

## Lecture 4: Interim Monitoring: Efficacy

## Introduction to Efficacy Monitoring

### Historical Boundaries

Haybittle-Peto

Pocock

O'Brien-Fleming

### Unified Approach: Information, Z-Scores, B-Values

### Alpha Spending Functions

Definition

O'Brien-Fleming-Like Spending Function

Pocock-Like Spending Function

Families of Spending Functions

### Effect of Monitoring on Power

### Case Study: CAST

### Small Sample Sizes

### Panoply of Problems Post-Monitoring

# Table of Contents

## Introduction to Efficacy Monitoring

### Historical Boundaries

Haybittle-Peto

Pocock

O'Brien-Fleming

### Unified Approach: Information, Z-Scores, B-Values

### Alpha Spending Functions

Definition

O'Brien-Fleming-Like Spending Function

Pocock-Like Spending Function

Families of Spending Functions

### Effect of Monitoring on Power

### Case Study: CAST

### Small Sample Sizes

### Panoply of Problems Post-Monitoring

# Introduction to Efficacy Monitoring

- ▶ Clinical trials are monitored by a ***Data and Safety Monitoring Board (DSMB)*** (also called a DSMC or DMC).
- ▶ Committee of 3-9 external experts (MDs, 1-2 statisticians, an ethicist).
- ▶ Typically meet 1-2 times a year.
- ▶ Review general trial conduct (accrual, data quality, missing data, etc.), safety (serious adverse events, unexpected events, etc.), futility, and efficacy.
- ▶ Make recommendations to trial sponsor, sponsor makes final decision (but almost always accepts recommendations).

# Introduction to Efficacy Monitoring

- ▶ Futility: Are data so unpromising or is trial conduct so poor that a null result is almost assured?
- ▶ Efficacy: If one arm is clearly superior, may stop trial or recommend change (e.g., announce result, make the superior treatment the new control, etc.).
- ▶ Problem: If we reject  $H_0$  whenever **nominal p-value** (not adjusted for monitoring) is  $\leq \alpha$ , type 1 error rate (probability of rejecting  $H_0$  at some point) is inflated.
- ▶ Even with only 1 interim and 1 final analysis,  
 $P(\text{rejecting } H_0 | H_0) = 0.083$  for 2-sided test at  $\alpha = 0.05$  if looks are equally-spaced.
- ▶ Armitage et al. (1969) showed inflation of type 1 error rate.

# Introduction to Efficacy Monitoring

- ▶ Situation can be worse if looks are not equally-spaced.
- ▶ Example: Suppose looks are after 10 and 10,000 observations.
  - ▶ The 2 p-values are nearly independent because the overlap is only 10 out of 10,000 people.
  - ▶ Independence is the worst case; the type 1 error rate is

$$\begin{aligned}P(\text{Reject } H_0 | H_0) &= P(P_1 \leq 0.05 \cup P_2 \leq 0.05) \\&= 1 - P(P_1 > 0.05 \cap P_2 > 0.05) \\&\approx 1 - P(P_1 > 0.05)P(P_2 > 0.05) \\&= 1 - (1 - 0.05)^2 = 0.0975. \quad (1)\end{aligned}$$

# Introduction to Efficacy Monitoring

- ▶ Situation can be worse if looks are not equally-spaced.
- ▶ Example: Suppose looks are after 10 and 10,000 observations.
  - ▶ The 2 p-values are nearly independent because the overlap is only 10 out of 10,000 people.
  - ▶ Independence is the worst case; the type 1 error rate is

$$\begin{aligned}P(\text{Reject } H_0 | H_0) &= P(P_1 \leq 0.05 \cup P_2 \leq 0.05) \\&= 1 - P(P_1 > 0.05 \cap P_2 > 0.05) \\&\approx 1 - P(P_1 > 0.05)P(P_2 > 0.05) \\&= 1 - (1 - 0.05)^2 = 0.0975. \quad (1)\end{aligned}$$

# Introduction to Efficacy Monitoring

- ▶ Situation can be worse if looks are not equally-spaced.
- ▶ Example: Suppose looks are after 10 and 10,000 observations.
  - ▶ The 2 p-values are nearly independent because the overlap is only 10 out of 10,000 people.
  - ▶ Independence is the worst case; the type 1 error rate is

$$\begin{aligned}P(\text{Reject } H_0 | H_0) &= P(P_1 \leq 0.05 \cup P_2 \leq 0.05) \\&= 1 - P(P_1 > 0.05 \cap P_2 > 0.05) \\&\approx 1 - P(P_1 > 0.05)P(P_2 > 0.05) \\&= 1 - (1 - 0.05)^2 = 0.0975.\end{aligned}\tag{1}$$



# Introduction to Efficacy Monitoring

**Table:** Type 1 error rate for unadjusted monitoring for 2-sided test at  $\alpha = 0.01$  or  $\alpha = 0.05$ . Note:  $2p \leq 0.05$  means 2-sided p-value  $\leq 0.05$ .

# Looks ( $k$ )	Reject $H_0$ if $2p \leq 0.01$		Reject $H_0$ if $2p \leq 0.05$	
	Equally Spaced	Worst Case	Equally Spaced	Worst Case
2	0.018	0.020	0.083	0.098
3	0.024	0.030	0.107	0.143
4	0.029	0.039	0.126	0.185
5	0.033	0.049	0.142	0.226
10	0.047	0.096	0.193	0.401
20	0.064	0.182	0.248	0.642
$\infty$	1	1	1	1

- ▶ This table applies to many different tests: t-test, test of proportions, logrank test, Cox model, etc.

# Introduction to Efficacy Monitoring

- ▶ Note that preceding table applies to 2-sided tests at  $\alpha = 0.01$  and  $\alpha = 0.05$ , but also applies to 1-sided tests at  $\alpha = 0.005$  and  $\alpha = 0.025$ .
- ▶ Because efficacy (upper) boundary could differ from “harm” (lower) boundary, we focus **for the rest of this lecture on 1-sided efficacy (upper) boundaries.**
- ▶ For symmetric, 2-sided z-score boundaries at level  $\alpha$ , use  $\pm c_i$ , where  $c_i$  is 1-sided boundary at level  $\alpha/2$ .
  - ▶ This is slightly conservative. Actual 2-sided error rate is infinitesimally less than  $\alpha$ .

# Table of Contents

Introduction to Efficacy Monitoring

Historical Boundaries

Haybittle-Peto

Pocock

O'Brien-Fleming

Unified Approach: Information, Z-Scores, B-Values

Alpha Spending Functions

Definition

O'Brien-Fleming-Like Spending Function

Pocock-Like Spending Function

Families of Spending Functions

Effect of Monitoring on Power

Case Study: CAST

Small Sample Sizes

Panoply of Problems Post-Monitoring

# Historical Boundaries: Haybittle-Peto

- ▶ The earliest boundary was the Haybittle-Peto boundary (Haybittle (1971)).
- ▶ Original suggestion used a very large z-statistic boundary (3) for interim looks, and 1.96 for final look.
- ▶ Haybittle-Peto was modified using the Bonferroni inequality:
  - ▶ Use p-value threshold 0.001 at interim looks.
  - ▶ Use p-value threshold  $\alpha - (k - 1)(0.001)$  at final look.
  - ▶ E.g., with 3 interim and 1 final analysis, reject at interim if  $p \leq 0.001$ , and at end if  $p \leq 0.025 - 3(0.001) = 0.022$ .
- ▶ By Bonferroni, the type 1 error rate,  $P(\text{reject } H_0 \text{ sometime})$ , is
$$\leq 0.001 + 0.001 + \dots + 0.001 + \alpha - (k - 1)(0.001) = \alpha.$$

# Historical Boundaries: Haybittle-Peto

- ▶ Desirable properties of Haybittle-Peto.
  - ▶ Simple to implement.
  - ▶ Can use regardless of timing of analyses.
  - ▶ Valid for **any** test statistic (don't need to know joint distribution of test statistic over time because Bonferroni inequality is used).
  - ▶ Final z-statistic boundary is close to what it would be with no monitoring (for a reasonable number of analyses).
- ▶ Undesirable property of Haybittle-Peto: Reversal of fortune
  - ▶ Z-statistic boundary drops drastically at the end, causing a logical inconsistency: Could be under boundary at penultimate look, see a partial reversal, and be over boundary at end. How could you be convinced now that you've seen a partial reversal?

# Historical Boundaries: Haybittle-Peto

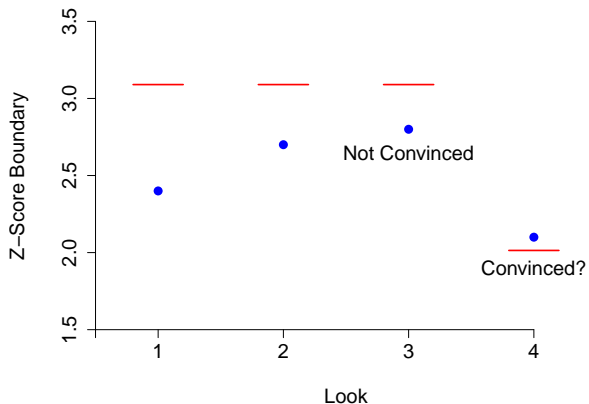


Figure: Reversal of fortune problem with Haybittle-Peto.

# Historical Boundaries: Pocock

- ▶ Pocock (1977) raised the z-statistic boundary by the same amount for each look.
- ▶ Pre-specify number of looks,  $k$ , and **assume they are equally spaced.**
- ▶ Use z-score boundary  $c = c(k)$  such that

$$P\left(\bigcup_{i=1}^k Z_i \geq c\right) = \alpha.$$

# Historical Boundaries: Pocock

Table: 1-sided z-score/p-value boundaries for Pocock procedure.

# Looks ( $k$ )	$\alpha = 0.005$	$\alpha = 0.025$	$\alpha = 0.05$
1	2.576 0.0050	1.960 0.0250	1.645 0.0500
2	2.772 0.0028	2.178 0.147	1.875 0.0304
3	2.873 0.0020	2.289 0.0110	1.992 0.0232
4	2.939 0.0016	2.361 0.0091	2.067 0.0194
5	2.986 0.0014	2.413 0.0079	2.122 0.0169
6	3.023 0.0013	2.453 0.0071	2.164 0.0152
7	3.053 0.0011	2.485 0.0065	2.197 0.0140
8	3.078 0.0010	2.512 0.0060	2.225 0.0130
9	3.099 0.0010	2.535 0.0056	2.249 0.0123
10	3.117 0.0009	2.555 0.0053	2.270 0.0116



# Historical Boundaries: Pocock

- ▶ Problem with Pocock: Z-statistic boundary at end is too high (equivalently, p-value boundary is too low).
- ▶ Example: for  $k = 5$  looks, 1-sided 0.025 final boundary for z-statistic (p-value) is 2.413 (0.0079).
- ▶ **Causes loss of power, requiring larger sample sizes.**
- ▶ Also practical reasons for wanting high early boundaries and lower late boundaries. Early in trial, staff may not understand protocol.
- ▶ Pocock recommends against his own procedure.

## Historical Boundaries: O'Brien-Fleming

- ▶ Haybittle-Peto had a desirable final z-statistic boundary, but dropped so abruptly that it allowed a logical inconsistency.
- ▶ **Assume looks are equally-spaced.**
- ▶ What is the steepest descending boundary that avoids the logical inconsistency? Answer: O'Brien and Fleming (1979) boundary.
- ▶ Z-statistic boundary at look  $i$  is proportional to  $1/\sqrt{i}$ .
- ▶ Z-statistic boundaries at looks  $1, 2, \dots, k$  are

$$\frac{a}{\sqrt{1}} = a, \frac{a}{\sqrt{2}}, \dots, \frac{a}{\sqrt{k}},$$

where  $a = a(k)$  is a constant making the type 1 error rate  $\alpha$ .

# Historical Boundaries: O'Brien-Fleming

Table: 1-sided O'Brien-Fleming z-score/p-value boundaries for  $\alpha = 0.025$ .

$k$	1	2	3	4	5	6	7	8	9	10
1	1.960 0.0250									
2	2.796 0.0026	1.977 0.0240								
3	3.471 0.0003	2.454 0.0071	2.004 0.0225							
4	4.048 $2.6 \times 10^{-5}$	2.862 0.0021	2.337 0.0097	2.024 0.0215						
5	4.562 $2.5 \times 10^{-6}$	3.226 0.0006	2.634 0.0042	2.281 0.0113	2.040 0.0207					
6	5.029 $2.5 \times 10^{-7}$	3.556 0.0002	2.903 0.0018	2.514 0.0060	2.249 0.0123	2.053 0.0200				
7	5.458 $2.4 \times 10^{-8}$	3.860 0.0001	3.151 0.0008	2.729 0.0032	2.441 0.0073	2.228 0.0129	2.063 0.0196			
8	5.861 $2.3 \times 10^{-9}$	4.144 $1.7 \times 10^{-5}$	3.384 0.0004	2.930 0.0017	2.621 0.0044	2.393 0.0084	2.215 0.0134	2.072 0.0191		
9	6.240 $2.2 \times 10^{-10}$	4.412 $5.1 \times 10^{-6}$	3.603 0.0002	3.120 0.0009	2.791 0.0026	2.547 0.0054	2.358 0.0092	2.206 0.0137	2.080 0.0188	
10	6.600 $2.1 \times 10^{-11}$	4.667 $1.5 \times 10^{-6}$	3.810 0.0001	3.300 0.0005	2.951 0.0016	2.694 0.0035	2.494 0.0063	2.333 0.0098	2.200 0.0139	2.087 0.0184

# Historical Boundaries

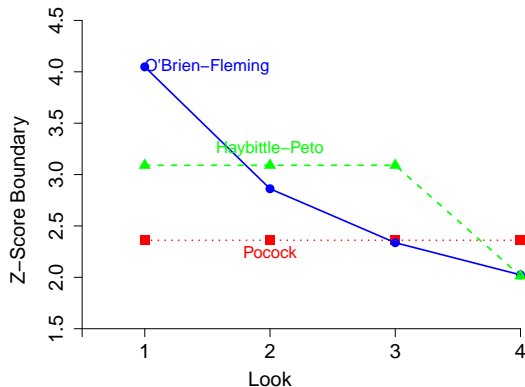


Figure: Haybittle-Peto, Pocock, and O'Brien-Fleming boundaries for 4 equally-spaced looks.

