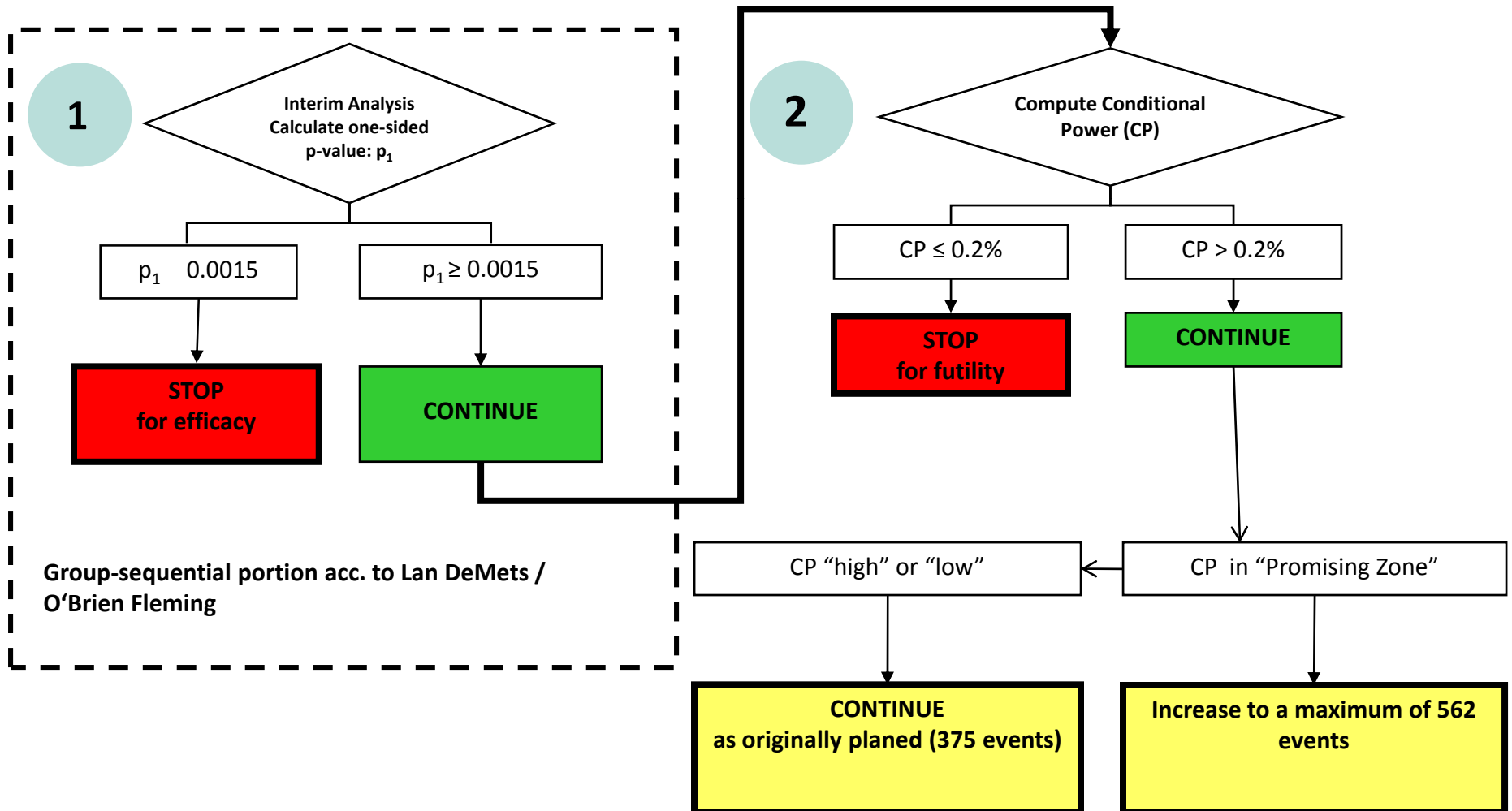
- Design optimistically (HR=0.71; 375 events; 450 subjects @ 19/month)
- One interim analysis after 50% information
 - Stop early if overwhelming evidence of efficacy
 - Stop early for futility if low conditional power
 - Increase number of events, sample size and (if possible) rate of recruitment at the interim **if results are promising**
- Promising zone is $(30\% \leq \text{CP} < 90\%)$ or equivalently $(0.84 \geq \text{estimated HR} > 0.74)$

The Adaptive Decision Rules

- Conditional Power = Prob of success at end of trial given interim results
- Increase the number of events if conditional power is in the Promising Zone



Flow chart for the Adaptive Decision Rules



Operating Characteristics

1. Under Pessimistic Scenario, HR = 0.77 (10,000 simulations)

Zone	P(Zone)	Power		Duration (months)		SampSize	
		NonAdpt	Adapt	NonAdpt	Adapt	NonAdpt	Adapt
Unf	25%	33%	35%	28	28	436	439
Prom	34%	71%	90%	29	38	453	680
Fav	41%	95%	95%	26	26	414	413

