10 Early Stopping of Clinical Trials

10.1 General Issues in Monitoring Clinical Trials

Clinical trials are monitored periodically and may be stopped early due to

- Serious toxicity or adverse events
- Established benefit
- No trend of interest
- Design or logistical difficulties too serious to fix

Clinical trials will be monitored by an independent Data Safety Monitoring Board (DSMB).

- Primary responsibility of a DSMB: Ensure the safety and well being of the patients in the trial.
- Members of a DSMB:
  - ⋆ Clinical
  - ⋆ Laboratory
  - ⋆ Epidemiology
  - ⋆ Biostatistics
  - ⋆ Data Management
  - ⋆ Ethics
- DSMB should have no conflict of interest with the study or studies they are monitoring; All the discussions of the DSMB are confidential.

DSMB's duty:

- Protocol review
- Interim reviews
  - ⋆ study progress
  - ⋆ quality of data
  - ⋆ safety
  - ⋆ efficacy and benefit
- Manuscript review

During the early stages of a clinical trial the focus is on administrative issues regarding the conduct of the study. These include:

- Recruitment/Entry Criteria
- Baseline comparisons
- Design assumptions and modifications
  - ⋆ entry criteria

- ⋆ treatment dose
- ⋆ sample size adjustments
- ⋆ frequency of measurements

- Quality and timeliness of data collection

- Focus on possible early termination due to established treatment difference.
- **Group-sequential** methods: monitor data sequentially at some finite fixed time points
- Information-based design and monitoring of clinical trials

The typical scenario where these methods can be applied is as follows:

- Data in a study are collected over calendar time.
- Want to see if one treatment (new) is better than the other (old)
- conduct interim analysis to assess whether there is sufficient "strong evidence" to warrant early termination of the study
- At each monitoring time, a test statistic is computed and compared to a stopping boundary.
- The stopping boundaries are chosen to preserve certain operating characteristics that are desired; i.e. level and power

The methods we present are general enough to include problems where

- t-tests are used to compare the mean of continuous random variables between treatments
- proportion tests for dichotomous response variables
- logrank test for censored survival data
- tests based on likelihood methods for either discrete or continuous random variables; i.e. Score test, Likelihood ratio test, Wald test using maximum likelihood estimators

10.2 Information Based Design and Monitoring

Underlying structure:

- Data are generated from

$$f(y, \Delta, \theta),$$

  where

  - $\star$ $\Delta =$ treatment effect
  - $\star$ $\theta =$ nuisance parameters

- We want to test the null hypothesis

$$H_0 : \Delta = 0$$

  versus the alternative

$$H_A : \Delta \neq 0.$$

Note The methods discussed here can be modified to test

$$H_0 : \Delta \leq 0$$

versus

$$H_A : \Delta > 0.$$

- At any interim analysis time t, our decision making will be based on the test statistic

$$T(t) = \frac{\hat{\Delta}(t)}{se\{\hat{\Delta}(t)\}},$$

where $\hat{\Delta}(t)$ is an estimator for Δ and $se\{\hat{\Delta}(t)\}$ is the estimated standard error of $\hat{\Delta}(t)$ using all the data that have accumulated up to time t. For two-sided tests we would reject the null hypothesis if the absolute value of the test statistic $|T(t)|$ were sufficiently large and for one-sided tests if $T(t)$ were sufficiently large.

- **Some examples**:

  1. **Example 1.** (Dichotomous response)

     Let π_1, π_0 denote the population response rates for treatments 1 (new treatment) and 0 (control) respectively. Let the treatment difference be given by

     $$\Delta = \pi_1 - \pi_0$$

     The test of the null hypothesis at time t will be based on

     $$T(t) = \frac{p_1(t) - p_0(t)}{\sqrt{\bar{p}(t)\{1 - \bar{p}(t)\}\left\{\frac{1}{n_1(t)} + \frac{1}{n_2(t)}\right\}}},$$

     where using all the data available through time t, $p_j(t)$ denotes the sample proportion responding among the $n_j(t)$ individuals on treatment $j = 0, 1$.