# Table of Contents

Introduction to Efficacy Monitoring

Historical Boundaries

Haybittle-Peto

Pocock

O'Brien-Fleming

**Unified Approach: Information, Z-Scores, B-Values**

Alpha Spending Functions

Definition

O'Brien-Fleming-Like Spending Function

Pocock-Like Spending Function

Families of Spending Functions

Effect of Monitoring on Power

Case Study: CAST

Small Sample Sizes

Panoply of Problems Post-Monitoring

# Unified Approach: Information, Z-Scores, B-Values

- ▶ We want to unify monitoring so same boundaries apply to many different testing settings.
- ▶ We will see that for large sample sizes, joint distribution of test statistics over time is same for different tests.
  - ▶ Lan and Zucker (1993).
  - ▶ Proschan et al. (2006).
  - ▶ Jennison and Turnbull (1999)
- ▶ First step: Think about simple setting of iid  $N(\delta, 1)$  data. For  $n = 1, \dots, N$ ,
  - ▶ Estimator is  $\hat{\delta}_n = \bar{Y}_n = S_n/n$ .
  - ▶ Sample size  $n$  measures amount of information contained in  $\hat{\delta}_n$ .
    - ▶ Note:  $\text{var}(\hat{\delta}_n) = 1/n$ , so  $n = 1/\text{var}(\hat{\delta}_n)$  is information in  $\hat{\delta}_n$ .
  - ▶ Fraction of information at interim analysis,  $t = n/N$ , is called **information time** or **information fraction**.

# Unified Approach: Information, Z-Scores, B-Values

- ▶  $Z(t) = \frac{S_n}{\sqrt{n}}$ . Note that

$$\begin{aligned} E\{Z(t)\} &= E\left(\frac{S_n}{\sqrt{n}}\right) = \frac{n\delta}{\sqrt{n}} = \sqrt{n}\delta \\ &= (\sqrt{N}\delta) \sqrt{\frac{n}{N}} = \theta\sqrt{t}, \quad \theta = E\{Z(1)\}. \end{aligned} \quad (2)$$

- ▶ Can also find variances and covariances.  $Z(t)$  process is multivariate normal with
  - ▶  $E\{Z(t)\} = \theta\sqrt{t}$ , where  $\theta = E\{Z(1)\}$ .
  - ▶  $\text{var}\{Z(t)\} = 1$ .
  - ▶  $\text{cov}\{Z(s), Z(t)\} = \sqrt{s/t}$ ,  $s \leq t$ .
  - ▶ Higher correlation the closer  $s$  is to  $t$ .

# Unified Approach: Information, Z-Scores, B-Values

- ▶ Let  $B(t) = \sqrt{t}Z(t) = \sqrt{\frac{n}{N}} \left( \frac{S_n}{\sqrt{n}} \right) = \frac{S_n}{\sqrt{N}}$ .
- ▶  $B(t)$  is proportional to a sum of iid  $N(\delta, 1)$  observations, where the proportionality constant makes  $B(1) = Z(1)$ , (z-score at end).
- ▶  $E\{B(t)\} = \sqrt{t}E\{Z(t)\} = \sqrt{t}\theta\sqrt{t} = \theta t$ .  $\theta = E\{Z(1)\}$ .
- ▶ Sums of iid components have **independent increments**. For example, if  $i_1 < i_2$ ,  $S_{i_2} - S_{i_1} = \sum_{j=i_1+1}^{i_2} Y_j$  is independent of  $S_{i_1}$  because  $S_{i_1}$  involves only the first  $i_1$  components and  $S_{i_2} - S_{i_1}$  involves only the subsequent  $i_2 - i_1$  observations.
- ▶ Equivalently, if  $t_1 < t_2$ ,  $B(t_2) - B(t_1)$  is independent of  $B(t_1)$ . More generally,  
 $B(t_1), B(t_2) - B(t_1), \dots, B(t_k) - B(t_{k-1})$  are independent.



# Unified Approach: Information, Z-Scores, B-Values

- ▶  $B(t)$  has the following properties:
  - ▶ The joint distribution of  $B(t_1), \dots, B(t_k)$  is multivariate normal.
  - ▶  $E\{B(t)\} = \theta t$ ,  $\theta = E\{Z(1)\}$ .
  - ▶  $\text{var}\{B(t)\} = t$ .
  - ▶  $B(t)$  has independent increments. Consequently, for  $s \leq t$ ,  
 $\text{cov}\{B(s), B(t)\} = s$ .
- ▶ A stochastic process with these properties is called a ***Brownian motion with drift  $\theta$*** .
- ▶ Brownian motion has very jagged paths that are continuous everywhere but differentiable nowhere!

# Unified Approach: Information, Z-Scores, B-Values

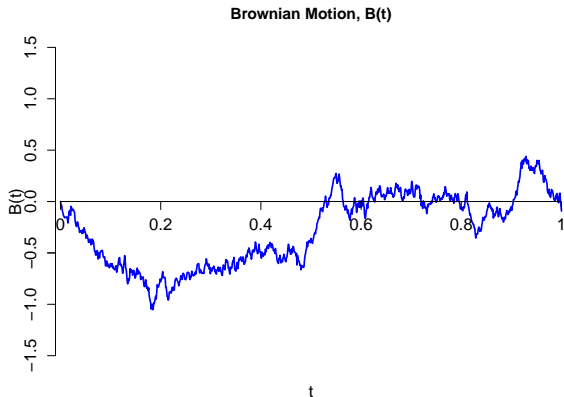


Figure: Brownian motion  $B(t)$  with drift 0.  $t$  is information time.

# Unified Approach: Information, Z-Scores, B-Values

Table: **Properties of  $Z(t)$  and  $B(t)$  processes, where  $\theta = E\{Z(1)\} = E\{B(1)\}$ . See Appendix for details.**

	<b><math>Z(t)</math></b>	<b><math>B(t)</math></b>
<b><math>E\{Z(t)\}</math> or <math>E\{B(t)\}</math></b>	<b><math>\theta \sqrt{t}</math></b>	<b><math>\theta t</math></b>
<b><math>\text{Var}\{Z(t)\}</math> or <math>\text{var}\{B(t)\}</math></b>	<b>1</b>	<b><math>t</math></b>
<b><math>\text{Cov}\{Z(s), Z(t)\}</math> or <math>\text{Cov}\{B(s), B(t)\}</math>, <math>s \leq t</math></b>	<b><math>\sqrt{\frac{s}{t}}</math></b>	<b><math>s</math></b>
<b>Independent increments?</b>	<b>No</b>	<b>Yes</b>

- ▶ We can monitor with  $B(t)$  or  $Z(t)$ , but calculations of probabilities are easier with  $B(t)$  because of independent increments and fact that mean of  $B(t)$  is linear function of  $t$ .

# Unified Approach: Information, Z-Scores, B-Values

Table: **Properties of  $Z(t)$  and  $B(t)$  processes, where  $\theta = E\{Z(1)\} = E\{B(1)\}$ . See Appendix for details.**

	<b><math>Z(t)</math></b>	<b><math>B(t)</math></b>
<b><math>E\{Z(t)\}</math> or <math>E\{B(t)\}</math></b>	<b><math>\theta \sqrt{t}</math></b>	<b><math>\theta t</math></b>
<b><math>\text{Var}\{Z(t)\}</math> or <math>\text{var}\{B(t)\}</math></b>	<b>1</b>	<b><math>t</math></b>
<b><math>\text{Cov}\{Z(s), Z(t)\}</math> or <math>\text{Cov}\{B(s), B(t)\}</math>, <math>s \leq t</math></b>	<b><math>\sqrt{\frac{s}{t}}</math></b>	<b><math>s</math></b>
<b>Independent increments?</b>	<b>No</b>	<b>Yes</b>

- ▶ We can monitor with  $B(t)$  or  $Z(t)$ , but calculations of probabilities are easier with  $B(t)$  because of independent increments and fact that mean of  $B(t)$  is linear function of  $t$ .

# Unified Approach: Information, Z-Scores, B-Values

Table: **Properties of  $Z(t)$  and  $B(t)$  processes, where  $\theta = E\{Z(1)\} = E\{B(1)\}$ . See Appendix for details.**

	<b><math>Z(t)</math></b>	<b><math>B(t)</math></b>
<b><math>E\{Z(t)\}</math> or <math>E\{B(t)\}</math></b>	<b><math>\theta \sqrt{t}</math></b>	<b><math>\theta t</math></b>
<b><math>\text{Var}\{Z(t)\}</math> or <math>\text{var}\{B(t)\}</math></b>	<b>1</b>	<b><math>t</math></b>
<b><math>\text{Cov}\{Z(s), Z(t)\}</math> or <math>\text{Cov}\{B(s), B(t)\}</math>, <math>s \leq t</math></b>	<b><math>\sqrt{\frac{s}{t}}</math></b>	<b><math>s</math></b>
<b>Independent increments?</b>	<b>No</b>	<b>Yes</b>

- ▶ We can monitor with  $B(t)$  or  $Z(t)$ , but calculations of probabilities are easier with  $B(t)$  because of independent increments and fact that mean of  $B(t)$  is linear function of  $t$ .

# Unified Approach: Information, Z-Scores, B-Values

- ▶ Key Fact: The same joint distribution of tests statistics holds for many different test statistics, provided we define information time correctly
  - ▶ One- and two-sample t-tests.
  - ▶ One and two-sample z-tests of proportions.
  - ▶ The logrank test and Cox model.
  - ▶ Large sample tests using an MLE.
  - ▶ Tests based on a complete, sufficient statistic.
  - ▶ Many more.
- ▶ **Information time is  $n/N$ , except for survival, when it is  $d/D$ , where  $n$  and  $N$  are interim and final sample sizes and  $d$  and  $D$  are interim and final numbers of people with events.**
- ▶ See Lan and Zucker (1993); chapter 2 of Proschan et al. (2006); chapters 3 and 11 of Jennison and Turnbull (1999).

# Unified Approach: Information, Z-Scores, B-Values

- ▶ Key idea is that many estimators  $\hat{\delta}$  behave just like a mean of *some number,  $I$* , of iid  $N(\delta, 1)$  observations.
- ▶ Just as  $\hat{\delta}$  behaves like a mean,  $I\hat{\delta}$  behaves like a sum.
- ▶ We just have to define  $I$  (information) appropriately;  $I = 1/\text{var}(\hat{\delta})$ .
- ▶ What really matters is information fraction,  $t = I/I_{\text{end}}$ , where  $I$  and  $I_{\text{end}}$  are the information at interim analysis and end of trial.
- ▶ For most estimators, information is proportional to sample size, so  $t = n/N$ , where  $n$  and  $N$  are interim and final sample sizes.
- ▶ In survival, information is proportional to # events, so  $t = d/D$ , where  $d$  and  $D$  are interim and final # people with events.

# Another Interpretation of Brien-Fleming Procedure

- ▶ B-values help us understand another motivation for O'Brien-Fleming.
- ▶ Pocock is a constant boundary for  $Z(t)$ .
- ▶ O'Brien-Fleming is a constant boundary for  $B(t)$ .

$$B(t_i) \geq c \Leftrightarrow Z(i/k) \geq \frac{c}{\sqrt{i/k}} \Leftrightarrow Z(i/k) \geq \frac{a}{\sqrt{i}}, \quad a = c\sqrt{k}.$$



# Table of Contents

Introduction to Efficacy Monitoring

Historical Boundaries

Haybittle-Peto

Pocock

O'Brien-Fleming

Unified Approach: Information, Z-Scores, B-Values

**Alpha Spending Functions**

Definition

O'Brien-Fleming-Like Spending Function

Pocock-Like Spending Function

Families of Spending Functions

Effect of Monitoring on Power

Case Study: CAST

Small Sample Sizes

Panoply of Problems Post-Monitoring

# Alpha Spending Functions

- ▶ For any z-score boundary  $c_1, \dots, c_k$ , we can compute the cumulative crossing probabilities by different times.
- ▶ Likewise, if we know cumulative crossing probabilities, we can re-construct boundaries.
- ▶ Lan and DeMets (1983): Instead of specifying boundaries, specify an **alpha spending function**  $\alpha^*(t)$  giving cumulative alpha spent by information time  $t$ .
  - ▶  $\alpha^*(t) \uparrow$  as  $t \uparrow$ .
  - ▶  $\alpha^*(0) = 0, \alpha^*(1) = \alpha$ .
- ▶ Then use information times to construct boundaries.

# Alpha Spending Functions

- ▶ Properties depend on which spending function we choose.

- ▶ **Most popular spending function for a 1-sided test at level  $\alpha$ :**

$$\alpha^*(t) = 2 \left\{ 1 - \Phi \left( \frac{z_{\alpha/2}}{\sqrt{t}} \right) \right\}, \quad (3)$$

where  $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ .

- ▶ For 1-sided,  $\alpha = 0.025$ ,

$$\alpha^*(t) = 2 \left\{ 1 - \Phi \left( \frac{2.2414}{\sqrt{t}} \right) \right\}. \quad (4)$$

- ▶ Generates boundaries similar to O'Brien-Fleming's for equally-spaced  $t$ , but **spending function approach does not require equal spacing**

# Alpha Spending Functions

- ▶ We illustrate boundary construction using this spending function and 3 looks for a survival trial planned for 200 deaths by end.
- ▶ First look occurs at 58th death,  $t = 58/200 = 0.29$ .
- ▶ Cumulative alpha to spend by  $t = 0.29$  is

$$\alpha^*(0.29) = 2 \left\{ 1 - \Phi \left( \frac{2.2414}{\sqrt{0.29}} \right) \right\} = 3.15 \times 10^{-5}. \quad (5)$$

- ▶ Need to find  $c_1$  such that  $P\{Z(0.29) \geq c_1\} = 3.15 \times 10^{-5}$ .
  - ▶  $c_1 = z_{3.15 \times 10^{-5}} = \Phi^{-1}(1 - 3.15 \times 10^{-5}) = 4.0011$ .
  - ▶ In R, use `qnorm(1-0.0000315)`.

# Alpha Spending Functions

- ▶ We illustrate boundary construction using this spending function and 3 looks for a survival trial planned for 200 deaths by end.
- ▶ First look occurs at 58th death,  $t = 58/200 = 0.29$ .
- ▶ Cumulative alpha to spend by  $t = 0.29$  is

$$\alpha^*(0.29) = 2 \left\{ 1 - \Phi \left( \frac{2.2414}{\sqrt{0.29}} \right) \right\} = 3.15 \times 10^{-5}. \quad (5)$$

- ▶ Need to find  $c_1$  such that  $P\{Z(0.29) \geq c_1\} = 3.15 \times 10^{-5}$ .
  - ▶  $c_1 = z_{3.15 \times 10^{-5}} = \Phi^{-1}(1 - 3.15 \times 10^{-5}) = 4.0011$ .
  - ▶ In R, use `qnorm(1-0.0000315)`.