2. Under Optimistic Scenario, HR = 0.71 (10,000 simulations)

Zone	P(Zone)	Power		Duration		SampSize	
		NonAdpt	Adapt	NonAdpt	Adapt	NonAdpt	Adapt
Unf	12%	57%	53%	29	29	441	443
Prom	28%	87%	99%	30	39	453	680
Fav	60%	99%	98%	29	25	402	400

Preservation of type-1 Error

- Design a K -look group sequential design with boundaries b_1, b_2, \dots, b_K at **cumulative** sample sizes n_1, n_2, \dots, n_K
- Suppose these sample sizes are changed over the course of the trial to $n_1^*, n_2^*, \dots, n_K^*$
- Define the **incremental** sample sizes $n^{(j)} = (n_j - n_{j-1})$ and $n^{*(j)} = (n_j^* - n_{j-1}^*)$, $j = 1, 2, \dots, K$
- Define the weights $w_j = (n^{(j)} / n_K)$ and $w_j^* = (n^{*(j)} / n_K^*)$
- Let $Z^{*(j)}$ be the incremental Wald statistic based only on the $n^{*(j)}$ new observations between looks $(j - 1)$ and j

The Weighted Wald Statistic

- The weighted Wald statistic is defined as

$$Z_{j,\mathbf{w}}^* = \frac{\sqrt{w^{(1)}} Z^{*(1)} + \sqrt{w^{(2)}} Z^{*(2)} + \dots + \sqrt{w^{(j)}} Z^{*(j)}}{\sqrt{w^{(1)} + w^{(2)} + \dots + w^{(j)}}}$$

- This statistic is asymptotically normally distributed with mean

$$E(Z_{j,\mathbf{w}}^*) = \frac{\delta \sum_{l=1}^j \sqrt{w^{(l)} I^{*(l)}}}{\sqrt{\sum_{l=1}^j w^{(l)}}}$$

and unit variance, where $I^{*(l)}$ is the incremental information at look l

- Cui, Hung and Wang (1999) and Lehmacher and Wassmer (1999) have shown that

$$P_0\left(\bigcup_{j=1}^K |Z_{j,\mathbf{w}}^*| \geq b_j\right) = \alpha .$$

- Note: If no sample size change, then $Z_{j,\mathbf{w}} = Z_{j,\text{wald}}$

Objection to Weighted Statistic

- Weights must be pre-specified instead of reflecting the sample sizes that were actually used in the trial
 - This has the undesirable effect of down-weighting future cohorts of patients. (Since $n^{*(j)} > n^{(j)}$, it follows that $w_j^* > w_j$)
 - Weighted statistics are non-intuitive, and difficult to explain to non-statisticians
- Large increases in sample size based on tiny interim estimates of δ lead to inefficient designs (Jennison and Turnbull, 2003). (This is not a problem if sample size increases are modest (up to a doubling, say) and only adopted in a reasonably conservative promising zone)

Dispensing with Weighted Statistic

- Works for K -stage designs with adaptation at stage $K - 1$
- Consider the 2-stage design. Define

$$Z = \frac{\hat{\delta}}{\text{se}(\hat{\delta})}; \quad Z_1 = \frac{\hat{\delta}_1}{\text{se}(\hat{\delta}_1)}; \quad Z_2 = \frac{\hat{\delta}_2}{\text{se}(\hat{\delta}_2)}$$

- Conventional Final Analysis: $Z \geq z_\alpha$

Non-conventional Final Analysis: $w_1 Z_1 + w_2 Z_2 \geq z_\alpha$

- w_1 and w_2 are **pre-specified** weights with $w_1^2 + w_2^2 = 1$
(typically $w_j = \sqrt{\frac{n_j}{n_1 + n_2}}$ even if n_2 is increased to n_2^*)
 - Stage 2 subjects contribute less than Stage 1 subjects
 - Estimation of δ is complicated by the weighted statistic