2. **Example 2.** (Time to event)

Suppose we assume a proportional hazards model:

$$\frac{\lambda_1(t)}{\lambda_0(t)} = \exp(-\Delta),$$

and we want to test the null hypothesis of no treatment difference

$$H_0 : \Delta = 0 \quad vs. \quad H_A : \Delta \neq 0$$

$$(\text{or } H_0 : \Delta \leq 0 \quad vs. \quad H_A : \Delta > 0.)$$

At time t, we would compute the test statistic

$$T(t) = \frac{\hat{\Delta}(t)}{se\{\hat{\Delta}(t)\}},$$

where $\hat{\Delta}(t)$ is the maximum partial likelihood estimator of Δ. For the two-sided test we would reject the null hypothesis if $|T(t)|$ were sufficiently large and for the one-sided test if $T(t)$ were sufficiently large.

Remark: The material on the use and the properties of the maximum partial likelihood estimator are taught in the classes on Survival Analysis. We note, however, that the logrank test computed using all the data up to time t is equivalent asymptotically to the test based on $T(t)$.

3. **Example 3.** (Any parametric models $p(z; \Delta, \theta)$):
   Find MLE $\hat{\Delta}(t)$ of Δ and its $se\{\hat{\Delta}(t)\}$ and conduct testing using

$$T(t) = \frac{\hat{\Delta}(t)}{se\{\hat{\Delta}(t)\}}$$

- Usually,
  - ⋆ Under $H_0 : \Delta = 0$,

$$T(t) = \frac{\hat{\Delta}(t)}{se\{\hat{\Delta}(t)\}} \overset{a}{\sim} \mathrm{N}(0,1).$$

  - ⋆ Under $H_A : \Delta = \Delta^* \neq 0$,

$$T(t) = \frac{\hat{\Delta}(t)}{se\{\hat{\Delta}(t)\}} \overset{\Delta=\Delta^*}{\sim} N(\Delta^* I^{1/2}(t, \Delta^*), 1),$$

where $I(t, \Delta^*)$ is the statistical information (for Δ) at t, and

$$I(t, \Delta^*) \approx \{se(\hat{\Delta}(t)\}^{-2}.$$

Group sequential test

- Determine the total number of interim analysis K.
- Determine boundary values $b(t_j)$ for $j = 1, 2, ..., K$.
- Reject $H_0 : \Delta = 0$ at the first time t_j when

$$|T(t_j)| \geq b(t_j),$$

  if we consider two-sided test.

- Question: What values $b(t_j)$ should be used?
- Suppose we want to control the overall type I error prob at α, can we use $b(t_j) = \mathcal{Z}_{\alpha/2}$?

- The actual type I error if $b(t_j) = 1.96$ is used:

K	false positive rate
1	0.050
2	0.083
3	0.107
4	0.126
5	0.142
10	0.193
20	0.246
50	0.320
100	0.274
1,000	0.530
∞	1.000

- $\Longrightarrow$ For given overall type I error prob α, need to find out $b(t_j)$ for $j = 1, 2, ..., K$ such that

$$P[\text{reject } H_0 | H_0] = \alpha,$$

or equivalently

$$P[\text{accept } H_0 | H_0] = 1 - \alpha$$

- Sequential test implies that we accept H_0 if

$$|T(t_j)| \leq b(t_j), \text{ for all } j = 1, 2, ...K.$$

$\Longrightarrow$

$$P_{\Delta=0}\{|T(t_1)| < b(t_1), \ldots, |T(t_K)| < b(t_K)\} = 1 - \alpha. \qquad (10.1)$$

- Need to know the joint distribution of $T(t_1), T(t_2), ..., T(t_K)$.

- Fundamental result:

  "Any efficient based test or estimator for Δ, properly normalized, when computed sequentially over time, has, asymptotically, a normal independent increments process whose distribution depends only on the parameter Δ and the statistical information."

  Scharfstein, Tsiatis and Robins (1997). JASA. 1342-1350.