# Alpha Spending Functions

- ▶ We illustrate boundary construction using this spending function and 3 looks for a survival trial planned for 200 deaths by end.
- ▶ First look occurs at 58th death,  $t = 58/200 = 0.29$ .
- ▶ Cumulative alpha to spend by  $t = 0.29$  is

$$\alpha^*(0.29) = 2 \left\{ 1 - \Phi \left( \frac{2.2414}{\sqrt{0.29}} \right) \right\} = 3.15 \times 10^{-5}. \quad (5)$$

- ▶ Need to find  $c_1$  such that  $P\{Z(0.29) \geq c_1\} = 3.15 \times 10^{-5}$ .
  - ▶  $c_1 = z_{3.15 \times 10^{-5}} = \Phi^{-1}(1 - 3.15 \times 10^{-5}) = 4.0011$ .
  - ▶ In R, use `qnorm(1-0.0000315)`.

# Alpha Spending Functions

- ▶ We illustrate boundary construction using this spending function and 3 looks for a survival trial planned for 200 deaths by end.
- ▶ First look occurs at 58th death,  $t = 58/200 = 0.29$ .
- ▶ Cumulative alpha to spend by  $t = 0.29$  is

$$\alpha^*(0.29) = 2 \left\{ 1 - \Phi \left( \frac{2.2414}{\sqrt{0.29}} \right) \right\} = 3.15 \times 10^{-5}. \quad (5)$$

- ▶ Need to find  $c_1$  such that  $P\{Z(0.29) \geq c_1\} = 3.15 \times 10^{-5}$ .
  - ▶  $c_1 = z_{3.15 \times 10^{-5}} = \Phi^{-1}(1 - 3.15 \times 10^{-5}) = 4.0011$ .
  - ▶ In R, use `qnorm(1-0.0000315)`.

# Alpha Spending Functions

- ▶ Don't cross boundary and 2nd look is at 110th death,  
 $t = 110/200 = 0.55$ .
- ▶ Cumulative alpha to spend by  $t = 0.55$  is

$$\alpha^*(0.55) = 2 \left\{ 1 - \Phi \left( \frac{2.2414}{\sqrt{0.55}} \right) \right\} = 0.0025. \quad (6)$$

- ▶ Need to find  $c_2$  such that

$$P\{Z(0.29) \geq 4.0011 \cup Z(0.55) \geq c_2\} = 0.0025.$$

- ▶ Not so easy to do! Use simulation or numerical integration.
- ▶ Lan/DeMets free software at University of Wisconsin.



# Alpha Spending Functions

- ▶ Don't cross boundary and 2nd look is at 110th death,  
 $t = 110/200 = 0.55$ .
- ▶ Cumulative alpha to spend by  $t = 0.55$  is

$$\alpha^*(0.55) = 2 \left\{ 1 - \Phi \left( \frac{2.2414}{\sqrt{0.55}} \right) \right\} = 0.0025. \quad (6)$$

- ▶ Need to find  $c_2$  such that

$$P\{Z(0.29) \geq 4.0011 \cup Z(0.55) \geq c_2\} = 0.0025.$$

- ▶ Not so easy to do! Use simulation or numerical integration.
- ▶ Lan/DeMets free software at University of Wisconsin.

# Alpha Spending Functions

- ▶ Don't cross boundary and 2nd look is at 110th death,  
 $t = 110/200 = 0.55$ .
- ▶ Cumulative alpha to spend by  $t = 0.55$  is

$$\alpha^*(0.55) = 2 \left\{ 1 - \Phi \left( \frac{2.2414}{\sqrt{0.55}} \right) \right\} = 0.0025. \quad (6)$$

- ▶ Need to find  $c_2$  such that

$$P\{Z(0.29) \geq 4.0011 \cup Z(0.55) \geq c_2\} = 0.0025.$$

- ▶ Not so easy to do! Use simulation or numerical integration.
- ▶ Lan/DeMets free software at University of Wisconsin.

# Alpha Spending Functions

- ▶ Don't cross boundary and 2nd look is at 110th death,  
 $t = 110/200 = 0.55$ .
- ▶ Cumulative alpha to spend by  $t = 0.55$  is

$$\alpha^*(0.55) = 2 \left\{ 1 - \Phi \left( \frac{2.2414}{\sqrt{0.55}} \right) \right\} = 0.0025. \quad (6)$$

- ▶ Need to find  $c_2$  such that

$$P\{Z(0.29) \geq 4.0011 \cup Z(0.55) \geq c_2\} = 0.0025.$$

- ▶ Not so easy to do! Use simulation or numerical integration.
- ▶ Lan/DeMets free software at University of Wisconsin.

# Alpha Spending Functions

Lan: DeMets Group Sequential Calculations

File **Compute** Help

No Computation Selected

### Analysis Parameters

Interim Analyses (k):  ( $1 < k \leq 25$ )

Information times( $t$ ):  ( $0 < t \leq 1$ )

Test Boundaries:

Calculate

Time	Lower Bound	Upper Bound		
1	0.20			
2	0.40			
3	0.60			
4	0.80			
5	1.00			

This is the statusbar



# Alpha Spending Functions

Lan-DeMets Group Sequential Calculations

File Compute Help

No Computation Selected

Bounds Alt-b  
Drift Alt-d  
Probability Alt-p  
Confidence Alt-i

(0 <  $\alpha$  ≤ 1)

Time	Lower Bound	Upper Bound		
1	0.20			
2	0.40			
3	0.60			
4	0.80			
5	1.00			

Calculate

# Alpha Spending Functions

Lan-DeMets Group Sequential Calculations

File Compute Help

## Compute Bounds

### Analysis Parameters

Interim Analyses (k):  (k ≤ 25)

Information times (t):  (0 < t ≤ 1)

Test Boundaries:

### Z-Score

Observed Z?

### Spending Function

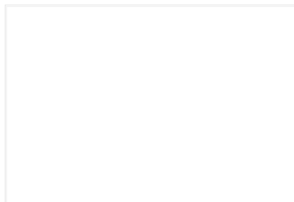
Overall Alpha:  (0 < α ≤ 1.0)

Function:

Truncate bounds?

	Time	Lower Bound	Upper Bound	Nominal Upr Alpha	Cum Alpha
1	0.20				
2	0.40				
3	0.60				
4	0.80				
5	1.00				

Calculate



# Alpha Spending Functions

Lan-DeMets Group Sequential Calculations

File Compute Help

## Compute Bounds

### Analysis Parameters

Interim Analyses (k): 3 (0 < k ≤ 25)  
Information times (t): Equally Spaced (0 < t ≤ 1)  
Test Boundaries: Two-Sided Symmetric

### Z-Score

Observed Z? No

### Spending Function

Overall Alpha: 0.05 (0 < α ≤ 1.0)  
Function: Blin-Rimm  
Truncate bounds? No

Calculate

	Time	Lower Bound	Upper Bound	Nominal Upr Alpha	Cum Alpha
1	0.33				
2	0.67				
3	1.00				

# Alpha Spending Functions

Lan-DeMets Group Sequential Calculations

File Compute Help

## Compute Bounds

### Analysis Parameters

Interim Analyses (k):  ( $1 < k \leq 25$ )

Information times (t):  ( $0 < t \leq 1$ )

Test Boundaries:

### Z-Score

Observed Z?

### Spending Function

Overall Alpha:  ( $0 < \alpha \leq 1.0$ )

Function:

Truncate bounds?

Calculate

Time	Lower Bound	Upper Bound	Nominal Upr Alpha	Cum Alpha
1				
2				
3				



# Alpha Spending Functions

Lan-DeMets Group Sequential Calculations

File Compute Help

## Compute Bounds

### Analysis Parameters

Interim Analyses (k):  ( $1 < k \leq 25$ )  
Information times (t):  ( $0 < t \leq 1$ )  
Test Boundaries:

### Z-Score

Observed Z?

### Spending Function

Overall Alpha:  ( $0 < \alpha \leq 1.0$ )  
Function:   
Truncate bounds?

	Time	Lower Bound	Upper Bound	Nominal Upr Alpha	Cum Alpha
1	0.29				
2	0.55				
3	1.00				

Calculate

# Alpha Spending Functions

Lan-DeMets Group Sequential Calculations

File Compute Help

## Compute Bounds

### Analysis Parameters

Interim Analysis (k):  (1 < k ≤ 25)

Information Times (t):  (0 < t ≤ 1)

Test Boundaries:

### Z-Score

Observed Z?

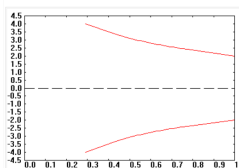
### Spending Function

Overall Alpha:  (0 < α ≤ 1.0)

Function:

Truncate bounds?

Calculate



Time	Lower Bound	Upper Bound	Nominal Upr Alpha	Cum Alpha	
1	0.29	-4.0011	-4.0011	0.00003	0.00006
2	0.55	-2.8074	2.8074	0.00250	0.00502
3	1.00	-1.9740	1.9740	0.02419	0.05000

# Alpha Spending Functions

- ▶ Free software says  $c_2 = 2.8074$ .
- ▶ Last look is at 200th death,  $t = 200/200 = 1$ .
- ▶ Cumulative alpha to spend by  $t = 1$  is

$$\alpha^*(1) = 2 \left\{ 1 - \Phi \left( \frac{2.2414}{\sqrt{1}} \right) \right\} = 0.025. \quad (7)$$

- ▶ Need to find  $c_3$  such that

$$P[\{Z(0.29) \geq 4.0011\} \cup \{Z(0.55) \geq 2.8074\} \cup \{Z(1) \geq c_3\}] = 0.025.$$

- ▶ Free software says  $c = 1.9740$ .

# Alpha Spending Functions

- ▶ Free software says  $c_2 = 2.8074$ .
- ▶ Last look is at 200th death,  $t = 200/200 = 1$ .
- ▶ Cumulative alpha to spend by  $t = 1$  is

$$\alpha^*(1) = 2 \left\{ 1 - \Phi \left( \frac{2.2414}{\sqrt{1}} \right) \right\} = 0.025. \quad (7)$$

- ▶ Need to find  $c_3$  such that

$$P[\{Z(0.29) \geq 4.0011\} \cup \{Z(0.55) \geq 2.8074\} \cup \{Z(1) \geq c_3\}] = 0.025.$$

- ▶ Free software says  $c = 1.9740$ .

# Alpha Spending Functions

- ▶ Free software says  $c_2 = 2.8074$ .
- ▶ Last look is at 200th death,  $t = 200/200 = 1$ .
- ▶ Cumulative alpha to spend by  $t = 1$  is

$$\alpha^*(1) = 2 \left\{ 1 - \Phi \left( \frac{2.2414}{\sqrt{1}} \right) \right\} = 0.025. \quad (7)$$

- ▶ Need to find  $c_3$  such that

$$P[\{Z(0.29) \geq 4.0011\} \cup \{Z(0.55) \geq 2.8074\} \cup \{Z(1) \geq c_3\}] = 0.025.$$

- ▶ Free software says  $c = 1.9740$ .

# Alpha Spending Functions

## Lan-DeMets Group Sequential Boundaries Calculations

### Compute Spending

Analysis Parameters

Interim Analyses (k): 3

Information times(t): User Input

Test Boundaries: Two-Sided Symmetric

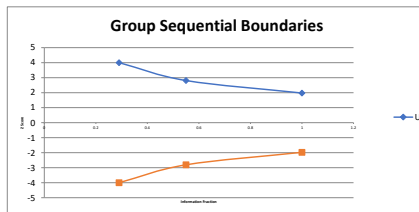
Spending Function

Overall Significance Level: 0.05

Spending Function: O'Brien-Fleming

Truncate Bounds? No

Time	Lower	Upper	Nominal Upr Alpha	Cum Alpha
0.29	-4.0011	4.0011	0.00003	0.00006
0.55	-2.8074	2.8074	0.00250	0.00502
1.00	-1.9740	1.9740	0.02419	0.05000



# Alpha Spending Functions

- ▶ Note: An equivalent way to compute  $c_i$  is to make

$$P(Z(t_1) < c_1 \cap \dots \cap Z(t_{i-1}) < c_{i-1} \cap Z(t_i) \geq c_i) = \alpha^*(t_i) - \alpha^*(t_{i-1}).$$

- ▶ That is, probability of **first** crossing boundary at time  $t_i$  is  $\alpha^*(t_i) - \alpha^*(t_{i-1})$ .

- ▶ Then cumulative probability of crossing by  $t_i$  is

$$\alpha^*(t_1) + \{\alpha^*(t_2) - \alpha^*(t_1)\} + \dots + \{\alpha^*(t_i) - \alpha^*(t_{i-1})\} = \alpha^*(t_i).$$

# Alpha Spending Functions

- ▶ **Big advantages of spending function approach:**
  - ▶ **Looks need not be equally-spaced.**
  - ▶ **Don't even have to pre-specify number of looks (but number and timing of looks assumed independent of data).**
- ▶ Nonetheless, pre-specification of number and timing of looks is advisable.



# Alpha Spending Functions

- ▶ Pocock-like spending function: Lan and DeMets noticed that amount spent by Pocock boundaries looked like a log function for a large number of looks.
- ▶ To get similar spending function:

$$\alpha^*(t) = \alpha \ln(a + bt),$$

$$\alpha^*(0) = 0 \Rightarrow a = 1$$

$$\alpha^*(1) = \alpha \Rightarrow b = e - 1$$

$$\alpha^*(t) = \alpha \ln\{1 + (e - 1)t\}. \quad (8)$$

(8) is Pocock-like spending function. Generates boundaries similar to those of Pocock's when looks are equally spaced.

# Alpha Spending Functions

- ▶ Pocock-like spending function: Lan and DeMets noticed that amount spent by Pocock boundaries looked like a log function for a large number of looks.
- ▶ To get similar spending function:

$$\alpha^*(t) = \alpha \ln(a + bt),$$

$$\alpha^*(0) = 0 \Rightarrow a = 1$$

$$\alpha^*(1) = \alpha \Rightarrow b = e - 1$$

$$\alpha^*(t) = \alpha \ln\{1 + (e - 1)t\}. \quad (8)$$

(8) is Pocock-like spending function. Generates boundaries similar to those of Pocock's when looks are equally spaced.

