Introduction to Clinical Trials

Lecture 6: Study Monitoring I

Giorgos Bakoyannis, PhD
Associate Professor
Department of Biostatistics and Health Data Science
Indiana University Indianapolis

Notes by Prof. Constantin T. Yiannoutsos, City University of New York

Monitoring of treatment effects

A clinical trial should not continue only by virtue of the fact that it has begun. Establishing that the ethical considerations that were present at its initiation continue to be present at every point in its implementation is paramount.

In this lecture we discuss what constitutes appropriate monitoring of a clinical trial. We focus, as in the textbook, on randomized comparative trials with a single primary endpoint.

Early stopping of a clinical trial

There are two decision landmarks for monitoring a clinical trial: when to terminate accrual in the study and when to disseminate the results. While these would be the same in many situations, accrual and follow-up time can frequently be manipulated in some trial designs (e.g., in studies of time to an event).

The decision to stop is frequently viewed as "symetrical". That is, whether treatment A is better than treatment B or the converse. However, if A is the standard and B the experimental treatment, a study can be stopped when,

- B cannot be shown to be better than A
- B is shown to be worse than A ("futility")
- B is much better than A

Stopping guidelines should thus reflects the inherent "asymmetry" of the implications of the final study result.

Reasons for early stopping of a clinical trial

The following are some reasons to stop a clinical trial early

- Treatments are found to be different by experts
- Treatments are found to be not different by experts
- Side effects are too severe to continue in light of benefits
- Accrual too slow to complete study in a timely fashion
- The data are of poor quality
- Definitive information about the treatment becomes available making the study unnecessary or unethical
- The scientific questions are no longer important
- Adherence to treatment is unacceptably poor
- Resources to perform the study are lost or are no longer available
- The study integrity has been undermined by fraud or misconduct

Competing interests in the decision to stop

Frequently there are competing interests in stopping versus continuing a clinical trial. There are pressures to terminate a trial as soon as possible to minimize the size and duration of the study and the number of patients receiving an inferior treatment as well as disseminating the results.

On the other hand, benefits of longer, larger trials lead to increased precision of the estimates, increased power, examining important subgroups, and gathering information on secondary endpoints.

This tension reflects the needs of the collective good versus those of the individual patient (individual good). There is an ethical mandate to continue the trial until it provides a standard of evidence appropriate for the setting.

Monitoring committees

An area where all components of monitoring comes together is with monitoring committees. These are called "Data Safety Monitoring Boards (DSMB) or Committees (DSMC) or Treatment Effects Monitoring Committees (TEMC).

In the following we will use the term "DSMB" instead of "TEMC", as the textbook does, because of the widespread use of the former term (despite the fact that the latter term is more accurately descriptive of the role of these committees).

Functions performed by monitoring committees

These monitoring committees perform a number of important oversight functions:

- Assist the study team with study protocol design
- Consider data quality and timeliness
- Review drug toxicity and adverse events (patient safety)
- Assess treatment efficacy
- Provide guidelines about the continuation of the study, its modification or early closing and dissemination of the results

Note: Monitoring committees serve an advisory role. These committees cannot make executive decisions about the studies.

Reasons for the popularity of the DSMB

DSMB have become very popular as ways to monitor clinical trials. Some of the reasons for this include:

- A workable mechanism for protecting the interests of the patients while preserving the integrity of the trial
- The DSMB is intellectually and financially independent from the study investigators and ensure the appearance of objectivity
- Many trial sponsors support or require the DSMB mechanism. For example,
  - The NCI policy is given at http://deainfo.nci.nih.gov/grantspolicies/datasafety.htm
  - The NIH policy is included at http://grants.nih.gov/grants/guide/notice-files/not98-084.html
  - The FDA policy is igiven at http://www.fda.gov/OHRMS/DOCKETS/98fr/01d-0489-gdl0003.pdf

Relationship of the DSMB to the investigators

There are two models of the structural relationship of the DSMB to the investigators.

- The DSMB is advisory to the trial sponsors
- The DSMB is advisory to the study investigators

In the former model, has the advantage that the decision of the DSMB will be free of investigator opinion. The disadvantage is that the study sponsor may not be the best filter of the DSMB decisions and transmit important information to the investigators.

The latter model is preferable, especially if this is done through a study steering committee. Thus, guidance on a number of important ethical issues will be efficiently transmitted to the investigator, the person most closely responsible for the safety of the patients.

Membership of the DSMB

DSMB usually consist of 3 to 10 members, based on the complexity of the trial and the issues involved. The DSMB must include experts but also it must include people that are experienced in clinical trials (which may not be experts in the specific area). A statistician is also an important member of the DSMB.

The DSMB should also involve a clinical trial investigator. While investigators affiliated with the trial are important to provide a unique perspective to the rest of the Board with respect to appropriately summarized information on patient safety and data quality, their participation in discussions on treatment efficacy and decisions to stop the trial early is inappropriate. Usually an “open” session that includes study investigators and a “closed” session, excluding study investigators, take place during the DSMB meeting.

Objectivity versus expertise in choosing DSMB members

When deciding whom to include as a DSMB member, it frequently happens that the investigators most knowledgeable of the field are the ones that may have less objectivity in rendering decisions. Thus, there is frequently a conflict between expertise and objectivity.

Ethical considerations however prescribe that expertise should be chosen over objectivity when experts can be relied upon to be reasonably objective in the collective sense and employ objective methods.

The most serious lack of expertise is ignorance of the study protocol, or lack of interaction with the patients. This suggests that one of the principal investigators should be included as a member of the DSMB as long as they do not become privy of "unblinded interim comparisons" from the study.

Blinding (masking)

Masking or blinding of DSMB members from treatment allocation has been recommended. In these cases, dummy treatment indicators are provided to the Board while the DSMB has the ability to request unblinding during the meeting.

While this could conceivably protect the integrity of the study, it is not a good idea because the Board might wrongly attribute unexpected efficacy trends in the wrong direction (consider the fact that a study should not be stopped only when the experimental treatment is demonstrably worse than the standard treatment).