- Define

$$W(t) = I^{1/2}(t, \Delta^*)T(t)$$

- Since when $\Delta = \Delta^*$,

$$T(t) = \frac{\hat{\Delta}(t)}{se\{\hat{\Delta}(t)\}} \sim N(\Delta^* I^{1/2}(t, \Delta^*), 1),$$

$$\Longrightarrow$$

$$w(t) \sim N(\Delta^* I(t, \Delta^*), I(t, \Delta^*)).$$

- The joint distribution of the multivariate vector $\{W(t_1), \ldots, W(t_K)\}$ is asymptotically normal with mean vector $\{\Delta^* I(t_1, \Delta^*), \ldots, \Delta^* I(t_K, \Delta^*)\}$ and covariance matrix where

$$var\{W(t_j)\} = I(t_j, \Delta^*),\ j = 1, \ldots, K$$

and

$$cov[W(t_j), \{W(t_\ell) - W(t_j)\}] = 0,\ j < \ell, j, \ell = 1, \ldots, K.$$

- That is,
  - ⋆ The statistic $W(t_j)$ has mean and variance proportional to the statistical information at time t_j
  - ⋆ Has independent increments; that is

$$\begin{aligned} W(t_1) &= W(t_1) \\ W(t_2) &= W(t_1) + \{W(t_2) - W(t_1)\} \\ &\vdots \\ W(t_j) &= W(t_1) + \{W(t_2) - W(t_1)\} + \ldots + \{W(t_j) - W(t_{j-1})\} \end{aligned}$$

    has the same distribution as a partial sum of independent normal random variables

  This structure implies that the covariance matrix of

$\{W(t_1), \ldots, W(t_K)\}$ is given by

$$var\{W(t_j)\} = I(t_j, \Delta^*),\ j = 1, \ldots, K$$

and for $j < \ell$

$$\begin{aligned}
& cov\{W(t_j), W(t_\ell)\} \\
= \;& cov[W(t_j), \{W(t_\ell) - W(t_j)\} + W(t_j)] \\
= \;& cov[W(t_j), \{W(t_\ell) - W(t_j)\}] + cov\{W(t_j), W(t_j)\} \\
= \;& 0 + var\{W(t_j)\} \\
= \;& I(t_j, \Delta^*).
\end{aligned}$$

- Since

$$T(t_j) = I^{-1/2}(t_j, \Delta^*)W(t_j),\ j = 1, \ldots, K$$

so $\{T(t_1), \ldots, T(t_K)\}$ is also multivariate normal with mean

$$E\{T(t_j)\} = \Delta^* I^{1/2}(t_j, \Delta^*),\ j = 1, \ldots, K \tag{10.2}$$

and the covariance matrix

$$var\{T(t_j)\} = 1,\ j = 1, \ldots, K \tag{10.3}$$

and for $j < \ell$, the covariances are

$$\begin{aligned} cov\{T(t_j), T(t_\ell)\} &= cov\{I^{-1/2}(t_j, \Delta^*)W(t_j), I^{-1/2}(t_\ell, \Delta^*)W(t_\ell)\} \\ &= I^{-1/2}(t_j, \Delta^*)I^{-1/2}(t_\ell, \Delta^*)cov\{W(t_j), W(t_\ell)\} \\ &= I^{-1/2}(t_j, \Delta^*)I^{-1/2}(t_\ell, \Delta^*)I(t_j, \Delta^*) \\ &= \frac{I^{1/2}(t_j, \Delta^*)}{I^{1/2}(t_\ell, \Delta^*)} = \sqrt{\frac{I(t_j, \Delta^*)}{I(t_\ell, \Delta^*)}}. \end{aligned} \tag{10.4}$$

- Under $H_0 : \Delta = 0$, $\{T(t_1), \ldots, T(t_K)\}$ is multivariate normal with mean vector zero and covariance (correlation) matrix

$$V_T = \left[\sqrt{\frac{I(t_j, 0)}{I(t_\ell, 0)}} \right], \ j \leq \ell. \tag{10.5}$$

- When we conduct interim analysis with equal increment of information (usually equal # of patients)

$$I(t_1, \cdot) = I, \ I(T_2, \cdot) = 2I, \ \ldots, \ I(t_K, \cdot) = KI,$$

then

$$\begin{bmatrix} T(t_1) \\ T(t_2) \\ \vdots \\ T(t_K) \end{bmatrix} \overset{H_0}{\sim} \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sqrt{\frac{1}{2}} & \cdots & \sqrt{\frac{1}{K}} \\ \sqrt{\frac{1}{2}} & 1 & \cdots & \sqrt{\frac{2}{K}} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\frac{1}{K}} & \sqrt{\frac{2}{K}} & \cdots & 1 \end{bmatrix} \right)$$

- Given this structure, for given boundary values $b_j = b(t_j), j = 1, 2, ..., K$ we can iteratively calculate

$$P_{\Delta=0}\{|T(t_1)| < b_1, \ldots, |T(t_K)| < b_K\}.$$

  Please see Armitage, McPherson and Rowe (1969, JRSS-A) for more detail.