# Alpha Spending Functions

- ▶ Pocock-like spending function: Lan and DeMets noticed that amount spent by Pocock boundaries looked like a log function for a large number of looks.
- ▶ To get similar spending function:

$$\alpha^*(t) = \alpha \ln(a + bt),$$

$$\alpha^*(0) = 0 \Rightarrow a = 1$$

$$\alpha^*(1) = \alpha \Rightarrow b = e - 1$$

$$\alpha^*(t) = \alpha \ln\{1 + (e - 1)t\}. \quad (8)$$

(8) is Pocock-like spending function. Generates boundaries similar to those of Pocock's when looks are equally spaced.

# Alpha Spending Functions

- ▶ Pocock-like spending function: Lan and DeMets noticed that amount spent by Pocock boundaries looked like a log function for a large number of looks.
- ▶ To get similar spending function:

$$\alpha^*(t) = \alpha \ln(a + bt),$$

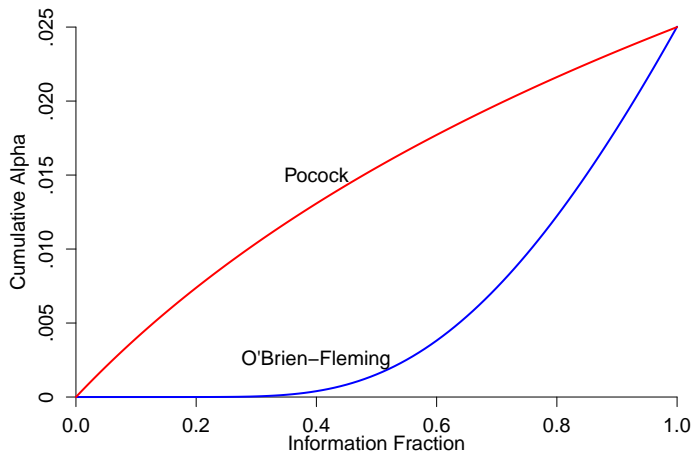
$$\alpha^*(0) = 0 \Rightarrow a = 1$$

$$\alpha^*(1) = \alpha \Rightarrow b = e - 1$$

$$\alpha^*(t) = \alpha \ln\{1 + (e - 1)t\}. \quad (8)$$

(8) is Pocock-like spending function. Generates boundaries similar to those of Pocock's when looks are equally spaced.

# Alpha Spending Functions



# Alpha Spending Functions

- ▶ O'Brien-Fleming-like spending function is convex and spends almost no  $\alpha$  early in trial, then rises quickly toward end (desirable).
- ▶ In contrast, Pocock-like spending function is concave and spends more aggressively early (undesirable).
- ▶ Consequence: O'Brien-Fleming-like spending function (good) creates high early boundaries and boundaries close to 1.96 at end, whereas Pocock (bad) has lower early boundaries but higher late boundaries.

# Alpha Spending Functions

- ▶ There are also families of spending functions like the power family:

$$\alpha^*(t) = \alpha t^\phi,$$

where small values of power parameter  $\phi$  spend  $\alpha$  aggressively early, whereas larger values spend alpha conservatively until close to  $t = 1$ .

- ▶ Another family is the Hwang, Shi-DeCani family:

$$\alpha^*(0) = 0, \quad \alpha^*(t) = \alpha \times \left\{ \frac{1 - \exp(-\gamma t)}{1 - \exp(-\gamma)} \right\}, \quad t > 0.$$

- ▶  $\gamma$  can be positive (aggressive early spending) or negative (conservative early spending).
- ▶  $\gamma = -4$  is like O'Brien-Fleming.

# Alpha Spending Functions

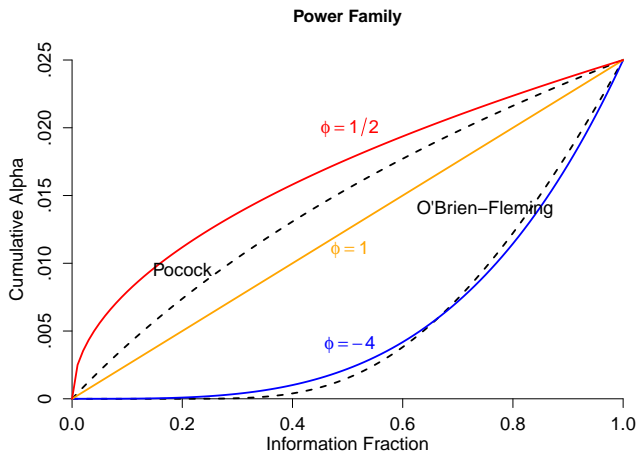


Figure: The power family  $\alpha^*(t) = 0.025t^\phi$  for different values of  $\phi$ .



# Alpha Spending Functions

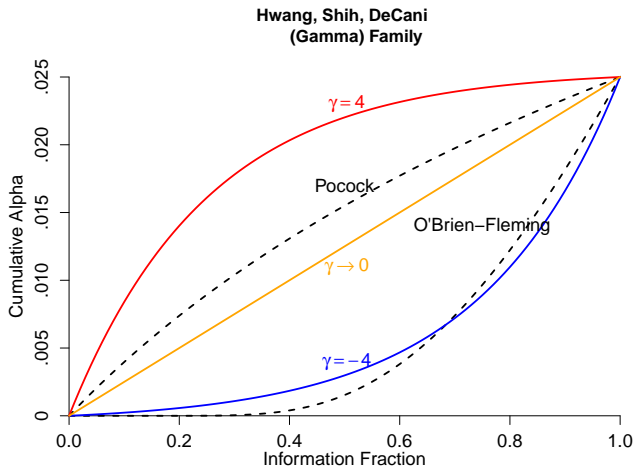


Figure: Wang, Shih, DeCani family of spending functions,  $\{1 - \exp(-\gamma t)\} / \{1 - \exp(-\gamma)\}$  for different values of  $\gamma$ .

# Table of Contents

Introduction to Efficacy Monitoring

Historical Boundaries

Haybittle-Peto

Pocock

O'Brien-Fleming

Unified Approach: Information, Z-Scores, B-Values

Alpha Spending Functions

Definition

O'Brien-Fleming-Like Spending Function

Pocock-Like Spending Function

Families of Spending Functions

**Effect of Monitoring on Power**

Case Study: CAST

Small Sample Sizes

Panoply of Problems Post-Monitoring

# Effect of Monitoring on Power

- ▶ Just like in non-monitoring setting, power in monitoring setting depends only on  $E\{Z(1)\} = \theta$ , the expected z-score at end (drift parameter of Brownian motion).
- ▶ Monitoring must incur a sample size penalty; always more powerful to spend all alpha at end (Neyman-Pearson lemma).
- ▶ The size of the penalty depends on the spending function
  - ▶ Pocock–big penalty
  - ▶ O'Brien-Fleming–small penalty.
- ▶ Why?

# Effect of Monitoring on Power

- ▶ Power for 1-sided test at level  $\alpha$  always satisfies following inequality, where  $Z$  is z-statistic at end of trial

$$P_{\theta}(Z \geq c_k) \leq \text{Power, monitoring} \leq P_{\theta}(Z \geq z_{\alpha})$$

$$P_{\theta}(Z \geq c_k) \leq \text{Power, monitoring} \leq \text{Power, no monitoring.}$$

where  $c_k$  is boundary at end and  $z_{\alpha}$  is the  $1 - \alpha$ th quantile of standard normal.

- ▶ For O'Brien-Fleming-like boundaries,  $c_k$  is close to  $z_{\alpha}$  and left side is close to power with no monitoring.

# Effect of Monitoring on Power

- ▶ We can compute sample size/power for monitoring using the EZ principle.
- ▶ Just like in non-monitoring setting, power depends on EZ, namely  $E\{Z(1)\}$  (the drift parameter,  $\theta$ ).
- ▶ Can use Wisconsin's WinLD program to compute drift parameter  $\theta$  for given power.
- ▶ Then equate  $E\{Z(1)\}$  to given value of drift and solve for  $N$ .

# Alpha Spending Functions

Lan-DeMets Group Sequential Calculations

File Compute Help

No Computation Selected

Bounds Alt-b  
Drift Alt-d  
Probability Alt-p  
Confidence Alt-i

( $0 < \tau \leq 1$ )

Calculate

Time	Lower Bound	Upper Bound		
1	0.20			
2	0.40			
3	0.60			
4	0.80			
5	1.00			

# Alpha Spending Functions

Lan-DeMets Group Sequential Calculations

File Compute Help

## Compute Drift

### Analysis Parameters

Interim Analyses (k)  ( $1 < k \leq 25$ )

Information times (t)  ( $0 < t \leq 1$ )

Test Boundaries:

### Power and Bounds Parameters

Determine Bounds:

Power:  ( $0 < (1 - \beta) \leq 1$ )

Drift

### Spending Function

Overall Alpha:  ( $0 < \alpha \leq 1.0$ )

Function:

Truncate bounds?

Calculate

	Time	Lower Bound	Upper Bound	Nominal Upr Alpha	Cum exit pr
1	0.20				
2	0.40				
3	0.60				
4	0.80				
5	1.00				

# Alpha Spending Functions

Lan-DeMets Group Sequential Calculations

File Compute Help

Compute Drift

### Analysis Parameters

Interim Analyses (k):  ( $1 < k \leq 25$ )  
Information times(t):  ( $0 < t \leq 1$ )  
Test Boundaries:

### Power and Bounds Parameters

Determine Bounds:   
Power:  ( $0 < (1 - \beta) \leq 1$ )

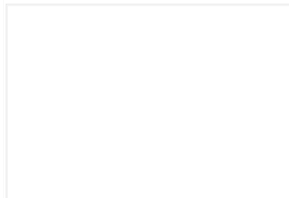
Drift:

### Spending Function

Overall Alpha:  ( $0 < \alpha \leq 1.0$ )  
Function:   
Truncate bounds?

Time	Lower Bound	Upper Bound	Nominal Upr Alpha	Cum exit pr
1	0.25			
2	0.50			
3	0.75			
4	1.00			

Calculate





# Alpha Spending Functions

Lan-DeMets Group Sequential Calculations

File Compute Help

## Compute Drift

### Analysis Parameters

Interim Analyses (k):  ( $1 < k \leq 25$ )  
Information times (t):  ( $0 < t \leq 1$ )  
Test Boundaries:

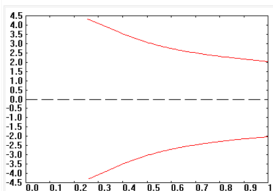
### Power and Bounds Parameters

Determine Bounds:   
Power:  ( $0 < (1 - \beta) \leq 1$ )

### Drift

### Spending Function

Overall Alpha:  ( $0 < \alpha \leq 1.0$ )  
Function:   
Truncate bounds?



Time	Lower Bound	Upper Bound	Nominal Upr Alpha	Cum exit pr	
1	0.25	-4.3326	4.3326	0.00001	0.00350
2	0.50	-2.9631	2.9631	0.00152	0.25788
3	0.75	-2.3590	2.3590	0.00916	0.68526
4	1.00	-2.0141	2.0141	0.02200	0.90000

# Effect of Monitoring on Power

- ▶ Use WinLD software at U. of Wisconsin.
- ▶ Click on “Compute” menu and click on “Drift”. Choose “Interim Analyses (k)”, select 4 and hit enter (must hit enter).
- ▶ Table at upper right shows 4 equally-spaced looks (to change to unequal spacings, use “Use Input” under “Information Times”).
- ▶ Choose power level (default is 0.90) under “Power and Bounds Parameters”.
- ▶ Choose spending function under “Spending Function” (default is O’Brien-Fleming).

## Effect of Monitoring on Power

- ▶ Click on the “Calculate” button.
- ▶ In lower left, under “Drift” is number 3.2711.
- ▶ Tells us that in sample size calculations, make expected z-score 3.2711 instead of  $1.96 + 1.28 = 3.24$ .
- ▶ Example: For a t-test, expected z-score at end is

$$\frac{\delta}{\sqrt{\frac{2\sigma^2}{N}}}$$

- ▶ Equate to 3.2711 and solve for  $N$ . Per-arm sample size is

$$N = \frac{2\sigma^2(3.2711)^2}{\delta^2} \text{ per arm.}$$

# Effect of Monitoring on Power

- ▶ If  $\delta = 5$  and  $\sigma = 14$ ,  $N \approx 168$  per arm.
- ▶ Compare to 165 per arm with no monitoring.
- ▶ Now choose “Pocock” spending function and hit “Calculate”.
  - ▶ New drift is 3.5177.
  - ▶ New per-arm sample size is

$$N = \frac{2(14)^2(3.5177)^2}{5^2} \approx 195.$$

- ▶ More substantial sample size penalty.

# Table of Contents

Introduction to Efficacy Monitoring

Historical Boundaries

Haybittle-Peto

Pocock

O'Brien-Fleming

Unified Approach: Information, Z-Scores, B-Values

Alpha Spending Functions

Definition

O'Brien-Fleming-Like Spending Function

Pocock-Like Spending Function

Families of Spending Functions

Effect of Monitoring on Power

**Case Study: CAST**

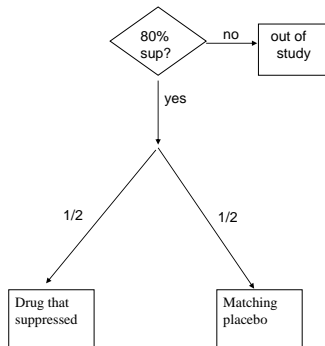
Small Sample Sizes

Panoply of Problems Post-Monitoring

# CAST

- ▶ Recall the Cardiac Arrhythmia Suppression Trial (CAST).
- ▶ Patients with prior heart attack and arrhythmias.
- ▶ Several papers showed arrhythmias are associated with increased risk of death after heart attack (see, e.g., CDP (1973), Ruberman et al. (1977)).
- ▶ Class I antiarrhythmic drugs encainide, flecainide, and moricizine were approved based on surrogate arrhythmia endpoint.
- ▶ Goal of CAST: Determine whether suppressing arrhythmias leads to fewer sudden deaths/cardiac arrests.
- ▶ CAST titrated encainide, flecainide, and moricizine until they found one that worked (did not care which drug used).
- ▶ Wanted to ensure arrhythmias could be suppressed, so patients whose arrhythmias not suppressed were not randomized.

## CAST DESIGN



1

Figure: CAST design.