There is inherent asymmetry in the nature of the decisions of the study (i.e., the study should be stopped if the experimental therapy is either much better or there is evidence of being worse than the standard therapy).

The DSMB review

During its meeting, the DSMB considers a number of components of the evolving study:

- *Baseline compatibility*

  The DSMB assesses the compatibility of various ancillary factors between the treatment groups
- *Review design assumptions*

  Several assumptions made during the design of the study such as accrual or dropout rate must be considered to ensure that the study can be completed in a timely manner. In addition, the continued availability or resources, principally funds and drug availability must be ensured.
- *Data quality and timeliness*
  There is always lag in data submission from the sites to the data coordinating center. However, the rate of submitted forms should be very high (typically over 90%). Also, all the events as of the specified cutoff date are included in the report, even if additional data submissions are required past the cutoff date to accomplish this.

The DSMB review (continued)

- *Patient eligibility and protocol deviations*
  All considerations pertaining to the establishment of patient eligibility must be performed and documented in the database. Treatment adherence is also an important consideration to ensure compliance with the protocol. Adherence can break down when
  - Side effects are serious
  - Patient inability to tolerate the treatment
  - Poor quality control is exercised
  - Pressures external to the study mount (e.g., availability of other treatments)

The DSMB review
Review of safety and toxicity data

Toxicity data must be reviewed carefully and in the context of the particular study. For example, mild events that can be reversed with dose modifications are not of concern.

On the other hand, serious or fatal events might be unacceptable in a study of healthy patients that otherwise have long life expectancy.

Finally, serious adverse events might be acceptable in AIDS studies or cytotoxic clinical trials in cancer.

Note: Events related to pre-existing conditions (e.g., death from cancer in a cancer trial) do not constitute an adverse event (although they have great significance in the assessment of efficacy of the study treatment).

The DSMB review
Review of efficacy comparisons

Efficacy considerations are probably the most important issues assessed by the DSBM for a number of reasons such as ethics and resource utilization.

Data quality and baseline comparability of threatment groups will be assessed, as mentioned earlier, and statistical guidelines will be very helpful in assisting the DSMB in its decision.

One complication of the decision process is the frequently observed situation of convincingly increased efficacy accompanied by also increasing toxicity events.

The DSMB review
Specific questions addressed by the DSMB

The DSMB, in its effort to assist the investigators with information to effectively carry out the study, will address the following questions:

- *Should the study continue?*
  Usually, trials provide much more information in addition to whether one treatment is more effective than the other. Thus, closing a study and, potentially closing the window of opportunity of conducting comparative studies in this area, must be weighed carefully against the individual good of study participants
- *Should the study protocol be modified?*

The DSMB review
Specific questions addressed by the DSMB

A number of aspects of the protocol may need modification after experience with the study. These include:

- Changes with dosing due to AE
- Change the frequency and timing of diagnostic tests
- Discontinuation of one or more arms or treatment combinations in multi-arm studies
- Modification of consent documents
- Improvement in data quality and timeliness
- Change in treatment and eligibility criteria to enhance adherence and increase accrual

The DSMB review
Specific questions addressed by the DSMB

- *Are additional views of the data required?*
  The DSMB may request additional analyses to help its members in their decision process.
- *Should the DSMB meeting schedule be modified?*
  Interesting trends in the data might prompt the DSMB to consider meeting more frequently. Statistical monitoring plans must be flexible enough (see subsequent discussion on this issue) to accommodate modified or emergency meetings by the Board
- *Are there other recommendations by the DSMB?*
  The DSMB is a group of objective experts interested with the best interests of all parties associated with the study. In this capacity the Board frequently makes a number of ancillary recommendations (e.g., measures to increase accrual, changing its own membership to ensure the presence of appropriate expertise, etc.)

Section 1

Statistical considerations in study monitoring

Background

Suppose that we compare the means of two groups. At the end of the study, and after $2N$ total subjects have been enrolled, we compute the statistic

$$Z_N = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2\sigma^2}{N}}} = \frac{S_N}{\sqrt{\nu_N}}$$

where $D_i = X_{1i} - X_{2i}$, $S_N = \sum_{i=1}^{N} D_i$ and $\nu_N = \mathrm{Var}(S_N) = 2N\sigma^2$.

Here we assume a balanced subject allocation, i.e., $N_1 = N_2 = N$ and a common variance $\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \sigma^2$.

Introducing interim monitoring of the study

Now consider carrying out an interim analysis of the data when $2n$ subjects have been enrolled and that we compute an interim z-score

$$Z(t) = \frac{Sn}{\sqrt{\nu_n}}$$

where $S_n = \sum_{i=1}^{n} D_i$ and $\nu_n = \mathrm{Var}(S_n) = 2n\sigma^2$ and t is the trial fraction

$$\begin{aligned} t &= \frac{\nu_n}{\nu_N} \\ &= \frac{2n\sigma^2}{2N\sigma^2} = \frac{n}{N} \end{aligned}$$

Note that t is the study *fraction* (i.e., how far through the study we are at any point). In this situation, $Z(1) = Z_N$.

The problem with multiple interim analyses

Consider a study of sample N where the treatment comparison is based on a one-sided test. Then, the null hypothesis (with one-sided alternative $H_A : \mu_1 > \mu_2$) is rejected when $Z(1) \geq z_{1-\alpha}$ where α is the type-I error rate.

Consider what would happen if we introduced k analyses at trial fractions $\tau_1, \tau_2, \cdots, \tau_k = 1$ and stopped the study any time that the interim z-score $Z(\tau_j) \geq z_{1-\alpha}$ (for $j = 1, 2, \cdots, k$).

The probability that $Z(1) \geq z_{1-\alpha}$ is

$$\begin{aligned} P\left(\cup_{j=1}^{k}\{Z(\tau_j) \geq z_{1-\alpha}\}\middle|H_0\right) &= 1 - P\left(\cap_{j=1}^{k}\{Z(\tau_j) \leq z_{1-\alpha}\}\middle|H_0\right) \\ &\leq 1 - (1-\alpha)^k \end{aligned}$$

under the null hypothesis. The less than or equal sign results since $\mathrm{Cov}\{Z(\tau_j), Z(\tau_{j'})\} \geq 0$.