- For given type I error α we want to control, too many b_j's satisfy

$$P_{\Delta=0}\{|T(t_1)| < b_1, \ldots, |T(t_K)| < b_K\} = 1 - \alpha.$$

- How do we choose b_j's?

10.3 Choice of Boundaries

- Wang and Tsiatis (1987) *Biometrics* proposed a flexible class of boundaries

$$b_j = (\text{constant}) \times j^{(\Phi - .5)},$$

where Φ determines the shape of a boundary (so Φ is called the shape parameter).

- For given α, K, and Φ, the constant c can be computed so that

$$P_{\Delta=0}\{\bigcap_{j=1}^{K} |T(t_j)| < cj^{(\Phi-.5)}\} = 1 - \alpha.$$

- Denote this c by $c(\alpha, K, \Phi)$.

Table 1: $c(\alpha, K, \Phi)$ *for some selected values of* α, K, Φ

	$\alpha = 0.05$				$\alpha = 0.01$			
	K				K			
Φ	2	3	4	5	2	3	4	5
0.0	2.7967	3.4712	4.0486	4.5618	3.6494	4.4957	5.2189	5.8672
0.1	2.6316	3.1444	3.5693	3.9374	3.4149	4.0506	4.5771	5.0308
0.2	2.4879	2.8639	3.1647	3.4175	3.2071	3.6633	4.0276	4.3372
0.3	2.3653	2.6300	2.8312	2.9945	3.0296	3.3355	3.5706	3.7634
0.4	2.2636	2.4400	2.5652	2.6628	2.8848	3.0718	3.2071	3.3137
0.5	2.1784	2.2896	2.3616	2.4135	2.7728	2.8738	2.9395	2.9869

- **Pocock boundaries**:

$$\Phi = 0.5, \implies b_j = c(\alpha, K, 0.5),\ j = 1, 2, ..., K$$

For example, if $K = 5$ and $\alpha = .05$, then $c(.05, 5, 0.5) = 2.41$.

That is, we reject $H_0 : \Delta = 0$ the first time where

$$|T(t_j)| \geq 2.41$$

Equivalently, we reject $H_0 : \Delta = 0$ the first time where

$$\mathrm{P-value} \leq 0.0158$$

- **O'Brien-Fleming boundaries**:

$$\Phi = 0, \Longrightarrow \; b_j = c(\alpha, K, 0) j^{-1/2} = c(\alpha, K, 0)/\sqrt{j}, \; j = 1, 2, ..., K$$

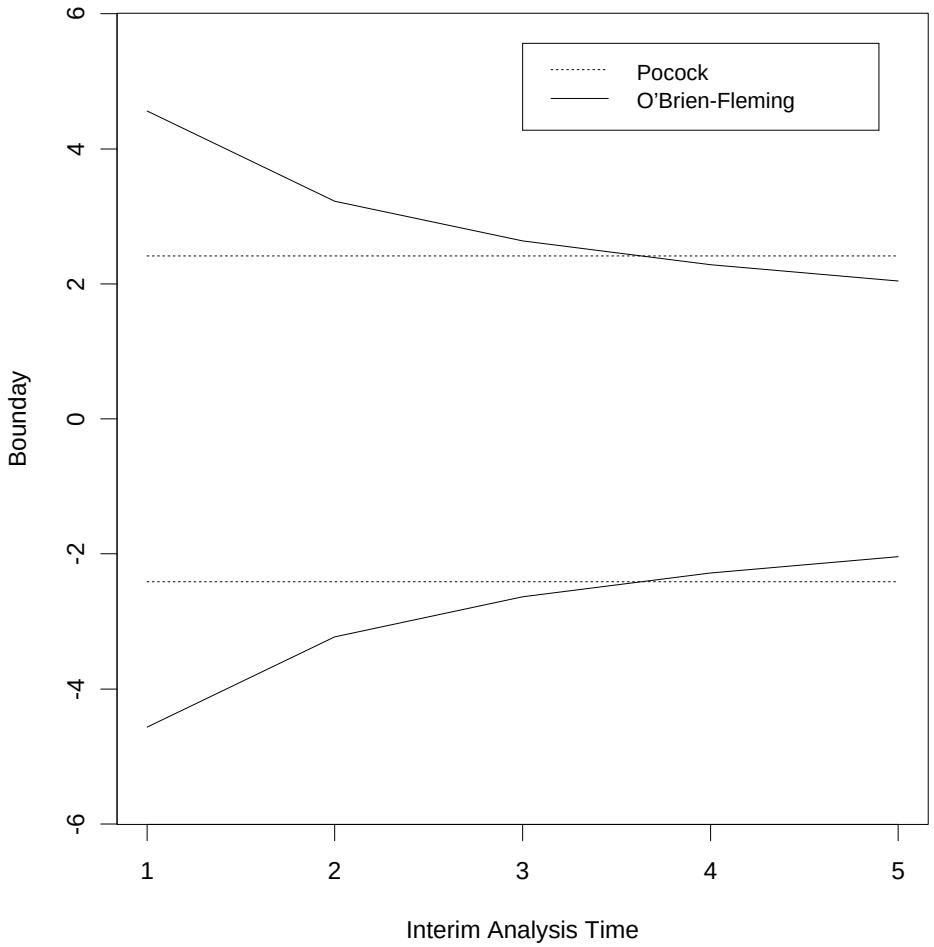
For example, if $K = 5$ and $\alpha = .05$, then $c(.05, 5, 0.0) = 4.56$. Then $b_j = c(\alpha, K, 0)/\sqrt{j} = 4.56/\sqrt{j}$ gives 5 boundary values:

$$b_1 = 4.56, \; b_2 = 3.22, \; b_3 = 2.63, \; b_4 = 2.28, \; b_5 = 2.04$$

Table 2: *Nominal p-values for* $K = 5$ *and* $\alpha = .05$

	Nominal p-value	
j	Pocock	O'Brien-Fleming
1	0.0158	0.000005
2	0.0158	0.00125
3	0.0158	0.00843
4	0.0158	0.0225
5	0.0158	0.0413

Figure 1: *Pocock and O'Brien-Fleming Boundaries*



10.4 Power and Sample Size in Terms of Information

- The test statistic $T(t)$:

$$T(t) \overset{\Delta=0}{\sim} N(0,1)$$

and under a clinically important alternative $\Delta = \Delta_A$

$$T(t) \overset{\Delta=\Delta_A}{\sim} N(\Delta_A \; I^{1/2}(t, \Delta_A), 1),$$

where $I(t, \Delta_A)$ denotes statistical information which can be approximated by $[se\{\hat{\Delta}(t)\}]^{-2}$

- In order that a two-sided level-α test have power $1-\beta$ to detect Δ_A, the noncentrality parameter $\Delta_A I^{1/2}(t,\Delta_A)$ must satisfy

$$\Delta_A I^{1/2}(t,\Delta_A) = \mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta},$$

or

$$I(t,\Delta_A) = \left\{ \frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}}{\Delta_A} \right\}^2 . \tag{10.6}$$

- Since $I(t,\Delta_A) = [se\{\hat{\Delta}(t)\}]^{-2}$, we need to collect enough data so that

$$[se\{\hat{\Delta}(t)\}]^{-2} = \left\{ \frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}}{\Delta_A} \right\}^2 .$$

- One strategy would monitor $se\{\hat{\Delta}(t)\}$ until it satisfies the above condition (at t^F and do data analysis). Reject $H_0 : \Delta = 0$ if

$$|T(t^F)| \geq \mathcal{Z}_{\alpha/2}.$$

- **Example**: Compare two response rates using $\Delta(t) = p_1(t) - p_0(t)$ at t:

$$se\{\hat{\Delta}(t)\} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1(t)} + \frac{\pi_0(1-\pi_0)}{n_0(t)}}.$$

Therefore, to obtain the desired power of $1-\beta$ to detect the alternative where the population response rates were π_1 and π_0, with $\pi_1 - \pi_0 = \Delta_A$, we would need the sample sizes $n_1(t^F)$ and $n_0(t^F)$ to satisfy

$$\left\{\frac{\pi_1(1-\pi_1)}{n_1(t^F)} + \frac{\pi_0(1-\pi_0)}{n_0(t^F)}\right\}^{-1} = \left\{\frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_\beta}{\Delta_A}\right\}^2.$$

Remark: This is essentially the same as what we did before.