- ▶ Some died during titration, but that's expected in sick population.
- ▶ Doctors already believed suppression hypothesis.
- ▶ In fact, some felt it was unethical to withhold treatment after finding a drug that suppressed arrhythmias.
  - ▶ Led to recruitment problems in CAST.
- ▶ So convinced were doctors that results could only be beneficial that CAST investigators proposed a 1-sided test at  $\alpha = 0.05$ .
- ▶ At first DSMB meeting 3/14/87, DSMB reviewed protocol and recommended using 1-sided  $\alpha = 0.025$  instead of 0.05 for efficacy. Asked for monitoring guideline in future.
- ▶ Next meeting January 1988: Board chose to be blinded.



- ▶ Next DSMB meeting: 9/16/88. DSMB discussed and approved monitoring plan.
- ▶ Plan used  $\alpha = 0.025$  for efficacy.
- ▶ DSMB decided to use symmetric lower 0.025 boundary for harm.
- ▶ Spending function spent  $\alpha = 0.0125$  linearly until just before end, then spent remaining 0.0125 at end:

$$\alpha^*(t) = \begin{cases} 0.0125t & \text{if } t < 1, \\ 0.025 & \text{if } t = 1. \end{cases} \quad (9)$$

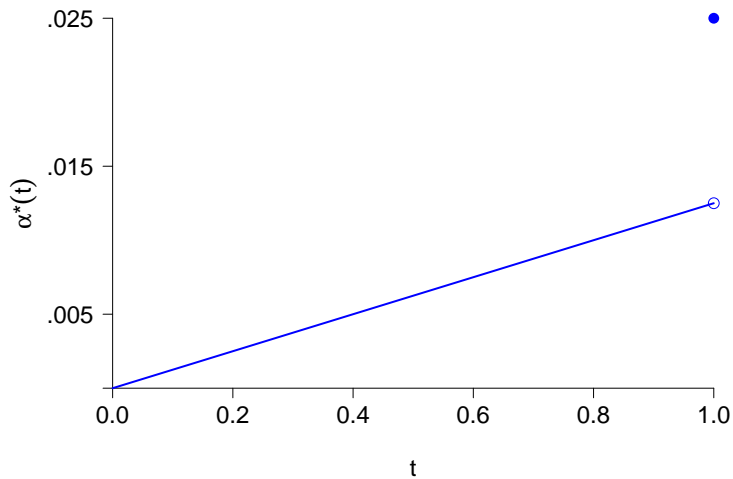


Figure: Spending function for efficacy used in CAST. Used symmetric lower boundary.

- ▶ After approving monitoring plan, the DSMB examined data.
- ▶ Proportion with sudden death/cardiac arrest:
  - ▶ Arm X: 3/576.
  - ▶ Arm Y: 19/571.
- ▶ Information fraction after 22 of 425 expected events:  
 $t = 22/425 = 0.05$ .
- ▶ No boundary was yet in place, but suppose it were.
  - ▶ Can spend  $\alpha^*(0.05) = 0.0125(0.05) = 0.000625$ . Boundary would have been  $-3.22$  and logrank z-score was  $-3.43$ .
  - ▶ Would have crossed harm boundary!
- ▶ Board decided no matter which arm was placebo, they wouldn't stop. Remained blinded.

- ▶ Next DSMB meeting: 4/16-4/17, 1989.
- ▶ There were now 48 sudden deaths/cardiac arrests, 35 of which were on arm Y!
- ▶ Now clear they overestimated final number events. Now expected to be 300, so  $t = 48/300 = 0.16$ .
- ▶ Boundary  $\pm 2.97$ .
- ▶ DSMB unblinded and discovered arm Y was active.
- ▶ Logrank z-score:  $Z(0.16) = -3.22$ . Harm boundary crossed!
- ▶ DSMB was shocked.

- ▶ Maybe pills were mis-labeled? No. In preparation for meeting, they analyzed pills and found no mis-labeling.
- ▶ Maybe randomization failed? No. Baseline characteristics were similar across arms.
- ▶ Maybe harm was confined to a certain subgroup? No. Results were consistent across subgroups and secondary endpoint of mortality.
- ▶ Maybe harm was confined to 1 or 2 drugs? It appeared that encainide and flecainide were the “bad actors” and moricizine was good.
- ▶ Decided to discontinue encainide and flecainide and continue CAST II with moricizine.

- ▶ DSMB chose to be blinded again. Saw results by Arm P versus Arm Q.
- ▶ Changed entry criteria to get sicker patients: Thought drugs should work in sicker patients.
- ▶ Decided to spend 0.05 for harm and 0.025 for benefit.
- ▶ Included 2-week placebo titration phase to see if too many patients were dying while titrating drugs.
- ▶ On July 31, 1991, there were 3 events on arm P and 15 on arm Q in 2-week titration phase.
- ▶ Arm Q was moricizine!
- ▶ CAST II ended.

- ▶ Lessons from CAST:
  - ▶ Can have benefit on surrogate (arrhythmias) and harm on endpoint of real interest (sudden death/cardiac arrest).
  - ▶ Harm can never be ruled out. Use 2-sided tests in clinical trials.
  - ▶ May make sense to use asymmetric boundaries for harm versus benefit.
  - ▶ Blinding DSMBs is ill-advised. Members think this makes them more objective, but the opposite is true.
    - ▶ Pre-conceived ideas enter maximally if blinded.
    - ▶ Why waste time pondering what you would do if results were in one direction or the other, when they are only in one direction?
- ▶ CAST references: CAST (1989) and CAST (1992).
- ▶ Moore (1995) gives a history of the suppression hypothesis and development and testing of antiarrhythmic drugs.

# Table of Contents

Introduction to Efficacy Monitoring

Historical Boundaries

Haybittle-Peto

Pocock

O'Brien-Fleming

Unified Approach: Information, Z-Scores, B-Values

Alpha Spending Functions

Definition

O'Brien-Fleming-Like Spending Function

Pocock-Like Spending Function

Families of Spending Functions

Effect of Monitoring on Power

Case Study: CAST

**Small Sample Sizes**

Panoply of Problems Post-Monitoring



## Small Sample Sizes

- ▶ Methods so far have considered large samples sizes.
- ▶ What can we do if sample sizes are small?
- ▶ One improvement for continuous outcomes: Apply p-value boundary to p-value computed using t-distribution.
- ▶ Example: Suppose you monitor continuous outcome using t-statistic with 4 equally-spaced looks and Pocock boundary for 1-sided  $\alpha = 0.025$ .
- ▶ At first look after 5 people per-arm, compute t-statistic

$$T = \frac{\bar{Y}_T - \bar{Y}_C}{\sqrt{2s^2/5}}$$

# Small Sample Sizes

- ▶ The z-score (1-sided p-value) boundary at each look is 2.361 (0.0091).
- ▶ **Bad idea: Apply z-score boundary to  $T$ .** Inaccurate because at 1st look,  $T$  has only 8 d.f., very different from  $N(0,1)$ .
- ▶ **Much better idea: Compute 1-sided p-value using t-distribution with 8 df. Then apply boundary 0.0091 to p-value.**
  - ▶ **Reject  $H_0$  at any analysis if  $p < 0.0091$ , where  $p$  is 1-sided p-value using t-distribution with the given df.**

# Small Sample Sizes

- ▶ Another approach with small sample sizes: Use permutation test.
- ▶ E.g, with spending function  $\alpha^*(t)$ , determine boundary  $c_i$  such that.

$$P_{\text{perm}}(Z(t_1) \geq c_1 \cup \dots \cup Z(t_i) \geq c_i) = \alpha^*(t_i),$$

where  $P_{\text{perm}}(A)$  denotes permutation probability of  $A$ :

$$P_{\text{perm}}(A) = \frac{\# \text{ permutations such that event } A \text{ occurs}}{\text{total } \# \text{ permutations}}.$$

# Table of Contents

Introduction to Efficacy Monitoring

Historical Boundaries

Haybittle-Peto

Pocock

O'Brien-Fleming

Unified Approach: Information, Z-Scores, B-Values

Alpha Spending Functions

Definition

O'Brien-Fleming-Like Spending Function

Pocock-Like Spending Function

Families of Spending Functions

Effect of Monitoring on Power

Case Study: CAST

Small Sample Sizes

Panoply of Problems Post-Monitoring

# Post-Monitoring Inference: Panoply of Problems

- ▶ Assume nuisance parameters are known.
- ▶ With no monitoring, z-statistic is sufficient statistic.
  - ▶ Inferences should be based solely on  $Z$ .
  - ▶ Likelihood ratio for testing  $H_0 : \theta = 0$  versus  $H_1 : \theta > 0$  is monotone.
    - ▶ Most powerful test against each simple alternative  $H_1 : \theta = \theta_1$  is same for all  $\theta_1 > 0$ .
    - ▶  $Z > z_\alpha$  is UMP level  $\alpha$  test for  $H_1 : \theta > 0$ .
  - ▶ MLE of  $\theta$  is  $Z$  and is unbiased.

# Post-Monitoring Inference: Panoply of Problems

- ▶ These are all problematic with monitoring.

- ▶ If  $\tau$  is info time when trial stopped, likelihood ratio is

$$L(\theta)/L(0) = \exp \left\{ \theta \sqrt{\tau} Z(\tau) - (\theta^2/2)\tau \right\}.$$

- ▶ Consequently, sufficient statistic is **pair**  $\{\tau, Z(\tau)\}$ , so inferences should be based solely on  $\{\tau, Z(\tau)\}$  or, equivalently,  $\{\tau, B(\tau)\}$ .
  - ▶ No monotone likelihood ratio, so most powerful test against  $H_1 : \theta = 1$  could be different from most powerful test against  $H_1 : \theta = 2$  (no UMP test against  $H_1 : \theta > 0$ ).
  - ▶ Must specify how to order sample space to compute p-value to test  $H_1 : \theta > 0$ .
  - ▶ MLE of  $\theta$ ,  $B(\tau)/\tau$ , is biased high.

# Post-Monitoring Inference: Panoply of Problems

- ▶ Regarding calculating 1-sided p-value adjusted for monitoring, consider all outcomes consistent with boundaries, and order them in some way:
  - ▶ **MLE ordering** orders by  $B(\tau)/\tau$ .  $\{\tau_2, Z(\tau_2)\}$  more extreme than  $\{\tau_1, Z(\tau_1)\}$  if  $B(\tau_2)/\tau_2 \geq B(\tau_1)/\tau_1$ .
  - ▶ **Z-score ordering** orders by  $Z(\tau)$ .  $\{\tau_2, Z(\tau_2)\}$  more extreme than  $\{\tau_1, Z(\tau_1)\}$  if  $Z(\tau_2) \geq Z(\tau_1)$ .
  - ▶ **B-value ordering** orders by  $B(\tau)$ .  $\{\tau_2, Z(\tau_2)\}$  more extreme than  $\{\tau_1, Z(\tau_1)\}$  if  $B(\tau_2) \geq B(\tau_1)$ .
  - ▶ **Stagewise ordering** orders first by  $\tau$ , then by  $Z(\tau)$ .  $\{\tau_2, Z(\tau_2)\}$  more extreme than  $\{\tau_1, Z(\tau_1)\}$  if  $\tau_2 < \tau_1$  or if  $\tau_2 = \tau_1$  and  $Z(\tau_2) \geq Z(\tau_1)$ .
    - ▶ My favorite because does not force us to consider future events to compute p-value.

# Post-Monitoring Inference: Panoply of Problems

- ▶ Regarding estimation, suppose I have 2 independent unbiased estimators,  $\hat{\delta}_1$  and  $\hat{\delta}_2$ , of treatment effect  $\delta$ .
- ▶ I peek at  $\hat{\delta}_1$ .
  - ▶ If  $\hat{\delta}_1$  is very large, I report only  $\hat{\delta}_1$ .
  - ▶ If  $\hat{\delta}_1$  is not large, I average it with  $\hat{\delta}_2$  and report the average.
- ▶ Intuitively clear that this overestimates  $\delta$ .
- ▶ That is monitoring! Suppose 1 interim analysis at halfway point.
  - ▶ If interim estimate is very large, stop trial and report  $\hat{\delta}_1$ .
  - ▶ If interim estimate is not large, continue to end, so final estimate is average,  $(\hat{\delta}_1 + \hat{\delta}_2)/2$ , of 1st and 2nd half estimates.



# Summary

- ▶ Unified monitoring:
  - ▶ Information  $I$  in treatment effect estimator  $\hat{\delta}$  is  $1/\text{var}(\hat{\delta})$ , and information time is  $t = I_{\text{current}}/I_{\text{final}}$ .
  - ▶  $B(t)$  and  $\hat{\delta}(t)$  behave like sum and sample mean of  $I$  iid  $N(\delta, 1)$  observations.
  - ▶  $t$  often reduces to ratio of current to final sample size (non-survival) or current to final number of people with events (survival).
  - ▶ Same boundaries apply to different test statistics.
- ▶ Can monitor using  $Z(t)$  or Brownian motion  $B(t)$ . Joint distribution over time is multivariate normal with:
  - ▶  $E\{Z(t)\} = \theta\sqrt{t}$ ,  $\text{cov}\{Z(s), Z(t)\} = \sqrt{s/t}$ ,  $s \leq t$ .
  - ▶  $E\{B(t)\} = \theta t$ ,  $\text{cov}\{B(s), B(t)\} = s$ ,  $s \leq t$ . $\theta = E\{Z(1)\} = E\{B(1)\}$  is drift parameter.

# Summary

- ▶ For small sample sizes apply p-value boundary or use permutation test.
- ▶ classic boundaries (Haybittle-Peto, Pocock, O'Brien-Fleming) require equal spacing and pre-specification of number of looks.
- ▶ Desirable z-score boundaries (e.g., O'Brien-Fleming) are high early and close to  $z_\alpha$  at end. Effect on power is minimal (unlike Pocock).
- ▶ Haybittle-Peto simple and valid regardless of joint distribution of test statistic, but has reversal of fortune problem.
- ▶ Alpha spending functions are flexible and preferable to classic boundaries. Neither number nor timing of looks must be pre-specified. Can pick shape to ensure desired properties.
- ▶ Inference following monitoring is complicated.

# References I

- Armitage, P., McPherson, C., and Rowe, B. (1969). Repeated significance tests on accumulating data. Journal of the Royal Statistical Society: Series A (General) **132**, 235–244.
- CAST (1989). Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. New England Journal of Medicine **321**, 406–412.
- CAST (1992). Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. New England Journal of Medicine **327**, 227–233.
- CDP (1973). Prognostic importance of premature beats following myocardial infarction. JAMA **223**, 1116–1124.
- Haybittle, J. (1971). Repeated assessment of results in clinical trials of cancer treatment. The British journal of radiology **44**, 793–797.
- Jennison, C. and Turnbull, B. W. (1999). Group sequential methods with applications to clinical trials. CRC Press.