Sampling to a foregone conclusion

The implication of this on the Type-I error is given in the following table where the probability that a statistically significant difference will be observed based on the number of interim analyses k is as follows:

k	$\alpha = 0.01$	$\alpha = 0.05$
1	.0100000	.0500000
2	.0199000	.0975000
3	.0297010	.1426250
4	.0394040	.1854938
5	.0490099	.2262191
6	.0585199	.2649081
7	.0679347	.3016627
8	.0772553	.3365796
9	.0864828	.3697506
10	.0956179	.4012631
20	.1820931	.6415141
50	.3949939	.9230551
100	.6339676	.9940795
1000	.9999568	1.0000000
∞	1.0000000	1.0000000

Thus, the Type-I error typically gets inflated as the number of interim analyses increases.

Intuition

Let's for a second decipher the dense mathematical formulas from the previous slide. We read $P\left(\cup_{j=1}^{k}\{Z(\tau_j) \geq z_{1-\alpha}\}\middle|H_0\right)$ as "the probability that $Z(\tau_1) \geq z_{1-\alpha}$ OR $Z(\tau_2) \geq z_{1-\alpha}$ and so on up to the kth analysis, if the null hypothesis is true."

Through logical rules, this is equivalent to the complementary probability $1 - P(\cap_{j=1}^{k}\{Z(\tau_j) \leq z_{1-\alpha}\}|H_0)$, which means the opposite of not rejecting the null hypothesis in any of the $k-1$ interim and the final (kth) analysis.

The basic intuition is that every time we carry out an additional analysis we add the possibility of making a mistake and reject the null hypothesis when it is correct (i.e., making a type-I error). This inflates this type of error beyond the maximum acceptable α level.

Boundaries that adjust for multiple analyses

To adjust for multiple interim analyses we seek boundaries $c_1, \cdots, c_k$ such that

$$P\left(\cup_{j=1}^{k}\{Z(\tau_j) > c_j\}\Big|H_0\right) = \alpha$$

or

$$P\left(\cup_{j=1}^{k}\{Z(\tau_j) < -c_j\}\Big|H_0\right) = \alpha$$

for one-tailed tests and

$$P\left(\cup_{j=1}^{k}\{|Z(\tau_j)| > c_j\}\Big|H_0\right) = \alpha$$

for two-tailed tests.

In plain language, we seek boundaries that the total probability of rejecting in any of the analyses (both interim and final) is equal to α under the null hypothesis.

Section 2

Sequential boundaries

The Haybittle boundary

Haybittle (Br J Radiol, 1971) proposed the following procedure:

- Use critical value $c = 3$ at the interim analyses
- Use critical value $c = 1.96$ at the final analysis

The author showed by simulation that the alpha level of this procedure does not overly inflate the Type-I error if the number of interim analyses are not too numerous.

The Haybittle boundary
Advantages and disadvantages

The advantages of this approach are that

- It is simple to implement
- The final test is the same as in the case of no monitoring (note that $c = 1.96 = z_{1-\alpha/2}$ for $\alpha = 0.05$).

The disadvantages of the approach are that

- It is very difficult to stop before the final analysis (note that $c = 3$ is equivalent to a p-value of $p = 0.0013$).
- The procedure still produces a minor inflation of the Type-I error

The Haybittle boundary
Improvements

The latter disadvantage of the inflation of the Type-I error can be fixed by use of the Bonferroni procedure. That is, we proceed as follows:

- At first $k-1$ analyses use $p=0.001$, that is, reject the null hypothesis at the ith analysis if $|Z(t)|>3.29$.
- Use Bonferroni to fix last critical value.

For example, if we have $k=5$ (i.e., 4 interim analyses before the final), we use significance level 0.05 -4(0.001) $=$ 0.046 at the final analysis (i.e., reject the null if $|Z(1)|>1.995$).

This is a very nice procedure because it can be universally applied as long as p-values can be computed.

The Pocock procedure

Pocock (Biometrika, 1977) suggested the following procedure, based on equally-spaced analyses (i.e., analyses performed at times $t = j/k$ where $j = 1, \cdots, k$).

Determine c such that

$$P\left(\cup_{j=1}^{k}\{Z(j/k) > c\}\middle|H_0\right) = \alpha$$

The Pocock boundary
Advantages and disadvantages

The advantages of this procedure are

- It is a natural extension of the case without monitoring (i.e., going from $z_{1-\alpha}$ to c but still using a constant boundary)
- Uses same degree of evidence at each analysis
- c is typically smaller than the Haybittle boundary, so the Pocock procedure can stop earlier

The main problem with the Pocock approach is that the p-value at the final analysis should be very low to reject the null hypothesis.

What's more, Pocock now recommends against his own procedure!

Boundaries for the Pocock procedure

Table: Two-tailed boundaries for the Pocock procedure (Proshan, Lan & Wittes, 2006)

# of looks	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
1	2.576	1.960	1.645
2	2.772	2.178	1.875
3	2.873	2.289	1.992
4	2.939	2.361	2.067
5	2.986	2.413	2.122
10	3.117	2.550	2.270
20	3.225	2.672	2.392
∞	∞	∞	∞

The conclusion from the table is that the critical values increase significantly with the increase of the number of interim analyses.

Brownian motion

Now consider the related quantity $B(t) = \frac{S_n}{\sqrt{\nu_N}}$. The interim z-score $Z(t) = \frac{S_n}{\sqrt{\nu_n}}$ at time $t = n/N$ is related to $B(t)$ by the equation

$$\begin{aligned} Z(t) &= \frac{S_n}{\sqrt{\nu_n}} \\ &= \left(\frac{\sqrt{\nu_N}}{\sqrt{\nu_n}}\right)\frac{S_n}{\sqrt{\nu_N}} = \sqrt{\frac{N}{n}}B(t) = \frac{B(t)}{\sqrt{t}} \end{aligned}$$

The quantity $B(t)$ is related to the so-called "Brownian motion" (a stochastic process with a number of characteristics that help in the modeling of random events). While the Brownian motion is the source of fundamental theoretical results in study monitoring, we will not consider it further as it is beyond the scope of this course.