- Power of group sequential test:

$$1 - P[|T(t_1)| < b_1, \ldots, |T(t_K)| < b_K | \Delta = \Delta_A],$$

  where $b_j, j = 1, 2, ..., K$ are determined by α, K and Φ.

- need to find out the maximum information (MI), similar to maximum sample size.

- If we do interim analysis after equal increment of information, then

$$\begin{bmatrix} T(t_1) \\ T(t_2) \\ \vdots \\ T(t_K) \end{bmatrix} \overset{H_A}{\sim} \mathrm{N}\left(\begin{bmatrix} \Delta_A\sqrt{\frac{MI}{K}} \\ \Delta_A\sqrt{\frac{2\times MI}{K}} \\ \vdots \\ \Delta_A\sqrt{\frac{K\times MI}{K}} \end{bmatrix}, \begin{bmatrix} 1 & \sqrt{\frac{1}{2}} & \cdots & \sqrt{\frac{1}{K}} \\ \sqrt{\frac{1}{2}} & 1 & \cdots & \sqrt{\frac{2}{K}} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\frac{1}{K}} & \sqrt{\frac{2}{K}} & \cdots & 1 \end{bmatrix} \right)$$

- Denote $\delta = \Delta_A \sqrt{MI}$. Then the distribution of $\{T(t_1), T(t_2), ..., T(t_K)\}$ is determined by δ and K:

$$\begin{bmatrix} T(t_1) \\ T(t_2) \\ \vdots \\ T(t_K) \end{bmatrix} \overset{H_A}{\sim} \mathrm{N}\left(\begin{bmatrix} \delta\sqrt{\frac{1}{K}} \\ \delta\sqrt{\frac{2}{K}} \\ \vdots \\ \delta\sqrt{\frac{K}{K}} \end{bmatrix}, \begin{bmatrix} 1 & \sqrt{\frac{1}{2}} & \cdots & \sqrt{\frac{1}{K}} \\ \sqrt{\frac{1}{2}} & 1 & \cdots & \sqrt{\frac{2}{K}} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\frac{1}{K}} & \sqrt{\frac{2}{K}} & \cdots & 1 \end{bmatrix} \right)$$

- We can find out δ for given α, K and Φ so that the power

$$1 - P_\delta[|T(t_1)| < b_1, \ldots, |T(t_K)| < b_K|] = 1 - \beta$$

- Denote this δ by $\delta(\alpha, K, \Phi, \beta)$.

- In order the sequential test to have power $1-\beta$ to detect Δ_A, we need:

$$\Delta_A \sqrt{MI} = \delta(\alpha, K, \Phi, \beta)$$

or

$$MI = \frac{\delta^2(\alpha, K, \Phi, \beta)}{\Delta_A^2}$$

I. Inflation Factor

- If we do data analysis only once, then for level α test to have power $1-\beta$ to detect Δ_A, the information has to be

$$I^{FS} = \left\{ \frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}}{\Delta_A} \right\}^2 .$$

- The maximum information using a group sequential test is usually larger than I^{FS}
- Denote the inflation factor by

$$IF(\alpha, K, \Phi, \beta) = \frac{MI}{I^{FS}} = \left\{ \frac{\delta(\alpha, K, \Phi, \beta)}{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}} \right\}^2$$

- If we know the inflation factor, we know

$$MI = IF(\alpha, K, \Phi, \beta) I^{FS}$$

- This can be used to calculate maximum sample size needed.

Table 3: *Inflation factors as a function of* K, α, β *and* Φ

		α=0.05			α=0.01		
		Power=1-β			Power=1-β		
K	Boundary	0.80	0.90	0.95	0.80	0.90	0.95
2	Pocock	1.11	1.10	1.09	1.09	1.08	1.08
	O-F	1.01	1.01	1.01	1.00	1.00	1.00
3	Pocock	1.17	1.15	1.14	1.14	1.12	1.12
	O-F	1.02	1.02	1.02	1.01	1.01	1.01
4	Pocock	1.20	1.18	1.17	1.17	1.15	1.14
	O-F	1.02	1.02	1.02	1.01	1.01	1.01
5	Pocock	1.23	1.21	1.19	1.19	1.17	1.16
	O-F	1.03	1.03	1.02	1.02	1.01	1.01
6	Pocock	1.25	1.22	1.21	1.20	1.19	1.17
	O-F	1.03	1.03	1.03	1.02	1.02	1.02
7	Pocock	1.26	1.24	1.22	1.22	1.20	1.18
	O-F	1.03	1.03	1.03	1.02	1.02	1.02

- **Example with dichotomous endpoint**: Let π_1 and π_0 be the population response rates for treatments 1 and 0. $\Delta = \pi_1 - \pi_0$.

