Group sequential monitoring of clinical trials

Advanced topics

Constantin T. Yiannoutsos, Ph.D.
Indiana University Fairbanks School of Public Health
cyiannou@iu.edu

4 July, 2022

Outline of this unit

In this unit we will present a number of advanced topics in the monitoring of clinical trials:

- Impact of monitoring on the sample size

Outline of this unit

In this unit we will present a number of advanced topics in the monitoring of clinical trials:

- Impact of monitoring on the sample size
- Using alpha spending functions for flexible monitoring

Outline of this unit

In this unit we will present a number of advanced topics in the monitoring of clinical trials:

- Impact of monitoring on the sample size
- Using alpha spending functions for flexible monitoring
- Monitoring information versus sample size

Outline of this unit

In this unit we will present a number of advanced topics in the monitoring of clinical trials:

- Impact of monitoring on the sample size
- Using alpha spending functions for flexible monitoring
- Monitoring information versus sample size
- Wade briefly in the special case of time-to-event studies to cover issues like

Outline of this unit

In this unit we will present a number of advanced topics in the monitoring of clinical trials:

- Impact of monitoring on the sample size
- Using alpha spending functions for flexible monitoring
- Monitoring information versus sample size
- Wade briefly in the special case of time-to-event studies to cover issues like
  - Re-calibration of a study

Outline of this unit

In this unit we will present a number of advanced topics in the monitoring of clinical trials:

- Impact of monitoring on the sample size
- Using alpha spending functions for flexible monitoring
- Monitoring information versus sample size
- Wade briefly in the special case of time-to-event studies to cover issues like
  - Re-calibration of a study
  - Information versus calendar monitoring of a study

Outline of this unit

In this unit we will present a number of advanced topics in the monitoring of clinical trials:

- Impact of monitoring on the sample size
- Using alpha spending functions for flexible monitoring
- Monitoring information versus sample size
- Wade briefly in the special case of time-to-event studies to cover issues like
  - Re-calibration of a study
  - Information versus calendar monitoring of a study
- Beta spednding functions

Impact of interim monitoring on the sample size

Impact of interim monitoring on sample size

All the methods we just described preserve the alpha level in a clinical trial.

However, when incorporating interim monitoring into a clinical trial the sample size will need to be inflated in order to maintain the same power

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

For example, the sample size for a single-test design with $\alpha = 0.05$, power $1 - \beta = 0.8$ and effect size $f = 1$ is

$$n = [2(1.96 - 1.282)]^2 \approx 43$$

individuals per group.

When the number of tests is $g = 3$ we adjust the alpha level to $\alpha^* = \alpha/g = 0.0167$. Then the sample size everything else being equal is

$$n^* = [2(2.4 - 1.282)]^2 \approx 55$$

per group. This is an inflation of almost 28%.

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

The following output shows the inflation of the sample size from a Pocock and an O'Brien-Fleming procedures with equally and not equally-spaced analyses[1]:

```
##    OF.equal Pocock.equal OF.unequal Pocock.unequal
## 2  1.007126     1.100082   1.007126       1.100082
## 3  1.016100     1.150639   1.022111       1.117384
## 4  1.022163     1.183142   1.029460       1.125741
## 5  1.026486     1.206603   1.033885       1.130813
## 6  1.029746     1.224739   1.036865       1.134271
## 7  1.032298     1.239408   1.039017       1.136804
## 8  1.034351     1.251660   1.040650       1.138752
## 9  1.036040     1.262142   1.041933       1.140305
## 10 1.037456     1.271277   1.042970       1.141576
```

The Pocock procedure results in significant sample-size inflation compared to the O-F procedure. The inflation of the sample size is also related to the timing of the analysis. Notice that the O'Brien-Fleming routine is almost impervious to the timing of the analyses.

[1] In the case of unequal analyses, all k analyses occur after the 50% trial fraction.

Intuition from alpha spending functions

We notice that the timing of the analyses, as well as the type of spending function, has a significant impact on the sample size

We also note that the O'Brien-Fleming procedure has dramatically lower impact on the sample size compared to the Pocock procedure. This is another major advantage of the O'Brien-Fleming methodology that was not mentioned earlier.