## References II

- Lan, K. G. and Zucker, D. M. (1993). Sequential monitoring of clinical trials: the role of information and brownian motion. Statistics in Medicine **12**, 753–765.
- Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. Biometrika **70**, 659–663.
- Moore, T. J. (1995). Deadly Medicine: Why tens of thousands of heart patients died in America's worst drug disaster. Simon & Schuster.
- O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. Biometrics pages 549–556.
- Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. Biometrika **64**, 191–199.
- Proschan, M. A., Lan, K. G., and Wittes, J. T. (2006). Statistical monitoring of clinical trials: a unified approach. Springer Science & Business Media.
- Ruberman, W., Weinblatt, E., Goldberg, J. D., Frank, C. W., and Shapiro, S. (1977). Ventricular premature beats and mortality after myocardial infarction. New England Journal of Medicine **297**, 750–757.

## Appendix: Unified Approach

- ▶ Consider trial with continuous outcome and paired differences  $D_1, D_2, \dots, D_N$ , one member on treatment, other on control.

$$\text{Interim: } Z_n = \frac{S_n}{\sqrt{n\sigma^2}}, \quad \text{Final: } Z_N = \frac{S_N}{\sqrt{N\sigma^2}}.$$

$$\begin{aligned} \text{cov}(Z_n, Z_N) &= \text{cov}\left\{\frac{S_n}{\sqrt{n\sigma^2}}, \frac{S_N}{\sqrt{N\sigma^2}}\right\} \\ &= \frac{\text{cov}\{S_n, S_N\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \frac{\text{cov}\{S_n, S_n + (S_N - S_n)\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \end{aligned}$$

## Appendix: Unified Approach

- ▶ Consider trial with continuous outcome and paired differences  $D_1, D_2, \dots, D_N$ , one member on treatment, other on control.

$$\text{Interim: } Z_n = \frac{S_n}{\sqrt{n\sigma^2}}, \quad \text{Final: } Z_N = \frac{S_N}{\sqrt{N\sigma^2}}.$$

$$\begin{aligned} \text{cov}(Z_n, Z_N) &= \text{cov}\left\{\frac{S_n}{\sqrt{n\sigma^2}}, \frac{S_N}{\sqrt{N\sigma^2}}\right\} \\ &= \frac{\text{cov}\{S_n, S_N\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \frac{\text{cov}\{S_n, S_n + (S_N - S_n)\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \end{aligned}$$

## Appendix: Unified Approach

- ▶ Consider trial with continuous outcome and paired differences  $D_1, D_2, \dots, D_N$ , one member on treatment, other on control.

$$\text{Interim: } Z_n = \frac{S_n}{\sqrt{n\sigma^2}}, \quad \text{Final: } Z_N = \frac{S_N}{\sqrt{N\sigma^2}}.$$

$$\begin{aligned} \text{cov}(Z_n, Z_N) &= \text{cov}\left\{\frac{S_n}{\sqrt{n\sigma^2}}, \frac{S_N}{\sqrt{N\sigma^2}}\right\} \\ &= \frac{\text{cov}\{S_n, S_N\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \frac{\text{cov}\{S_n, S_n + (S_N - s_n)\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \end{aligned}$$

## Appendix: Unified Approach

$$\begin{aligned} &= \frac{\text{cov}\{S_n, S_n\} + \text{cov}\{S_n, (S_N - s_n)\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \frac{\text{var}(S_n) + 0}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} = \frac{n\sigma^2}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \sqrt{n/N} = \sqrt{t}, \quad t = n/N. \end{aligned}$$

- ▶  $t$  is called the **information fraction** or **information time** of the interim analysis.
- ▶ Note that  $t = 0$  and  $1$  at the beginning and end of the trial.



## Appendix: Unified Approach

$$\begin{aligned} &= \frac{\text{cov}\{S_n, S_n\} + \text{cov}\{S_n, (S_N - s_n)\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \frac{\text{var}(S_n) + 0}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} = \frac{n\sigma^2}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \sqrt{n/N} = \sqrt{t}, \quad t = n/N. \end{aligned}$$

- ▶  $t$  is called the **information fraction** or **information time** of the interim analysis.
- ▶ Note that  $t = 0$  and  $1$  at the beginning and end of the trial.

## Appendix: Unified Approach

$$\begin{aligned} &= \frac{\text{cov}\{S_n, S_n\} + \text{cov}\{S_n, (S_N - s_n)\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \frac{\text{var}(S_n) + 0}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} = \frac{n\sigma^2}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \sqrt{n/N} = \sqrt{t}, \quad t = n/N. \end{aligned}$$

- ▶  $t$  is called the **information fraction** or **information time** of the interim analysis.
- ▶ Note that  $t = 0$  and  $1$  at the beginning and end of the trial.

## Appendix: Unified Approach

- ▶ Similarly, at interim analyses with  $n_1$  and  $n_2$  observations,  $n_1 < n_2$ ,

$$\text{cov}(Z_{n_1}, Z_{n_2}) = \sqrt{\frac{n_1}{n_2}} = \sqrt{\frac{n_1/N}{n_2/N}} = \sqrt{t_1/t_2}. \quad (10)$$

- ▶ The closer the two interim analyses are, the higher the correlation of z-statistics. The mean of the z-statistic is:

$$\begin{aligned} E\{Z(t)\} &= E\left\{\frac{S_n}{\sqrt{n\sigma^2}}\right\} \\ &= \frac{n\mu}{\sqrt{n\sigma^2}} = \frac{\sqrt{n}\mu}{\sigma} \end{aligned}$$

$$\theta \equiv E\{Z(1)\} = \frac{\sqrt{N}\mu}{\sigma}, \text{ so}$$

$$E\{Z(t)\} = \theta\sqrt{t}. \quad (11)$$

## Appendix: Unified Approach

- ▶ Similarly, at interim analyses with  $n_1$  and  $n_2$  observations,  $n_1 < n_2$ ,

$$\text{cov}(Z_{n_1}, Z_{n_2}) = \sqrt{\frac{n_1}{n_2}} = \sqrt{\frac{n_1/N}{n_2/N}} = \sqrt{t_1/t_2}. \quad (10)$$

- ▶ The closer the two interim analyses are, the higher the correlation of z-statistics. The mean of the z-statistic is:

$$\begin{aligned} E\{Z(t)\} &= E\left\{\frac{S_n}{\sqrt{n\sigma^2}}\right\} \\ &= \frac{n\mu}{\sqrt{n\sigma^2}} = \frac{\sqrt{n}\mu}{\sigma} \end{aligned}$$

$$\theta \equiv E\{Z(1)\} = \frac{\sqrt{N}\mu}{\sigma}, \text{ so}$$

$$E\{Z(t)\} = \theta\sqrt{t}. \quad (11)$$

## Appendix: Unified Approach

- ▶ Similarly, at interim analyses with  $n_1$  and  $n_2$  observations,  $n_1 < n_2$ ,

$$\text{cov}(Z_{n_1}, Z_{n_2}) = \sqrt{\frac{n_1}{n_2}} = \sqrt{\frac{n_1/N}{n_2/N}} = \sqrt{t_1/t_2}. \quad (10)$$

- ▶ The closer the two interim analyses are, the higher the correlation of z-statistics. The mean of the z-statistic is:

$$\begin{aligned} E\{Z(t)\} &= E\left\{\frac{S_n}{\sqrt{n\sigma^2}}\right\} \\ &= \frac{n\mu}{\sqrt{n\sigma^2}} = \frac{\sqrt{n}\mu}{\sigma} \end{aligned}$$

$$\theta \equiv E\{Z(1)\} = \frac{\sqrt{N}\mu}{\sigma}, \text{ so}$$

$$E\{Z(t)\} = \theta\sqrt{t}. \quad (11)$$

## Appendix: Unified Approach

- ▶ By the central limit theorem,  $Z(t_i)$  is approximately normal for large sample size.
- ▶ Moreover, the joint distribution of  $\{Z(t_1), \dots, Z(t_k)\}$  is also approximately multivariate normal.
- ▶ Sometimes it is more convenient to look at the ***B-value*** rather than the z-statistic.

$$B(t) = \sqrt{t}Z(t). \quad (12)$$

## Appendix: Unified Approach

- ▶ The B-value has mean

$$\begin{aligned} E\{B(t)\} &= \sqrt{t}E\{Z(t)\} \\ &= \sqrt{t}(\theta\sqrt{t}) = \theta t. \end{aligned} \tag{13}$$

Also, for  $s \leq t$ ,

$$\begin{aligned} \text{cov}\{B(s), B(t)\} &= \text{cov}\{\sqrt{s}Z(s), \sqrt{t}Z(t)\} \\ &= \sqrt{s}\sqrt{t} \text{cov}\{Z(s), Z(t)\} \\ &= \sqrt{s}\sqrt{t} \sqrt{\frac{s}{t}} = s. \end{aligned} \tag{14}$$

## Appendix: Unified Approach

- ▶ The B-value has mean

$$\begin{aligned} E\{B(t)\} &= \sqrt{t}E\{Z(t)\} \\ &= \sqrt{t}(\theta\sqrt{t}) = \theta t. \end{aligned} \tag{13}$$

Also, for  $s \leq t$ ,

$$\begin{aligned} \text{cov}\{B(s), B(t)\} &= \text{cov}\{\sqrt{s}Z(s), \sqrt{t}Z(t)\} \\ &= \sqrt{s}\sqrt{t} \text{cov}\{Z(s), Z(t)\} \\ &= \sqrt{s}\sqrt{t} \sqrt{\frac{s}{t}} = s. \end{aligned} \tag{14}$$



## Appendix: Unified Approach

- ▶ The B-value has mean

$$\begin{aligned} E\{B(t)\} &= \sqrt{t}E\{Z(t)\} \\ &= \sqrt{t}(\theta\sqrt{t}) = \theta t. \end{aligned} \tag{13}$$

Also, for  $s \leq t$ ,

$$\begin{aligned} \text{cov}\{B(s), B(t)\} &= \text{cov}\{\sqrt{s}Z(s), \sqrt{t}Z(t)\} \\ &= \sqrt{s}\sqrt{t} \text{cov}\{Z(s), Z(t)\} \\ &= \sqrt{s}\sqrt{t}\sqrt{\frac{s}{t}} = s. \end{aligned} \tag{14}$$

## Appendix: Unified Approach

- ▶ Notice that, for  $s < t$ ,

$$\begin{aligned}\operatorname{cov}\{B(s), B(t) - B(s)\} &= \operatorname{cov}\{B(s), B(t)\} - \operatorname{cov}\{B(s), B(s)\} \\ &= s - s = 0.\end{aligned}\tag{15}$$

- ▶ Because  $B(s)$  and  $B(t) - B(s)$  are bivariate normal with 0 correlation,  $B(s)$  and  $B(t) - B(s)$  are independent.
- ▶ (***Independent increments property***) More generally, for  $t_1 < t_2 < \dots < t_k$ , the increments  $B(t_1)$ ,  $B(t_2) - B(t_1)$ ,  $\dots$ ,  $B(t_k) - B(t_{k-1})$  are independent.
- ▶  $B(t)$  is continuous everywhere, differentiable nowhere.

## Appendix: Unified Approach

- ▶ Notice that, for  $s < t$ ,

$$\begin{aligned}\operatorname{cov}\{B(s), B(t) - B(s)\} &= \operatorname{cov}\{B(s), B(t)\} - \operatorname{cov}\{B(s), B(s)\} \\ &= s - s = 0.\end{aligned}\tag{15}$$

- ▶ Because  $B(s)$  and  $B(t) - B(s)$  are bivariate normal with 0 correlation,  $B(s)$  and  $B(t) - B(s)$  are independent.
- ▶ (***Independent increments property***) More generally, for  $t_1 < t_2 < \dots < t_k$ , the increments  $B(t_1)$ ,  $B(t_2) - B(t_1)$ ,  $\dots$ ,  $B(t_k) - B(t_{k-1})$  are independent.
- ▶  $B(t)$  is continuous everywhere, differentiable nowhere.

## Lecture 5: Monitoring for Futility

# What is futility monitoring?

Interim look at the analysis of the primary endpoint for the purposes of examining whether the trial has a reasonable chance of providing useful scientific evidence

- ▶ Futility analyses often consider the current trend in the data and whether the trial has a reasonable chance of producing a statistically significant result at end of study
- ▶ There could be many factors contributing to possible futility
  - ▶ Lack of treatment effect
  - ▶ Lower than expected recruitment rate
  - ▶ Lower than expected event rate
  - ▶ Emerging evidence from other trials
- ▶ Futility caused by poor recruitment, too much loss to follow-up or non-adherence, is called *operational futility*

# When is futility monitoring worth considering?

- ▶ Futility monitoring gives a trial a chance to stop early if there appears to be little to no chance the trial will provide useful evidence
- ▶ Trials for which a failure to show an advantage for a new treatment would not lead to changes in medical practice would be candidates for interim futility assessments
- ▶ Stopping trials on a path to failure prevents patients from taking unnecessary risks
- ▶ Stopping trials on a path to failure saves resources that can be redirected to therapies with better potential
- ▶ Some have argued large publicly funded trials should always consider futility monitoring, saving costs and recruiting fewer patients to failed trials (Sully et al., 2014)

## Table 1. Trials That May Incorporate Interim Futility Analysis in Their Monitoring Plans.

### Trial Type

Placebo-controlled trials of an investigational treatment for a serious medical condition

Trials comparing an investigational treatment with a standard treatment

Trials comparing a drug combination with one or more of its components

Trials comparing a higher dose with the standard dose of an available treatment

Trials involving considerable participant burden or cost

# The Cardiovascular Inflammation Reduction Trial (CIRT)

## Background

- ▶ Inflammation plays a key role in atherothrombosis
- ▶ The Canakinumab Anti-inflammatory Thrombosis Outcomes Study (CANTOS) found use of a monoclonal antibody reduced cardiovascular events over placebo, without lowering blood pressure or lipids
  - ▶ Largest reduction in events were in subjects with largest reductions in interleukin-6 and high-sensitivity C-reactive protein
- ▶ There was interest to study another anti-inflammatory drug
  - ▶ Low-dose methotrexate is an inexpensive, effective, and widely used treatment for inflammatory conditions, including rheumatoid arthritis, psoriatic and juvenile arthritis
  - ▶ In observational studies, patients with rheumatoid/psoriatic arthritis who received low-dose methotrexate had fewer cardiovascular events than patients receiving other therapies/placebo.