The O'Brien-Fleming procedure

O'Brien and Fleming (Biometrics, 1979) proposed a related procedure with that of Pocock. The critical difference of the O'Brien-Fleming procedure is that the boundary is related to the $B(t)$ quantity rather than the interim z-score $Z(t)$.

In other words, the O'Brien-Fleming boundary is such that

$$P\left(\cup_{j=1}^{k}\{B(j/k) > c\}\Big|H_0\right) = \alpha$$

Given the relationship between $B(t)$ and $Z(t)$ the above procedure is equivalent to one in terms of $Z(t)$ as follows:

$$P\left(\cup_{j=1}^{k}\left\{Z(j/k) > c/\sqrt{j/k}\right\}\middle|H_0\right) = \alpha$$

Boundaries of the O'Brien-Fleming procdure

Table: Two-tailed boundaries for the O'Brien-Fleming procedure (Proshan, Lan & Wittes, 2006)

# of looks	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
1	2.576	1.960	1.645
2	2.580	1.977	1.678
3	2.595	2.004	1.710
4	2.609	2.024	1.733
5	2.621	2.040	1.751
10	2.660	2.087	1.801
20	2.695	2.126	1.842
∞	2.807	2.241	1.960

The O'Brien-Fleming procedure
Advantages and disadvantages

The big advantage of the O'Brien-Fleming procedure is that, at the final analysis, the alpha level is close to the original alpha level.

This is counter-balanced by the fact that the O-B procedure will stop the trial more infrequently early (when evidence is more limited) compared to the Pocock procedure.

This latter consideration may not be very problematic as, intuitively, there is great resistance for stopping early when information is limited.

Example: A study with $k = 5$ analyses
Boundaries of the 3 procedures

For a study with $k = 5$ total analyses, the three boundaries give the following critical values (Table 1) and corresponding p-value boundaries (Table 2):

Table: Two-tailed boundaries for the O'Brien-Fleming, Pocock and Haybittle-Peto procedures

# of looks	O'Brien-Fleming	Pocock	Haybittle-Peto
1	$\pm$ 4.562	$\pm$ 2.413	$\pm$ 3.290
2	$\pm$ 3.226	$\pm$ 2.413	$\pm$ 3.290
3	$\pm$ 2.634	$\pm$ 2.413	$\pm$ 3.290
4	$\pm$ 2.281	$\pm$ 2.413	$\pm$ 3.290
5	$\pm$ 2.040	$\pm$ 2.413	$\pm$ 1.995

Example: A study with $k = 5$ analyses
p-values

The p-values corresponding to the boundaries Table 3 are given in the following Table:

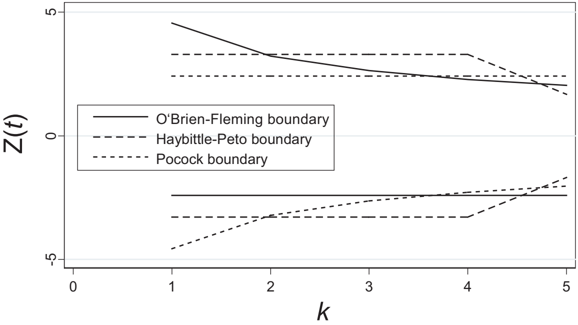
Table: Two-tailed p-values for the O'Brien-Fleming, Pocock and Haybittle- Peto procedures

# of looks	O'Brien-Fleming	Pocock	Haybittle-Peto
1	5.067e-06	0.016	0.001
2	0.001	0.016	0.001
3	0.008	0.016	0.001
4	0.023	0.016	0.001
5	0.041	0.016	0.046

Example: A study with $k = 5$ analyses
Graphical representation

The three boundaries are shown in the following Figure:

Figure: Two-tailed critical values for the O'Brien-Fleming, Pocock and Haybittle-Peto procedures



Statistical Information

We have mentioned that the definition of "statistical information" for parameter δ is the inverse of the variance of its estimator $\hat{\delta}$ i.e.,

$$I = \mathrm{Var}(\hat{\delta})^{-1}$$

The trial fraction $t = n/N$ is also the information fraction during the interim analysis. To see this, consider what the total information at the end of the study (i.e., $2N$ subjects have been accrued) is

$$I_N = (\nu_N)^{-1} = \frac{N}{2\sigma^2}$$

while the information at the interim analysis (and after $2n$ subjects have been accrued) is

$$I_n = (\nu_n)^{-1} = \frac{n}{2\sigma^2}$$

Trial fraction versus the information fraction

So the trial fraction t is

$$\begin{aligned} t &= n/N \\ &= \left(\frac{n}{2\sigma^2}\right)\left(\frac{2\sigma^2}{N}\right) \\ &= \frac{\nu_N}{\nu_n} \\ &= \frac{I_n}{I_N} \end{aligned}$$

This means that the trial fraction is also the information fraction.

Example: Ischemia trial

In an ischemia trial (Proshan, Lan & Wittes, 2006) that expects to enroll 200 subjects per arm, suppose that we have $n_T = 82$ control subjects and $n_C = 86$ treatment subjects.

The estimator of the difference of the proportion of events is $\hat{\delta} = \hat{p}_C - \hat{p}_T$ and, under the null hypothesis one can use a pooled estimate of the common proportion $\hat{p} = \frac{n_T \hat{p}_C + n_C \hat{p}_T}{n_C + n_T}$.

Ischemia trial: Information

The current information is

$$I_n = \mathrm{Var}(\hat{\delta}) = \left\{\hat{p}(1-\hat{p})\left(\frac{1}{82}+\frac{1}{86}\right)\right\}^{-1}$$

The total information at the end of the study is

$$I_N = \left\{\hat{p}(1-\hat{p})\left(\frac{2}{200}\right)\right\}^{-1}$$

The information fraction is

$$\frac{I_n}{I_N} = \frac{(82)(86)[168\hat{p}(1-\hat{p})]}{100/\hat{p}(1-\hat{p})} = 0.42$$

Information in studies of time to failure

We discussed information being proportional to the sample size in studies involving comparisons of means or proportions. But what about survival ("time-to-event") studies?