The timing of the analyses, has a different impact on the Pocock compared to the O'Brien-Fleming approach. When the analyses get bunched up at the end of the study, the sample size required for the Pocock procedure is lower while the one for the O'Brien-Fleming is slightly higher.

Unifying intuition

The reason is that the rate of spending of the alpha level is the undrelying factor which leads to the inflation of the sample size.

When we perform all the analyses at the end, the Pocock procedure spends its alpha level more slowly (since it was not able to spend any befor the 50% mark).

By contrast, the O'Brien-Fleming procedure spends more alpha in the unequal analysis timing case, since the early analyses (where O-F spends negligible alpha) occur later, (when the procedure spends substantial alpha level).

The unifying intuition is that, the faster the alpha level is spent, the larger the sample size must be to ensure that the power is mentained.

Using spending functions for flexible monitoring

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 $\alpha = 0.5$. The R code for this is

```
gsDesign(k=3, timing=c(0.2, 0.5, 1), sfu=sfHSD, sfupar=0, test.type=2)
```

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 $\alpha(\tau_3) = 0.0375$. Thus, the additional analysis will spend some of the remaining alpha of 1.25%. The revised 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

The R code to accomplish the insertion of the extra analysis at the 75% trial fraction is

```
gsDesign(k=4, timing=c(0.2, 0.5, .75, 1), sfu=sfHSD, sfupar=0, test.type=2)
```

Note the following:

- We simply redesign the study from the start, adding a fourth analysis at the 75% trial fraction
- The critical bounds before the third analysis did not change[2]
- 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) required to reject the null hypothesis in the final analysis.

[2] It would be a veritable disaster if they had to change!

Information-based monitoring

Statistical Information

The statistical information for parameter δ is the inverse of its variance i.e.,

$$I = \text{var}(\delta)^{-1}$$

Recall that, when the sample size for each treatment group is equal, 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 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

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

Information-based monitoring

However, if the sample size is not the same in the two arms, the fraction of the interim over the total sample size is not equal to the information fraction.

For example, if the total sample size is $2N = 200$ and the interim sample size is $n_1 = 50$ and $n_2 = 60$, the sample fraction is $t = 55\%$, but the information fraction is $t' = \left\{ \frac{\frac{2}{100}}{\frac{1}{50} + \frac{1}{60}} \right\} = 54.5\%$

Using information to monitor a trial obviates requiring equal sample sizes at each analysis. It also unifies the procedure of monitoring *regardless of the endpoint* (e.g., comparison of means, proportions or hazard ratios).

In fact, most software packages have an information monitoring computational core, which is then translated to the proper design by a user-defined wrapper.

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

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

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

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

Calculating the bound at the interim analysis

Suppose that we had designed the study so that an interim analysis were to be carried out at the 50% trial fraction, according to the O'Brien-Fleming spending function (say according to the HSD approach).

The original bounds would be

```
x<-gsDesign(k=2, test.type=2, sfu=sfHSD, sfupar=-4)
data.frame(`lower bounds`=x$lower$bound, `upper bounds`=x$upper$bound,
           `alpha spent`=cumsum(x$lower$spend))
```

```
##   lower.bounds upper.bounds alpha.spent
## 1    -2.749966     2.749966 0.002980073
## 2    -1.981131     1.981131 0.025000000
```

Revised bounds

What is the appropriate bound now?

All we need here is to know where we are. The information fraction provides a "Google map". Since $I_n = 0.42$ the proper bound for this analysis is

```
x<-gsDesign(k=2, test.type=2, sfu=sfHSD, sfupar=-4, timing=c(0.42,1))
data.frame(`lower bounds`=x$lower$bound, `upper bounds`=x$upper$bound,
           `alpha spent`=cumsum(x$lower$spend))
```

```
##   lower.bounds upper.bounds alpha.spent
## 1    -2.872492     2.872492 0.002036244
## 2    -1.976361     1.976361 0.025000000
```

So all we had to do was to revise the default `timing` of the analyses. It's that simple!

By the way, note here that, as we did not spend as much alpha (since by carrying out the analysis earlier, the O-F procedure spent less alpha), there is more alpha left for the final analysis, so, in the revised design, $Z(1)$ is smaller (the evidence required to reject the null hypothesis lower), compared to the original design.

Dealing with studies with time-to-event endpoints

Information in survival studies

Interim monitoring becomes complicated in survival analysis studies.