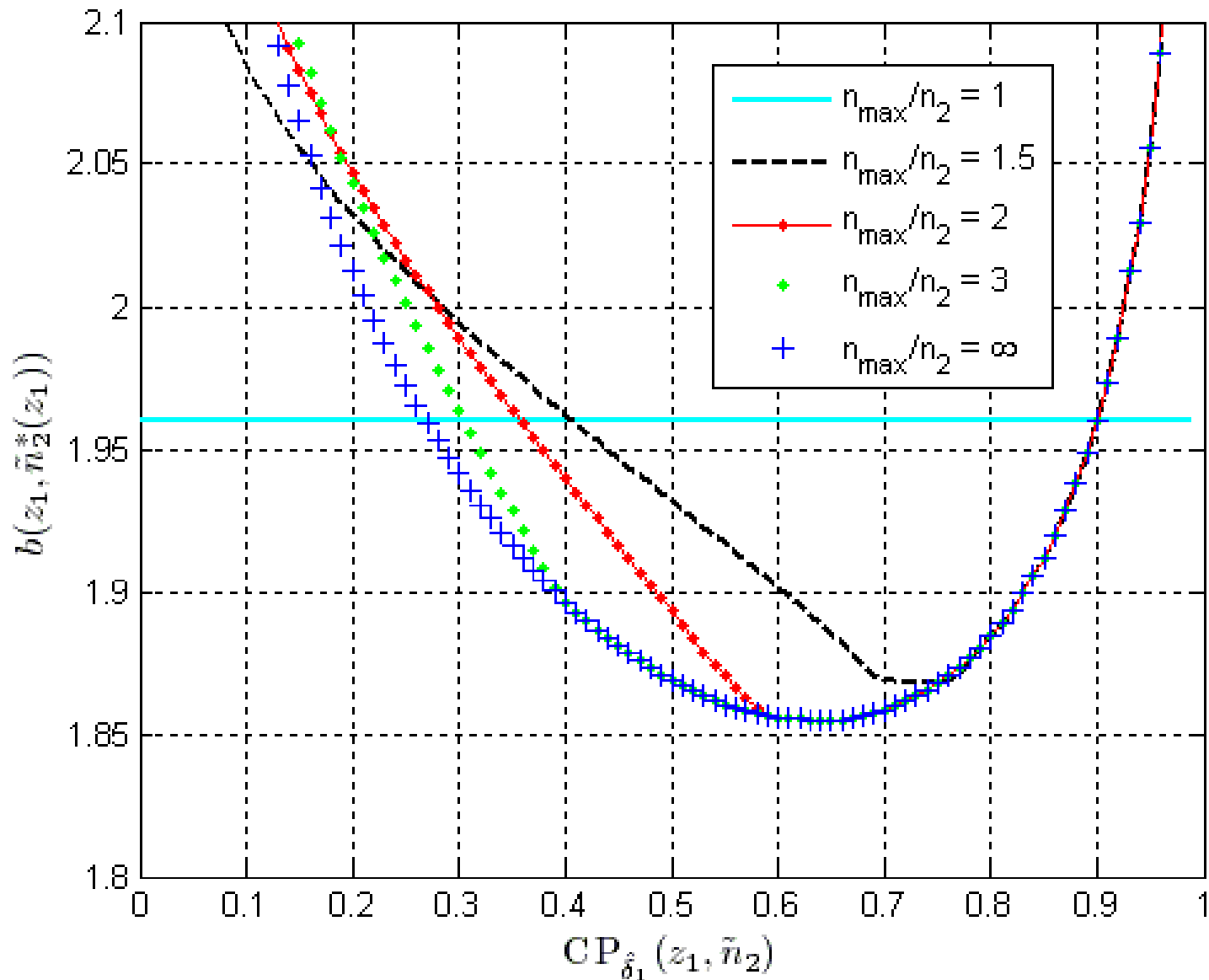
Theorem (Gao et. al., 2008)

$$P_0(w_1 Z_1 + w_2 Z_2 \geq z_\alpha) = P_0(Z \geq b(z_1, n_2^*))$$

where $b(z_1, n_2^*) = (n_2^*)^{-0.5} \left[\sqrt{\frac{n_2^*}{n_2}} (z_\alpha \sqrt{n_2} - z_1 \sqrt{n_1}) + z_1 \sqrt{n_1} \right]$

- OK to use the conventional statistic Z at the final analysis, if we adjust the critical region from z_α to $b(z_1, n_2^*)$
- Since z_1 is a random variable we won't know the final critical region until interim analysis
- But we can plot $b(z_1, n_2^*)$ versus z_1 ahead of time and study its behaviour

$b(z_1, n_2^*)$ vs. CP at Interim



CP_{min} for Various Design Options

Sample Size Ratios		CP _{min} Values for Targeted	
Maximum Allowed	At Interim Look	Conditional Powers	
(n_{\max}/n_2)	(n_1/n_2)	80%	90%
1.5	0.25	0.42	0.42
1.5	0.5	0.41	0.41
1.5	0.75	0.38	0.38
2	0.25	0.37	0.37
2	0.5	0.36	0.36
2	0.75	0.33	0.33
3	0.25	0.32	0.32
3	0.5	0.31	0.31
3	0.75	0.30	0.27
∞	0.25	0.32	0.28
∞	0.5	0.31	0.27
∞	0.75	0.30	0.25

Use of Conventional Test Statistic

- Can use the conventional test if promising zone is defined as $CP_{\min} \leq CP < \text{Targeted CP}$
- Negligible power loss due to conservatism
- Extends related work by Chen, DeMets and Lan (2004) who showed that weighted statistic can be dispensed with as long as $CP \geq 50\%$

Attractiveness of Approach

- Up-front sample size investment can be modest
- Additional investment is only made if interim results are promising
- If that happens, chances of success are dramatically increased
- Use of weighted statistics can be dispensed with. This greatly simplifies the adaptive design and makes it amenable to routine use

Implications of FDA Adaptive Guidance for Unblinded SSR

- This design falls into the category titled, “Adaptive Study Designs whose Properties are Less Well Understood”
- Guidance recommends using the method only for increasing sample size, not for decreasing
- Guidance recommends modest increases in sample size
- Guidance recommends using the method if the primary study objective cannot be achieved by other methods
- Guidance warns of “operational bias”. In the present context, one would have to address how the sponsor intends to prevent investigators from “reverse engineering” the treatment effect from knowledge of the adaptive decision