It turns out that information in this case is proportional to the number of events d and D associated with n and N participants at the interim and final analysis respectively.

Thus the information fraction at interim analysis after d events have occurred is

$$t \approx \frac{d}{D}$$

One problem with these studies is that one is not certain what the event rate will be even if the study is fully accrued (and, if the study stops early, it will never be confirmed whether the event rate were close to the one assumed at the time of the study design; we will discuss this further later in this lecture).

Information versus calendar time

Interim monitoring becomes complicated in survival analysis studies because information is proportional to the number of events and not the overall sample size. This is no problem in a study with no monitoring, because one may continue until all events have been observed.

However, usually there is a maximum duration of the study (i.e., total accrual time plus total follow-up time after completion of patient accrual).

It is thus difficult to figure out in practice where exactly in the information time you are at each interim analysis. Calendar time, by contrast, is unambiguous but may not correspond exactly (or even closely) with information time.

Section 3

Spending functions

Spending functions

Spending functions show the way that the total alpha is “spent” through the interim and final analyses. They are necessary for the following reasons:

- The Pocock and O'Brien-Fleming boundaries require equal spacing of the analyses but DSMB meet when the schedules permit
- Analysis times may not be easily predictable in advance
- Extra analyses may be scheduled during the implementation of the study

The seminal reference for this methodology is the paper by Lan & DeMets (Biometrika, 1983) that showed that boundaries can be computed without knowing the timing of the analyses in advance.

Alpha spending functions

An alpha spending function $\alpha(t)$, with $\alpha(0) = 0$ and $\alpha(1) = \alpha$ of the form

$$\alpha(t) = P\left(\cup_{j=1}^{k}\{|Z(t)| > c_j\}|H_0\right)$$

For a given schedule of analyses $\tau_1, \cdots, \tau_k$, this splits α in probabilities $\tilde{\alpha}(\tau_j)$, $j = 1, \cdots, k$,

$$\tilde{\alpha}(\tau_1) = P\left(|Z(\tau_1)| > c_1|H_0\right)$$

and for $j = 2, \cdots, k$

$$\tilde{\alpha}(\tau_j) = P\left(|Z(\tau_1)| \leq c_1, \cdots, |Z(\tau_{j-1})| \leq c_{j-1}, |Z(\tau_j)| > c_j|H_0\right)$$

with $\sum_{j=1}^{k} \tilde{\alpha}(\tau_j) = \alpha$. These probabilities are calculated by numerical methods (Armitage, McPherson & Rowe, JRSS A', 1969).

Example: Equal alpha spending over $k = 5$ interim analyses

Consider the situation where sample size calculations have determined that, with $\alpha = 0.05$ and $\beta = 0.2$ the requisite sample size for the fixed (i.e., one-analysis) design is $N = 100$ per group.

Suppose that we want to carry out $k = 5$ total analyses (i.e., four interim and one final analysis) at equal time points (i.e., after $n_1 = 20$ per group, $n_2 = 40$ per group and so on) and we want to spend $\alpha = 0.05$ equally, over these $k = 5$ analysis. In other words, we want

$$\begin{array}{c} \alpha(1) = 0.01 \\ \alpha(2) = 0.01 \\ \alpha(3) = 0.01 \\ \alpha(4) = 0.01 \\ \hline \alpha(5) = 0.01 \end{array}$$

Example: Equal alpha spending over $k = 5$ interim analyses *Bounds*

The critical values $c_1, \cdots, c_5$ are given from the following output:

```
Symmetric two-sided group sequential design with
80 % power and 2.5 % Type I Error.
Spending computations assume trial stops
if a bound is crossed.



  Analysis   N    Z   Nominal p Spend
         1  23 2.58    0.0050 0.005
         2  46 2.49    0.0064 0.005
         3  69 2.41    0.0080 0.005
         4  92 2.34    0.0097 0.005
         5 115 2.28    0.0114 0.005
     Total                    0.0250
```

This is the α spent for the scenarios where the experimental treatment is better. An equal amount (i.e., $\alpha/2 = 0.025$) is spent for scenarios leading to the standard treatment being superior (“futility”). Note also that the sample size N has been inflated. More on this later.

Example: Equal alpha spending over $k = 5$ interim analyses *Pictorial representation*

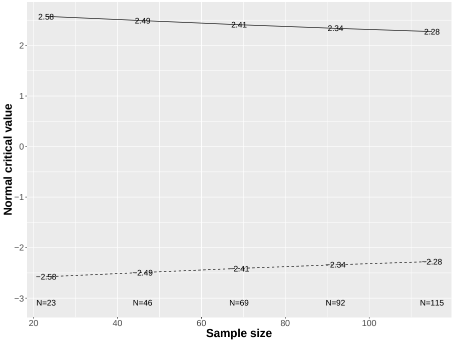


Figure: Boundaries for equal spending of the alpha level over $k = 5$ analyses

Continuous spending functions

The advantage of the alpha spending approach is that they can accommodate arbitrary interim analysis schedules.

The limitation of this approach is that the schedule of the interim analyses must be known *a priori*.

In a seminal paper, Lan & DeMets[1], proposed a continuous spending function approach. They suggested that alpha spending functions can be arbitrary as long as

- $\alpha(0) = 0$
- $\alpha(1) = \alpha$

Note: The main advantage of this approach is that the analyses do not have to be equally-spaced and the timing *or the number* of the analyses do not have to be known in advance!

[1] Lan KKG and DeMets DL. Discrete sequential boundaries for clinical trials, *Biometrika*, **70**, 659–663. 1983

O-F or Pocock-like spending functions

The main advantage of the continuous spending-function approach is that interim analyses can be undertaken at any point during the implementation of the study.