  Want to test $H_0 : \Delta = 0$ $vs.$ $H_A : \Delta \neq 0$ at level $\alpha = 0.05$ using a 4-look O'Brien-Fleming boundary ($\Phi = 0$).

  ⋆ we will reject H_0 the first time when

$$
\begin{aligned}
|T(t_j)| &= \left| \frac{p_1(t_j) - p_0(t_j)}{\sqrt{\bar{p}(t_j)\{1 - \bar{p}(t_j)\}\left\{\frac{1}{n_1(t_j)} + \frac{1}{n_0(t_j)}\right\}}} \right| \\
&\geq 4.049/\sqrt{j},\ j = 1, \ldots, 4.
\end{aligned}
$$

The boundaries are given by

Table 4: *Boundaries for a 4-look O-F test*

j	b_j	nominal p-value
1	4.05	.001
2	2.86	.004
3	2.34	.019
4	2.03	.043

- Suppose $\pi_0 = 0.3$ and we would like to have power 90% to detect $\pi_1 = 0.45$, how do we design the study?
- The fixed sample size design requires

$$n^{FS} = \left\{ \frac{1.96 + 1.28\sqrt{\frac{.3\times.7+.45\times.55}{2\times.375\times.625}}}{.15} \right\}^2 \times 4 \times .375 \times .625 = 434$$

- The inflation factor for $\alpha = 0.05$, power=0.9, $K = 4$ and $\Phi = 0$ is IF=1.02. The maximum sample size using a group sequential test is

$$1.02 \times 434 = 444,$$

  or 222 patients for each treatment
- Since $222/4 = 56$, we recruit 56 patients for each treatment first and then do interim analysis. If we dont reject H_0, then recruit additional 56 patients to each treatment and do interim analysis, etc.

Information based monitoring

- $\pi_0 = 0.3$ and $\pi_1 = 0.45$ are needed to derive the sample size. They might not be the case in practice.
- Suppose we would like a level 0.05 test to have power 0.9 to detect $\Delta = 0.15$, then the information needed for a fixed sample size design is

$$\left\{ \frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}}{\Delta_A} \right\}^2 = \left\{ \frac{1.96 + 1.28}{.15} \right\}^2 = 466.6$$

- So the MI for a 4-look O-F design is $466.6 \times 1.02 = 475.9$
- So the information required at the jth interim analysis

$$\frac{j \times 475.9}{4} = 119 \times j, \ j = 1, \ldots, 4$$

- The information available at the jth interim analysis is approximately

$$[se\{\hat{\Delta}(t)\}]^{-2} = \left[\frac{p_1(t)\{1-p_1(t)\}}{n_1(t)} + \frac{p_0(t)\{1-p_0(t)\}}{n_0(t)}\right]^{-1}$$

- This implies that we should do interim analysis when

$$\left[\frac{p_1(t_j)\{1-p_1(t_j)\}}{n_1(t_j)} + \frac{p_0(t_j)\{1-p_0(t_j)\}}{n_0(t_j)}\right]^{-1} = 119 \times j,\ j = 1, \ldots, 4.$$

  and use the test statistic and the boundary values given before.

- The information-based monitoring will maintain the overall type I error prob and desired power to detect a difference of interest even if the nuisance parameter values may be different than what we assume.

II. Average information

- For the same design characteristics, which boundary is better? Pocock or O-F?
- Pocock design has higher IF, but it is easier to stop using Pocock design.
- Compare them using average information (similar to average sample size) needed for the alternative
- If H_0 is true, the chance that H_0 will be rejected will be small (α is usually taken to be 0.05). So the chance the trial will be stopped is small too (at most α). So the average information under H_0 will be very close to the MI

- For example, if $K = 5$, $\alpha = 0.05$, power $= 90\%$ to detect an alternative of interest, then

Designs	Maximum information	Average information (H_A)
5-look Pocock	$I^{FS} \times 1.21$	$I^{FS} \times .68$
5-look O-F	$I^{FS} \times 1.03$	$I^{FS} \times .75$
Fixed-sample	I^{FS}	I^{FS}

  where

$$I^{FS} = \left\{ \frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}}{\Delta_A} \right\}^2$$

  is the information for fixed sample size design.