# The Cardiovascular Inflammation Reduction Trial (CIRT)

NCT01594333

- ▶ Randomized placebo-controlled trial examining whether low-dose methotrexate reduces heart attacks, strokes, or death
- ▶ NIH-sponsored trial launched in 2013 with goal of randomizing 7000 men and women
- ▶ Trial included medically stable participants with type 2 diabetes or metabolic syndrome and history of a heart attack or multiple coronary blockages
- ▶ Study protocol included a statistical plan for early termination for “futility”
  - ▶ If emerging data indicate trial unlikely to demonstrate benefit, this would not lead to changes in clinical practice

# CIRT Futility Analysis

NCT01594333

- ▶ There were two planned looks, when 50% and 75% of the planned primary events had accrued
- ▶ In March of 2018, the DSMB recommended trial termination.
  - ▶ Median follow-up of 2.3 years in 4786 participants
- ▶ The Methotrexate HR crossed a prespecified inefficacy boundary (Freidlin et al. (2010))
- ▶ Methotrexate did not reduce high-sensitivity C-reactive protein levels during the run-in phase
- ▶ Methotrexate elevated liver enzymes

# ACTIV-4B Example

A trial that ended for logistical reasons

- ▶ Accelerating Covid-19 Therapeutic Interventions and Vaccines (ACTIV-4B), a placebo-controlled trial testing antithrombotic agents given prophylactically to people with Covid-19 who had not yet been hospitalized
- ▶ Basis for trial was that thrombosis was a known risk for subjects with Covid-19
- ▶ DSMB recommended trial end for futility when it was observed event rate far too low to demonstrate benefit for treatment
  - ▶ 3/558 participants had a thrombotic event (Connors et al., 2021)
- ▶ Decision to conclude for futility also that such a low event rate does not justify the use of anticoagulant or antithrombotic drugs (Ellenberg and Shaw, 2022)

# When is futility monitoring **NOT** worth considering?

- ▶ Even if a treatment is showing little benefit partway through the trial, additional safety evidence may be desired
- ▶ Trials comparing two or more widely used therapies to see whether one has advantages over the other
  - ▶ Non-inferiority trials typically would not need non-inferiority monitoring
- ▶ Some settings would require full evidence to accrue in order to have a convincing null result (as long as ethical)
  - ▶ Trials of therapies already in use may need a convincing result to change practice
- ▶ Although futility monitoring may not have been pre-specified, unexpected events or trends in trial may lead a DSMB to recommend consideration of futility

# Testosterone Trials

Snyder et al. (2016)

- ▶ Testosterone Trials studied testosterone therapy in older men with documented subnormal testosterone levels
- ▶ Trial evaluated a widely used product not well studied for many of the functional outcomes it was advertised for
- ▶ It was deemed important to collect as much data as possible
- ▶ Stopping early for futility would be less likely to persuade providers and consumers than a larger database, and full safety information was particularly important given concerns about testosterone's effects on prostate and cardiovascular health
- ▶ No futility plan was provided by the study investigators

# Women's Health Initiative (WHI)

Women's Health Initiative Study Group and others (1998)

The WHI Hormone Replacement Therapy (HRT) Trials studied the risks and benefits of estrogen therapy in post-menopausal women.

- ▶ HRT came into wide use in 1960s, based largely on presumptions and observational data
  - ▶ Observational studies suggested HRT reduced a women's risk of coronary heart disease (CHD) by 40-50%
  - ▶ Evidence from trials suggested estrogen prevented hip fracture
  - ▶ Some concern regarding increased risk of breast cancer
- ▶ Two trials launched in 1993 that would enroll a 27,347 women.
  - ▶ 16,608 in the progestin+ estrogen trial (EP) and 10,739 in the estrogen alone (E alone) trial.
- ▶ Primary outcome was CHD; secondary outcome hip fracture; safety outcome breast cancer
- ▶ The WHI HRT trials monitored for efficacy and harm
- ▶ The results of these trials changed clinical practice (Rossouw et al., 2002; Anderson et al., 2004)

# Monitoring the Women's Health Initiative (WHI)

Wittes et al. (2007)

- ▶ There were a number of twists and turns in the monitoring of the WHI HRT, including early indications of increased risk of venous thrombosis and stroke
- ▶ In May 2001, the DSMB was convinced neither trial would show a benefit of HRT on CHD
  - ▶ Study continued because there would need to be unequivocal results in order to change practice
  - ▶ More data would also help elucidate some contrary trends: EP seemed to increase or breast cancer, E alone decreased risk
- ▶ In July 2002, 3 years before expected end, its EP arm was halted because compared to placebo, experimental developed more heart disease, invasive breast cancer, and other harmful outcomes such that risks outweighed benefits
- ▶ In Feb 2004, the E alone arm was halted due to concerns of stroke, with no apparent benefit on CHD

# PRECISION Trial

Nissen et al. (2016)

- ▶ Non-steroidal anti-inflammatory drugs were introduced in the 60s and became most widely prescribed drug in world
- ▶ Cox-2 inhibitors were a special class of NSAIDS thought to reduce gastrointestinal side effects but rofecoxib found to be associated with possible cardiovascular harm
- ▶ FDA mandated a trial of cardiovascular safety to evaluate another Cox-2 inhibitor celecoxib after another trial raised concern of increased risk
- ▶ PRECISION trial launched to establish non-inferiority of the Cox-2 inhibitor celecoxib to ibuprofen and naproxen with respect to cardiovascular death (including hemorrhagic death), nonfatal myocardial infarction, or nonfatal stroke
- ▶ Gastrointestinal and renal outcomes were also examined.
- ▶ An expected 20,000 randomized and 762 event needed for 90% power to establish non-inferiority with an upper margin of 1.33



# PRECISION Results

Nissen et al. (2016)

- ▶ 24,081 subjects requiring NSAIDs for osteoarthritis or rheumatoid arthritis were randomly assigned in a 1:1:1 fashion to celecoxib, naproxen and ibuprofen
- ▶ celecoxib non inferior to naproxen, 0.93; (95% confidence interval [CI], 0.76 to 1.13); and non-inferior to ibuprofen, 0.85; 95% CI, 0.70 to 1.04;  $P < 0.001$
- ▶ The risk of gastrointestinal events was significantly lower with celecoxib than with naproxen ( $P = 0.01$ ) or ibuprofen ( $P = 0.002$ )
- ▶ risk of renal events was significantly lower with celecoxib than with ibuprofen ( $P = 0.004$ ) but was not significantly lower with celecoxib than naproxen ( $P = 0.19$ )

## Methods for monitoring fertility

# Conditional Power

**Conditional Power (CP)** is the probability that a trial would successfully reject the null hypothesis if the trial continued until planned completion.

- ▶ CP useful in settings in which a convincing positive result at the end of the trial is needed to change clinical practice
- ▶ If the conditional power falls below a pre-specified threshold (commonly 10 to 20%), termination for futility may be considered
- ▶ CP must be computed by an unblinded statistician
  - ▶ Calculated using the observed trend in the data so far and a hypothesized trend for the data not yet collected
  - ▶ A DSMB may wish to see several estimates of conditional power based on a range of assumptions about the true treatment effect.
  - ▶ CP using the originally hypothesized trend generally the main estimate of CP
  - ▶ A binding rule w.r.t. CP (i.e., stop if  $CP \leq \gamma$  is called **stochastic curtailment**)

# Conditional Power (CP)

To calculate the CP at an interim analysis, we again can make use of the **B-value**  $B(t) = \sqrt{t}Z(t)$  and **information fraction**  $t$  (see Lecture 4)

- ▶ Where  $Z(t)$  is the Z-score at the interim analysis and
- ▶  $n$  observations out of the planned  $N$  are available at the interim analysis

# Conditional Power Calculation (part 1)

Proschan (2021)

Suppose  $c$  is the critical value needed for significance at end of trial

$$\begin{aligned} CP &= P\{Z(1) > c \mid Z(t) = z\} = P\{B(1) > c \mid B(t) = z\sqrt{t}\} \\ &= P\{B(1) - B(t) > c - z\sqrt{t} \mid B(t) = z\sqrt{t}\} \\ &= \\ &= P\{B(1) - B(t) > c - z\sqrt{t}\} \text{ (independent increments).} \end{aligned}$$

Also,  $B(1) - B(t)$  is normal with mean  $\theta \cdot 1 - \theta \cdot t = \theta(1 - t)$ , where  $\theta = E\{Z(1)\}$ , and variance

$$\begin{aligned} \text{var}\{B(1) - B(t)\} &= \text{var}\{B(1)\} + \text{var}\{B(t)\} - 2\text{cov}\{B(1), B(t)\} \\ &= 1 + t - 2t \\ &= 1 - t. \end{aligned}$$

## Conditional Power Calculation (part 2)

- ▶ Therefore, conditional power is

$$\begin{aligned} CP &= P \left\{ \frac{B(1) - B(t) - \theta(1-t)}{\sqrt{1-t}} > \frac{c - z\sqrt{t} - \theta(1-t)}{\sqrt{1-t}} \right\} \\ &= 1 - \Phi \left\{ \frac{c - z\sqrt{t} - \theta(1-t)}{\sqrt{1-t}} \right\} \\ &= \Phi \left\{ \frac{z\sqrt{t} + \theta(1-t) - c}{\sqrt{1-t}} \right\}, \end{aligned}$$

where  $z$  is value of the  $z$ -statistic at information time  $t$ .

$\theta = E(Z(1))$  and  $c$  is critical value at end of study

# CP Example: t-test

Proschan (2021)

Consider trial randomizing patients with hepatitis B to new drug (N) and standard (S). Primary outcome: change in  $\log_{10}$  viral load between baseline and 1 week expected SD = 2.8, N=250 per arm gives 85% power to detect 0.8 log difference.

At interim analysis:  $n_S=108$  and  $n_N = 111$ , mean change:

$\bar{Y}_S = -0.30$  and  $\bar{Y}_N = -0.18$ , sd of change:  $s_S = 2.35$  and  $s_N = 2.60$

so pooled variance =  $\frac{(108-1)s_S^2 + (111-1)s_N^2}{108+111-2} = 6.15$

information fraction  $t = \frac{1/\text{var}(\hat{\delta})}{1/\text{var}(\hat{\delta}_{end})} = \frac{1/\{\sigma^2(1/108+1/111)\}}{1/(2\sigma^2/250)}$

or  $t \approx (108 + 111)/2/250 = 0.438$

$Z(0.438) = \frac{\bar{Y}_S - \bar{Y}_N}{\sqrt{s^2(1/n_S + 1/n_N)}} = \frac{-0.30 - (-0.18)}{\sqrt{6.15(1/108 + 1/111)}} = -0.358$

$B(0.438) = \sqrt{t}Z(t) = \sqrt{0.438}(-0.358) = -0.237$

## CP Example: t-test (part 2)

Conditional power under original alternative hypothesis:

Because original power was 85%, expected Z score at end of trial is

$$\theta = 1.96 + 1.04 = 3$$

$$CP_3 = \Phi\left(\frac{-0.237 + 3(1 - 0.438) - 1.96}{\sqrt{1 - 0.438}}\right) = \Phi(-0.682) = 0.25$$

Conditional power under the null hypothesis:

$$CP_0 = \Phi\left(\frac{-0.237 + 0(1 - 0.438) - 1.96}{\sqrt{1 - 0.438}}\right) = \Phi(-2.931) = 0.002$$



# PREVAIL II: Binomial CP Example

Proschan (2021)

PREVAIL II was a trial of Ebola virus disease that compared triple monoclonal antibody product ZMapp + standard of care (SOC) with SOC alone

Primary endpoint was 28-day mortality, target ss was 100/arm that gave  $\approx 87\%$  power to detect a 50% reduction in mortality: 40% to 20% Trial ended with 71 observations because epidemic ended but an interesting question is would the results have been significant at end of trial?

There were: 13/35 deaths on SOC and 8/36 deaths on SOC+ZMapp.

$$t = \frac{1/\{p(1-p)(1/35+1/36)\}}{100/\{2p(1-p)\}} = 0.355$$

$$Z(.355) = \frac{13/35 - 8/36}{\sqrt{(21/71)(1-21/71)(1/35+1/36)}} = 1.377$$

$$B(.355) = \sqrt{(.355)}Z(.355) = .820$$

$$\theta = E(Z(1)) = \frac{0.40 - 0.20}{\sqrt{(2(.30)(1-.30))/100}} = 3.086$$

# PREVAIL II: Binomial CP Example (part 2)

Proschan (2021)

Conditional power under original alternative hypothesis:

$$CP_{orig} = \Phi\left(\frac{0.820 + 3.086(1 - 0.0355) - 1.96}{\sqrt{1 - 0.355}}\right) = \Phi(1.059) = 0.86$$

Conditional power under current trend

$$CP_{trend} = \Phi\left(\frac{0.820 + 2.31(1 - 0.355) - 1.96}{\sqrt{1 - 0.355}}\right) = \Phi(0.436) = 0.67$$

## PREVAIL II: Survival CP Example

Suppose PREVAIL II had planned to use the logrank test instead of test of proportions and that the trial had 80% power to detect a control to treatment hazard ratio of 1.6, with 142 events.

Suppose at interim analysis, there were 21 deaths observed and that the log-rank zscore was  $Z=0.83$

The expected Z score at end  $\theta = 1.96 + .84 = 2.80$

Information fraction is the ratio of current number of deaths to the number expected at end of trial  $d/D = 21/142 = 0.148$

Then,  $B(0.148) = \sqrt{0.148}(0.830) = 0.319$

$$CP_{2.8} \approx \Phi\left(\frac{0.319+2.8(1-0.148)-1.96}{\sqrt{1-0.148}}\right) = \Phi(0.807) = 0.79$$

When amount of information is so low (15% in this case), nearly impossible to stop for futility using conditional power. Note, conditional power is close to the original power in this case.

# Recall CAST Example

## Cardiac Arrhythmia Suppression Trial (CAST)

- ▶ Tested whether suppressing arrhythmias in patients with prior heart attack reduces composite of sudden deaths/cardiac arrests.
- ▶ Planned for 425 sudden deaths/cardiac arrests by end of trial.
- ▶ At DSMB meeting 4/17/1989, 35 sudden deaths/cardiac arrests in active arm, 13 in placebo arm.