Two spending functions, which are in the vein of the Pocock and the O'Brien-Fleming methods are given below:

- Pocock-like spending function

$$\alpha_P(t) = \alpha \log\{1 + (e-1)t\}$$

- O'Brien-Fleming-like spending function

$$\alpha_{OB}(t) = 2\left\{1 - \Phi\left(z_{1-\alpha/2}/\sqrt{t}\right)\right\}$$

Example: O-F and Pocock spending functions for $k = 5$ *Cum. alpha level spent*

For example, with $k = 5$ the Pocock and O'Brien-Fleming spending functions are

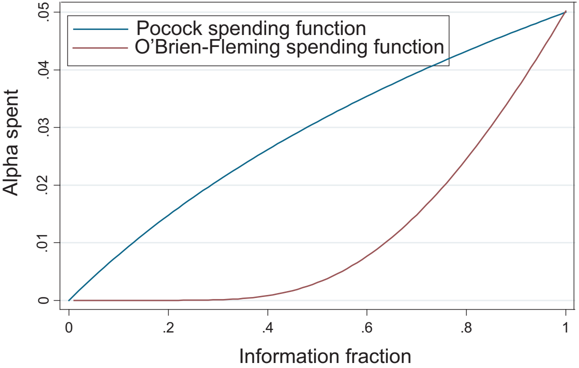
	Cumulative α spent		Bounds	
Information fraction	Pocock	O-F	Pocock	O-F
0.2	0.016	0.000	$\pm$ 2.41	$\pm$ 4.56
0.4	0.028	0.001	$\pm$ 2.41	$\pm$ 3.23
0.6	0.037	0.009	$\pm$ 2.41	$\pm$ 2.63
0.8	0.044	0.026	$\pm$ 2.41	$\pm$ 2.28
1.0	0.050	0.050	$\pm$ 2.41	$\pm$ 2.04

So, although both functions spend the same amount of α by the end of the study, the Pocock spending function spends alpha much faster than the O'Brien-Fleming spending function.

Spending functions
Pictorial representation

The cumulative rate of alpha spending of the two spending functions is given pictorially in the following figure:

Figure: Cum. α spending in Pocock and O'Brien-Fleming spending functions.



The Hwang, Shih & DeCani family of spending functions

In a 1990 paper, Hwang, Shih & DeCani[2] introduced the following general family of alpha spending functions:

$$\alpha(t) = \left\{ \begin{array}{cl} \alpha(1 - e^{-\gamma t})/(1 - e^{-\gamma}) & \mathrm{if} \gamma \neq 0 \\ \alpha t & \mathrm{if} \gamma = 0 \end{array} \right.$$

[2] Hwang IK, Shih WJ and DeCani JS. Group sequential designs using a family of Type I error probability spending functions, *Stat Med*, **9**:1439–1445. 1990

The HSD family of spending functions

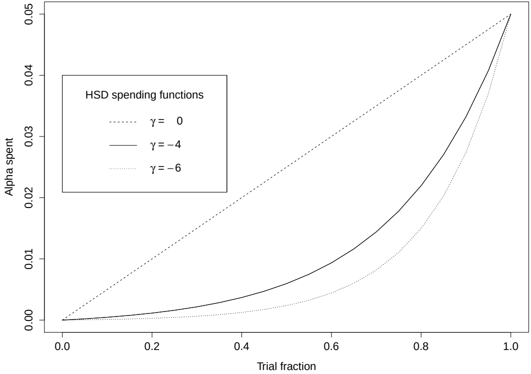


Figure: Hwang-Shih-DeCani family of spending functions

'From the figure it is clear that alpha is spent more quickly the larger the value of γ.

Example: Study with $k = 3$ analyses at unequal times

Suppose that you are carrying out two interim analyses, one at $\tau_1 = 0.2$ and one at $\tau_2 = 0.5$ and you are using the simple (linear) spending function $\alpha(t) = \alpha t$ with $\alpha = 0.5$.

This means that the critical values for this scenario will be

Fraction (t)	α spent	$Z(t)$
0.20	0.0100	2.58
0.50	0.0250	2.38
1.00	0.0500	2.14

What if, after the second interim analysis, you wanted to add a third interim look, say, at $\tau_3 = 0.75$ to the schedule?

Adding an interim analysis

The beauty of the spending-function approach is that, at any point, we only need to concern ourselves with what has happened so far and not on information about the remainder of the study. However, the addition of the interim analysis does have an impact on the boundary at the final analysis.

If we want to add an interim analysis at $\tau_3 = 0.75$ we realize that the alpha spent will be $\alpha(\tau_3) = 0.0375$. The bounds are as follows:

Fraction (t)	α spent	$Z(t)$
0.20	0.0100	2.58
0.50	0.0250	2.38
0.75	0.0375	2.32
1.00	0.0500	2.24

Comments

Note the following:

- The critical bounds before the third analysis did not change
- The critical bound at the final analysis is higher ($Z_3(1) = 2.14$ versus $Z_4(1) = 2.24$ in the case of the three-analysis and four-analysis scenaria respectively). This is because of the additional alpha spent to carry out this additional interim analysis.
- While adding an interim analysis to the schedule is straightforward, this should be undertaken with extreme care, since the immediate result will be to raise the level of evidence (lower the p-value threshold) in the final analysis.

Recalibration of a study

Spending functions can be used to react to information as it comes in within an ongoing study.

Suppose we have a study and we are using the Pocock spending function $\alpha_P(t) = 0.05 \log\{1 + (e-1)t\}$ (Proshan, Lan & DeMets, 2006). Now suppose that the first analysis happened at the $t = 1/10$ fraction. So, at this look, you spent $\alpha_P(1/10) = 0.008$ (using z-score boundaries ± 2.655 from the Pocock spending function).

At the second analysis, for reasons we will discuss later, you figure that the first analysis was actually at the trial fraction $t = 0.20$.

Question: What do you do?

Recalibration of the study
Adjusting the error spending rate

The information fraction at the first analysis should have been $t = 0.2$ and you should have spent $\alpha_P(0.20) = 0.015$ but you spent only $\alpha_P = 0.008$. So you are spending alpha much more slowly than you expected when the study was designed. So we need to adjust the rate of alpha spending to catch up.