- Pocock designs required smaller sample size on average if H_A is true.

- **Remarks**:
  - ⋆ If you want a design which, on average, stops the study with less information when there truly is a clinically important treatment difference, while preserving the level and power of the test, then a Pocock boundary is preferred to the O-F boundary.
  - ⋆ By a numerical search, one can derive the "optimal" shape parameter Φ which minimizes the average information under the clinically important alternative Δ_A for α, K, and power $(1-\beta)$. For example, when $K = 5$, $\alpha = .05$ and power of 90% the optimal shape parameter $\Phi = .45$, very close to the Pocock boundary (Wang and Tsiatis, 1987, *Biometrics*).
  - ⋆ However, the designs with smaller average information under H_A requires more information if the null hypothesis were true.
  - ⋆ Most clinical trials with a monitoring plan seem to favor more "conservative" designs such as the O-F design.

Statistical Reasons

1. Historically, most clinical trials fail to show a significant difference; hence, from a global perspective it is more cost efficient to use conservative designs (such as O-F design)
2. Even a conservative design, such as O-F, results in a substantial reduction in average information, under the alternative H_A, before a trial is completed as compared to a fixed-sample design (in our example .75 average information) with only a modest increase in the maximum information (1.03 in our example).

Non-statistical Reasons

3. In the early stages of a clinical trial, the data are less reliable and possibly unrepresentative for a variety of logistical reasons. It is therefore preferable to make it more difficult to stop early during these early stages.
4. Psychologically, it is preferable to have a nominal p-value at the end of the study which is close to .05. The nominal p-value at the final analysis for the 5-look O-F test is .041 as compare to .016 for the 5-look Pocock test. This minimizes the embarrassing situation where, say, a p-value of .03 at the final analysis would have to be declared not significant for those using a Pocock design.

III. Steps in the design and analysis of group-sequential tests with equal increments of information

Design

1. Decide the maximum number of looks K and the boundary Φ. K does not have to be very large.

Table 5: *O'Brien-Fleming boundaries* ($\Phi = 0$); $\alpha = .05$, *power=.90*

	Maximum	Average
K	Information	Information (H_A)
1	I^{FS}	I^{FS}
2	$I^{FS} \times 1.01$	$I^{FS} \times .85$
3	$I^{FS} \times 1.02$	$I^{FS} \times .80$
4	$I^{FS} \times 1.02$	$I^{FS} \times .77$
5	$I^{FS} \times 1.03$	$I^{FS} \times .75$

2. Compute I^{FS}, then translate it to the sample size or number of events.

3. Find the inflation factor $IF(\alpha, K, \Phi, \beta)$ and get

$$MI = I^{FS} \times IF(\alpha, K, \Phi, \beta).$$

Also calculate the maximum sample size or maximum number of events.

Analysis

4. Conduct data analysis after equal increment of MI/K information. This can be achieved by monitoring $[se\{\hat{\Delta}(t)\}]^{-2}$, although in practice, this is not generally how the analysis times are determined.

5. At the j-th interim analysis, the standardized test statistic

$$T(t_j) = \frac{\hat{\Delta}(t_j)}{se\{\hat{\Delta}(t_j)\}},$$

is computed using all the data accumulated until that time and the null hypothesis is rejected the first time the test statistic exceeds the corresponding boundary value.

Note: The procedure outlined above will have the correct level of significance as long as the interim analysis are conducted after equal increments of information.

However, in order for this test to have the desired power to detect Δ_A, it must be computed after equal increments of statistical information

MI/K where

$$MI = \left\{ \frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}}{\Delta_A} \right\}^2 IF(\alpha, K, \Phi, \beta).$$

If the initial guesses on the nuisance parameters were correct, then we would have the right power. Otherwise the study may be underpowered or overpowered.

We should monitor

$$[se\{\hat{\Delta}(t_j)\}]^{-2}$$

to see if it deviates significantly from the required information

$$j \times MI/K.$$

This helps detect the problem and fix it at the early stage.