# CP can help decision making

CAST Example (Proschan, 2021)

- ▶ Expected z-score at end if treatment reduces sudden deaths/cardiac arrests by 25%  $\theta = \ln(4/3) * \sqrt{425/4} = 2.965$ .  
Note, 25% reduction means treatment/control hazard ratio is 3/4, so control/treatment hazard ratio is  $1/(3/4)=4/3$ .
- ▶ Information time at interim analysis was  $(35+13)/425=0.113$ .
- ▶ Z-score was  $Z(0.113) = -3.22 = z$  (suggesting harm).
- ▶ Conditional power assuming 25% reduction is

$$\begin{aligned} \Phi \left\{ \frac{z\sqrt{t} + \theta(1-t) - c}{\sqrt{1-t}} \right\} &\approx \Phi \left\{ \frac{-3.22\sqrt{.113} + 2.965(1-.113) - 1.96}{\sqrt{1-.113}} \right\} \\ &= \Phi(-.438) = 0.33. \end{aligned}$$

- ▶ Only 33% chance of proving benefit at end, given current results and optimistic 25% reduction. CP under null hypothesis = 0.0006.
- ▶ Final # events now predicted to be 300, not 425.
- ▶ Using 300 final events and recomputing  $\theta$  and  $t$ ,  $CP \approx 0.10$  under 25% reduction and  $CP = 0.0002$  under null hypothesis.

# Planning for Futility Monitoring

- ▶ Though futility monitoring tends to be less formal, generally good idea to specify out a plan in the protocol
- ▶ How often to monitor will likely depend on setting
  - ▶ Timing often tied to achieving some threshold, such as when 50% and 75% of event rates have accrued
  - ▶ Hard to have low conditional power when less than 25% of data accrued
- ▶ DSMB may ask for an evaluation of futility partway thru the trial, even if not pre-specified

## Other factors at play

When evaluating futility, DSMB may consider emerging results from other trials.

- ▶ In the CIRT trial described earlier, the conditional power to detect the originally targeted effect size was 28%.
- ▶ The observed effect sizes in recently completed trials of anti-inflammatory agents in similar populations were much smaller than the targeted effect size in CIRT, increasing the DSMB's concern that CIRT was very unlikely to show a benefit.
- ▶ This emerging outside evidence together with the lack of effect on inflammatory markers, in addition to low CP, contributed to the DSMB's recommendation to terminate the trial for futility.

# Predicted Intervals

Predicted intervals predict the confidence interval that might be observed at trial's end under a given assumption about the future data. (Evans et al., 2007; Li et al., 2009)

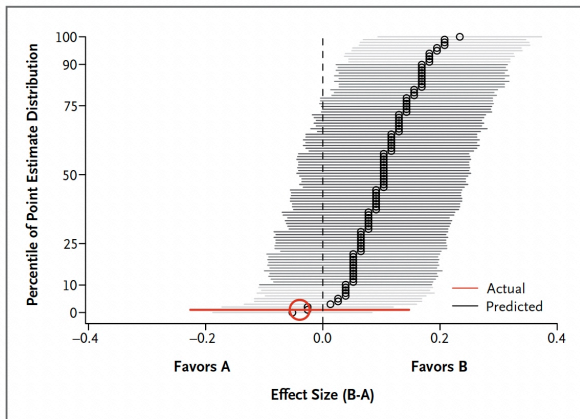
- ▶ Can enhance the interpretation of the conditional power at an interim analysis
- ▶ To calculate predicted intervals you simulate the remaining unobserved data for the trial, under a hypothesized trend, and combine with interim data
- ▶ Generate a large number of such confidence intervals (CI) and plot, ordering by effect size
- ▶ The comparisons of the width of the CI based on observed interim data alone with the width of the predicted interval sheds light on precision that could be gained with trial continuation, a potentially valuable tool for a DSMB
- ▶ Easily done in the R Software with package PIPS



# Predicted Intervals: Hypothetical Example

Ellenberg and Shaw (2022)

Predicted interval plot: 100 simulations of a hypothetical trial. At interim analysis: Deaths were 10/40 in group A vs 8/38 in group B halfway through trial. Hypothesized trend: 25% mortality for group A vs 50% mortality for group B for future data. Can see  $CP = 20\%$



Are there downsides to monitoring for fertility?

# Effect of monitoring for futility on Type 1 and Type II errors

- ▶ Monitoring for futility increases the probability that there will be a null result at the end of the trial (i.e. increased probability of failing to reject null hypothesis for primary endpoint)
- ▶ Type I error can not be inflated by monitoring for futility
- ▶ Type II error can be inflated

# The effect of stochastic curtailment on Power

- ▶ A **stochastic curtailment** rule like stop if CP under originally hypothesized treatment effect  $< 0.20$  lowers power because if you had continued, you might have gotten significant result at end
- ▶ Lan et al. (1982) showed that even if you monitored continuously, power is reduced only by small amount.
  - ▶ If type 2 error rate with no monitoring is  $\beta$  and stochastic curtailment rule stops if CP (**computed under original hypothesis**)  $\leq \gamma$ , then actual type 2 error rate is no greater than  $\beta/(1 - \gamma)$ .
- ▶ E.g., suppose stop if  $CP \leq 0.20$  and type 2 error rate was  $\beta = 0.10$  (power = 0.90) with no monitoring. Actual type 2 error rate is at most  $0.10/(1 - 0.20) = 0.125$ , so actual power is at least  $1 - 0.125 = 0.875$  even if you monitor continuously!

# Beta spending

- ▶ Another approach to futility monitoring is to consider a **beta-spending** approach. Analogous to alpha-spending, one could create a lower futility boundary which would guide stopping for futility if this boundary is crossed.
- ▶ The idea is that you can allow for repeated monitoring for futility while controlling the trial's false negative rate, thereby retaining trial power despite the multiple testing
- ▶ Advantage: Can choose the beta spending function of your liking (making it easier or harder to stop early)
- ▶ Tying upper boundary to lower futility boundary allows for slight increase in  $\alpha$ , to offset fact that futility monitoring lowers probability of falsely concluding significant result
  - ▶ This must be followed as a binding rule and so generally is NOT recommended.
- ▶ Better to not have upper tied to the lower boundary, since may ignore the lower boundary.

## Another downside to futility monitoring: Data in pipeline could change results

Whether the study team preparing report or DSMB member, need to think about a few things regarding stopping for futility

- ▶ Were the data used for the futility analysis the same for the endpoint used for final analysis: i.e. adjudicated endpoints?
- ▶ How much outstanding data was in pipeline at time of interim analysis?
- ▶ How much follow-up time/events will accrue between time of stopping for futility and final analysis?

# Example: LUME-2 Lung Trial

Hanna et al. (2016)

- ▶ LUME-Lung 2, a phase III trial of treatment of non–small cell lung cancer
- ▶ This trial was stopped early for futility on the basis of low conditional power (approximately 10%)
- ▶ Final trial results showed a significant benefit for the novel treatment on the primary end point of progression-free survival
- ▶ How did this happen?
  - ▶ Interim analysis based on the investigators' evaluation of disease progression, while final analysis based on centrally adjudicated determination of progression
  - ▶ There was also additional follow-up between stopping for futility and final analysis
  - ▶ Nice discussion by Lesaffre et al. (2017)
  - ▶ Conditional power might have been a useful tool to consider chances for result to turn around.

## Other Tools



# Revised Power

A **revised power** calculation can be done part-way through a trial using updated information on important design considerations, such as the variance of a continuous outcome, event rate, or the expected recruitment rate.

- ▶ The purpose of revised power is to determine whether a null result at the end of the trial would be informative under the updated assumptions.
- ▶ If the revised power is low, a null result would not rule out the original hypothesized treatment effect, suggesting that continuing the trial may be futile.
- ▶ **Revised power is done blinded**, unlike conditional power. It does not rely on interim trends in the data with respect to treatment effect.

# Revised Power: Don't always have to stop...

Love et al. (2015); Hade et al. (2019)

- ▶ An international breast cancer trial was launched in 2013 to study the effects of surgical timing during the menstrual phase on disease-free survival
  - ▶ Prior studies suggested adjuvant oophorectomy surgery during the luteal phase of the menstrual cycle may improve disease-free survival and overall survival compared to the follicular phase
- ▶ Concern arose from emergent data, including another trial, that the event rate assumed for the placebo arm during the planning stage may have been too high, potentially resulting in an underpowered trial
- ▶ In cooperation with the DSMB, investigators agreed to implement a blinded sample size re-estimation that relied on blinded trial data
- ▶ These calculations resulted in an increase in trial size from 340 to 510 due to the lower expected event rate.

# Bayesian Approaches

Snappin et al. (2006); Dmitrienko and Wang (2006)

There are a number of Bayesian approaches to futility monitoring

- ▶ One can compute **predictive power** (Dmitrienko and Wang, 2006), which is the conditional power averaged over a range of assumptions about the treatment difference that will be observed in the future data.
- ▶ The **predictive probability** framework is a fully Bayesian approach that specifies a prior probability for the treatment effect and, using the observed interim data, determines the posterior probability of a clinically important treatment difference
- ▶ For Bayesian approaches, careful thought must be given to specifying the prior probability.
  - ▶ Weak priors may give too much weight to early data
  - ▶ Dmitrienko and Wang (2006) argue that weak priors are advisable in large mortality trials to lessen the exposure of critically ill patients to ineffective interventions

# Summary

- ▶ Futility monitoring (FM) allows DSMB to evaluate mid-trial of whether scientifically useful information will be gained if trial continues until the planned end
- ▶ FM makes sense for some settings, not others
- ▶ Conditional Power one of the most common tools used for futility monitoring
- ▶ FM boundaries are generally considered advisory and not binding
- ▶ DSMB will always consider totality of evidence before recommending stopping for futility

# References I

- Anderson, G. L., Limacher, M., Assaf, A. R., Bassford, T., Beresford, S. A., Black, H., Bonds, D., Brunner, R., Brzyski, R., Caan, B., et al. (2004). Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: the women's health initiative randomized controlled trial. Jama **291**, 1701–1712.
- Clarke, C. A. and Glaser, S. L. (2007). Declines in breast cancer after the whi: apparent impact of hormone therapy. Cancer Causes & Control **18**, 847–852.
- Connors, J. M., Brooks, M. M., Sciruba, F. C., Krishnan, J. A., Bledsoe, J. R., Kindzelski, A., Baucom, A. L., Kirwan, B.-A., Eng, H., Martin, D., et al. (2021). Effect of antithrombotic therapy on clinical outcomes in outpatients with clinically stable symptomatic covid-19: the activ-4b randomized clinical trial. Jama **326**, 1703–1712.
- Dmitrienko, A. and Wang, M.-D. (2006). Bayesian predictive approach to interim monitoring in clinical trials. Statistics in medicine **25**, 2178–2195.
- Ellenberg, S. S. and Shaw, P. A. (2022). Early termination of clinical trials for futility—considerations for a data and safety monitoring board. NEJM Evidence **1**, EVIDctw2100020.
- Evans, S. R., Li, L., and Wei, L. (2007). Data monitoring in clinical trials using prediction. Drug information journal: DIJ/Drug Information Association **41**, 733–742.
- Freidlin, B., Korn, E. L., and Gray, R. (2010). A general inefficacy interim monitoring rule for randomized clinical trials. Clinical Trials **7**, 197–208.
- Hade, E. M., Young, G. S., and Love, R. R. (2019). Follow up after sample size re-estimation in a breast cancer randomized trial for disease-free survival. Trials **20**, 1–9.

# References II

- Hanna, N. H., Kaiser, R., Sullivan, R. N., Aren, O. R., Ahn, M.-J., Tiangco, B., Voccia, I., von Pawel, J., Kovcin, V., Agulnik, J., et al. (2016). Nintedanib plus pemetrexed versus placebo plus pemetrexed in patients with relapsed or refractory, advanced non-small cell lung cancer (lume-lung 2): a randomized, double-blind, phase iii trial. Lung Cancer **102**, 65–73.
- Lan, K., Simon, R., and Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. Communications in Statistics—Sequential Analysis **1**, 207–219.
- Lesaffre, E., Edelman, M., Hanna, N., Park, K., Thatcher, N., Willemsen, S., Gaschler-Markefski, B., Kaiser, R., and Manegold, C. (2017). Statistical controversies in clinical research: futility analyses in oncology—lessons on potential pitfalls from a randomized controlled trial. Annals of Oncology **28**, 1419–1426.
- Li, L., Evans, S. R., Uno, H., and Wei, L. (2009). Predicted interval plots (pips): a graphical tool for data monitoring of clinical trials. Statistics in biopharmaceutical research **1**, 348–355.
- Love, R. R., Laudico, A. V., Van Dinh, N., Allred, D. C., Uy, G. B., Quang, L. H., Salvador, J. D. S., Siguan, S. S. S., Mirasol-Lumague, M. R., Tung, N. D., et al. (2015). Timing of adjuvant surgical oophorectomy in the menstrual cycle and disease-free and overall survival in premenopausal women with operable breast cancer. JNCI: Journal of the National Cancer Institute **107**.
- Nissen, S. E., Yeomans, N. D., Solomon, D. H., Lüscher, T. F., Libby, P., Husni, M. E., Graham, D. Y., Borer, J. S., Wisniewski, L. M., Wolski, K. E., et al. (2016). Cardiovascular safety of celecoxib, naproxen, or ibuprofen for arthritis. New England Journal of Medicine **375**, 2519–2529.
- Proschan, M. A. (2021). Statistical Thinking in Clinical Trials. Chapman and Hall/CRC.

# References III

- Rossouw, J., Anderson, G., Prentice, R., et al. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. Jama **288**, 321–333.
- Snapinn, S., Chen, M.-G., Jiang, Q., and Koutsoukos, T. (2006). Assessment of fertility in clinical trials. Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry **5**, 273–281.
- Snyder, P. J., Bhasin, S., Cunningham, G. R., Matsumoto, A. M., Stephens-Shields, A. J., Cauley, J. A., Gill, T. M., Barrett-Connor, E., Swerdloff, R. S., Wang, C., et al. (2016). Effects of testosterone treatment in older men. New England Journal of Medicine **374**, 611–624.
- Sully, B. G., Julious, S. A., and Nicholl, J. (2014). An investigation of the impact of fertility analysis in publicly funded trials. Trials **15**, 1–9.
- Wittes, J., Barrett-Connor, E., Braunwald, E., Chesney, M., Cohen, H. J., DeMets, D., Dunn, L., Dwyer, J., Heaney, R. P., Vogel, V., et al. (2007). Monitoring the randomized trials of the women's health initiative: the experience of the data and safety monitoring board. Clinical Trials **4**, 218–234.
- Women's Health Initiative Study Group and others (1998). Design of the women's health initiative clinical trial and observational study. Control Clin Trials **19**, 61–109.