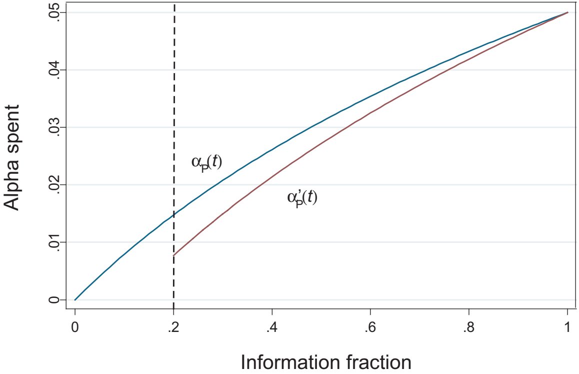
The alpha spent at the first look is gone. To adjust, you interpolate the spending function for $t \geq 0.20$ by

$$\alpha'_P(t) = 0.008 + \left(\frac{0.05 - 0.008}{0.05 - 0.015}\right)\{\alpha_P(t) - 0.015\}$$

Recalibration of the study
Pictorial representation

The two spending functions are given pictorially in the following figure:

Figure: Pocock-like alpha spending functions recalibrated after the first interim analysis



Impact of interim monitoring on sample size

In order to incorporate interim monitoring into a clinical trial the sample size will need to be inflated in order to achieve the same power.

This becomes immediately clear if we simply consider the Bonferroni adjustment for multiple comparisons.

The sample size for a single-test design with $\alpha = 0.05$, power $1 - \beta = 0.8$ and effect size $f = 1$ is $n = [2(1.96 - 1.282)]^2 \approx 43$.

When the number of tests is $k = 3$ we adjust the alpha level to $\alpha^* = \alpha/k = 0.0167$. Then the sample size everything else being equal is $n^* = [2(2.4 - 1.282)]^2 \approx 55$ individuals. This is an inflation of almost 28%.

Revision of sample size for the Pocock and O-F procedures

Table 5 shows the inflation of the sample size from a Pocock and an O'Brien-Fleming procedures with equally and not equally-spaced analyses:

Table: Sample size inflation for equally and not equally-spaced analyses for the O'Brien-Fleming and Pocock procedures

	Equispaced		non-equispaced	
k	O-F	Pocock	O-F	Pocock
2	1.004525	1.083919	1.013130	1.048586
3	1.011075	1.127630	1.019828	1.070271
4	1.015727	1.156074	1.023904	1.082788
5	1.019146	1.176742	1.026661	1.091020
6	1.021770	1.192787	1.028658	1.096883
7	1.023848	1.205804	1.030175	1.101290
8	1.025536	1.216702	1.031370	1.104736
9	1.026934	1.226042	1.032336	1.107509
10	1.028113	1.234192	1.033134	1.109795

The Pocock procedure results in significant sample-size inflation and the inflation is related to how quickly the alpha is spent. Notice that the O'Brien-Fleming routine is almost impervious to the timing of the analyses.

Section 4

Beta spending and stopping for futility

Beta spending

Akin to the idea of alpha spending we have the idea of beta spending. As before, for a series of $\tau_1, \cdots, \tau_k$ analyses, we have critical bounds $a_1, \cdots, a_k$ and $b_1, \cdots, b_k$ such that[3]

$$\tilde{\alpha}(\tau_1) = P(Z(\tau_1) > a_1 | H_0)$$

$$\tilde{\beta}(\tau_1) = P(Z(\tau_1) < b_1 | H_A)$$

for the first analysis and for $j = 2, \ldots, k$

$$\tilde{\alpha}(\tau_j) = P\left(\cap_{l=1}^{j-1}\{b_l \le Z(\tau_l) \le a_l\}, Z(\tau_j) > a_j | H_0\right)$$

$$\tilde{\beta}(\tau_j) = P\left(\cap_{l=1}^{j-1}\{b_l \le Z(\tau_l) \le a_l\}, Z(\tau_j) < b_j | H_A\right)$$

for analyses 2 through k with $\sum_{j=1}^{k} \tilde{\alpha}(\tau_j) = \alpha$ and $\sum_{j=1}^{k} \tilde{\beta}(\tau_j) = \beta$.

[3] In the sequel we will only deal with symmetric bounds, so it is implied that the lower bounds will be $-a_j$ and $-b_j$, $j = 1, \cdots, k$.

Example: A one-sided study with boundaries for futility

Suppose that we are designing a study with $k = 5$ analyses, to be carried out at the $\alpha = 0.05$ and with 90% power. Suppose also that we want to stop the study *both* if there is sufficient evidence to reject the null hypothesis as well as if there is evidence in favor of the null hypothesis (futility).

Using spending functions $\alpha(t) = \alpha t^\rho$ and $\beta(t) = \beta t^\rho$ with $\rho = 3$[4] we obtain the following upper and lower boundaries:

k	t	b_j	a_j	α spent	β spent
1	0.2000	-1.81629	3.35279	0.0004	0.0008
2	0.4000	-0.62004	2.75256	0.0032	0.0064
3	0.6000	0.24893	2.35028	0.0108	0.0216
4	0.8000	0.98426	2.01825	0.0156	0.0512
5	1.0000	1.68698	1.68698	0.0500	0.1000

The inflation of the sample size is about 4.8%.

[4] This is approximately the O'Brien-Fleming procedure.

One-sided alternative with boundaries for futility
Pictorial representation

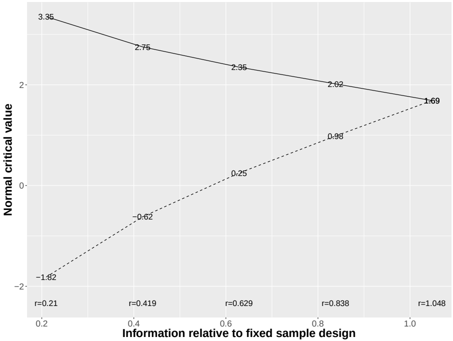


Figure: One-sided boundary with lower bound for futility

The trial continues while $b_j \leq Z(\tau_j) \leq a_j$, $j = 1, \cdots, k-1$.

Two-sided boundaries with futility bounds (inner wedge)

We can also have a situation where two-sided boundaries are generated with a two-sided region for futility as shown in the following Figure:

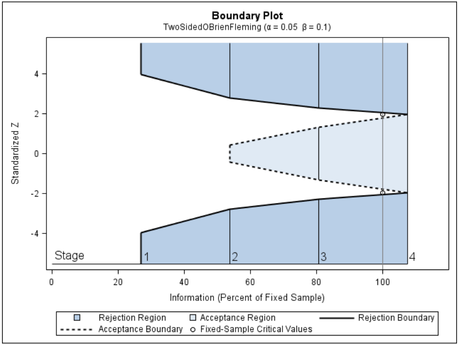


Figure: Two-sided boundaries with futility bounds (inner wedge) for an O-F study with four analyses

The unique nature of failure-time studies

In failure-time (survival) studies, the added complication, when designing them, is that the sample size, both for the fixed-sample and the study with interim analyses depends on the study duration and the accrual rate.

In the sequel, we will use two methods:

1. We will apply the Lachin & Foulkes [2] sample size method and extend it to group sequential design. This method fixes the duration of a study and varies enrollment rates to power a trial.
2. We also use the Lachin & Foulkes [2] basic power equation to compute sample size along the lines of Kim & Tsiatis [1] where enrollment rates are fixed and enrollment duration is allowed to vary to enroll a sufficient sample size to power a study.

Basic assumptions and fixed design sample size
Enrollment, randomization and dropout

When designing a failure-time study we first set up the enrollment and dropout information.

Usually enrollment rates are assumed to be uniform, but piece-wise uniform rates or other simplified patterns of patient recruitment can be accommodated by contemporary software.

We also need to provide information about the duration of the study. In this situation we fix the total duration of the study.

Finally, we would need the randomization ratio between the (two in this case) arms under comparison.

Basic assumptions and fixed design sample size
Failure parameters

Next we provide information about the median time to event in the control group as well as the accrual and dropout rate, hazard ratios under the null and alternate hypotheses for experimental therapy compared to control, and the desired Type I and II error rates.

Finally, we design a trial with no interim analyses under these assumptions.

Note that when calling `nSurv` in `R`, we transform the median time-to-event (m) to an exponential event rate (λ) with the formula

$$\lambda = \log(2)/m.$$

Example: Fixed design sample size for a failure-time trial

We consider the following parameters:

1. Planned duration of patient accrual is 24 months and the duration of follow-up is 12 months.
2. We consider piece-wise uniform enrollment rates
3. Median time to event in the control group $m = 12$ months
4. Exponential dropout rate per unit of time $\eta = 0.001$
5. Hypothesized experimental/control hazard ratio $\theta = 0.75$
6. Type-I error (1-sided) $\alpha = 0.025$
7. Type-II error (1-power) $\beta = 0.1$

Fixed design sample size

Using the Lachin & Foulkes[2] method we come up with

$$D = \frac{4(1.96 + 0.842)^2}{(\log 0.75)^2} = 380$$

events. Using the remaining information above, the total sample size is $N = 580$ patients are needed in the two arms ($N_1 = N_2 = 290$ per arm).

Adding the group sequential design

Now we move on to a group sequential design. We set up the number of analyses, timing and spending function parameters.

1. Number of analyses (interim + final), $k = 2$ with the interim analysis occurring at the 40% trial fraction.
2. We will use the Hwang-Shi-DeCani spending function for the efficacy bound with parameter $\gamma_\alpha = -10$.
3. For the futility bound, we will use the same family of spending functions with (beta) spending parameter $\gamma_\beta = 2$.

Group sequential design of a failure-time study
Output

The accrual rates are given in the following table:

Period (months)	Accrual rate
0-1	11.02665
1-3	16.53997
3-6	27.56662
6-24	44.10660

The upper and lower bounds are given in the following Table:

		Lower bounds			Upper bounds		
Analysis	N	Z	Nominal p	β spent	Z	Nominal p	α spent
1	241	0.71	0.7611	0.0637	3.84	0.0001	0.0001
2	603	1.96	0.9750	0.1000	1.96	0.0250	0.0250

Group sequential design of a failure-time study

The following Figure shows the above situation pictorially:

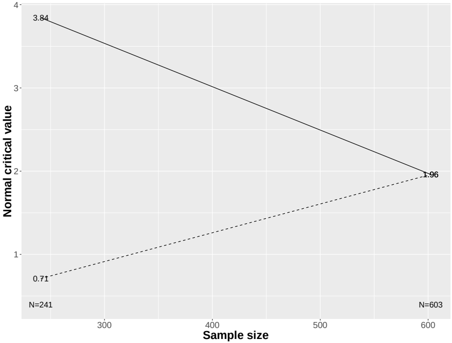


Figure: Group sequential design with one-sided futility bound

Monitoring a survival study using calendar time

Suppose that we have designed a study base on time-to-event (failure-time) data with maximum duration of six years and with a Pocock error spending function.

Then at the end of the first year (say) we are at $s = 1/6 = 0.17$ of the total calendar time but we may not be at $t = 1/6$ in terms of the final number of events.

If a DSMB review of the study is scheduled at the end of the first year, how can you proceed? (It is much more convenient to schedule meetings according to the yearly calendar than the study-fraction calendar).

One solution is to let $t = s$ and use a spending function $\alpha^*(s)$ (i.e., spend error at the rate of the calendar not the trial fraction).

Monitoring a survival study using calendar time (cont'd)

In the example above, at the end of the first year, the alpha to be spent should have been $\alpha^*(1/6) = 0.013$ (using the Pocock alpha spending function[5]). Then we can proceed identically as before by adjusting for this fraction.

This is an attractive way to do monitoring since the total number of events at the end is not necessarily known and thus we never really know t.

However, there is a big drawback: Events will be likely slow at the beginning, so it may actually be easier to stop the trial early using calendar versus information time.

[5] Note that $\alpha^*(0.17) = 0.05 \log(1 + (e-1)0.16) \approx 0.013$.

References

Kyungmann Kim and Anastasios A. Tsiatis.
Study duration for clinical trials with survival response and early stopping rule. *Biometrics*, 46:81–92, 1990.

John M. Lachin and Mary A. Foulkes.
Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics*, 42:507–519, 1986.