One complication is that the statistical information in studies with time-to-event endpoints is proportional to the number of events instead of the sample size.

This is for example the reason that prevention studies (e.g., studies of time to relapse in breast cancer) accrue tens of thousands of patients, as the number of endpoints are few and take a long time to be observed.

Information versus calendar time

Because information is proportional to the number of events and not the overall sample size, survival studies present logistical challenges as well.

Waiting until all events have been observed is no problem in a study with no monitoring, because one may continue until all events have been observed.

However, there's usually 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.

Monitoring a survival study through calendar time

You may decide that it is more convenient to monitor a study using the proportion of the maximum duration of the study rather than the information time.

This has the advantage of being able to schedule the review in a predictable manner, which in turn simplifies the logistics of a large number of people involved in the review synchronizing their calendars.

Alpha spending functions provide a way to do this pretty easily.

However, as the information time lags the calendar time (i.e., the number of observed events is low early and accelerates later), using calendar time to monitor the study will invariably result in overspending the alpha early.

Recalibrating monitoring of a survival 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 above).

At the second analysis however, you realize that the study will be unlikely to generate as many events.

Long story short, you figure that the first analysis actually occurred 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 $\alpha_P(0.20) = 0.015$.

However, 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.

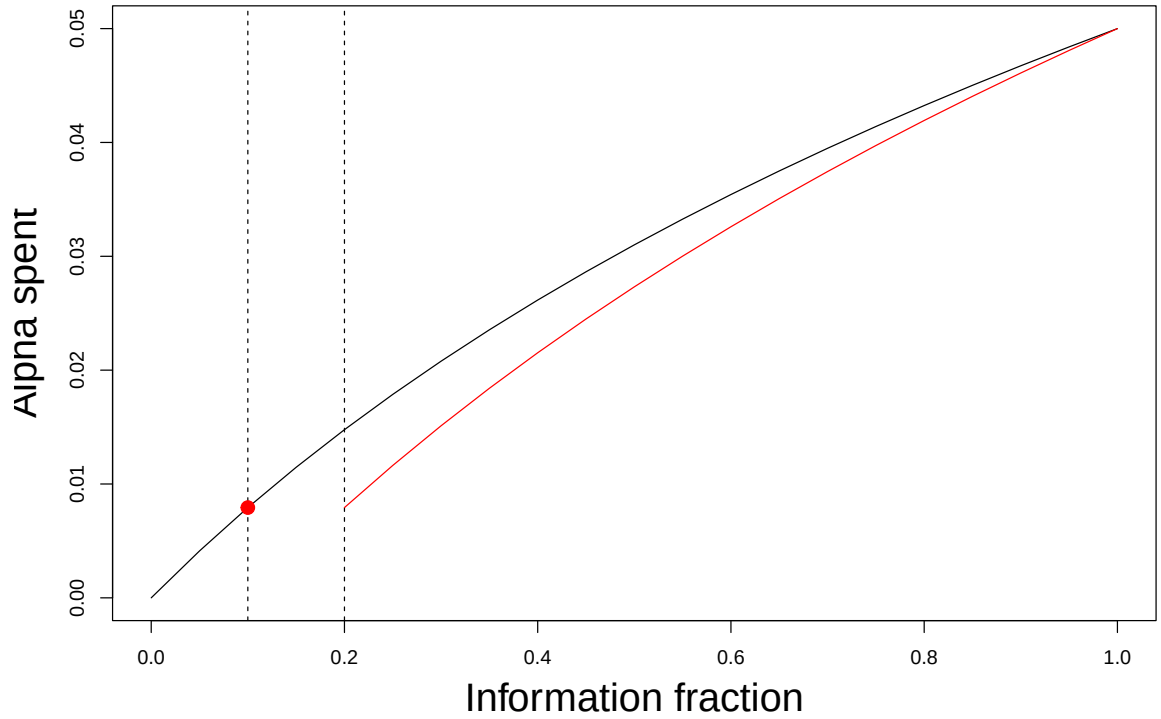
The alpha spent at the first look is gone; you cannot do anything about that. One way to adjust, is to interpolate the spending function for $t \geq 0.20$ by uniformly accelerating your alpha spending rate

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

Recalibration of the study: *Pictorial representation*

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

Figure 1: Pocock-like alpha spending functions recalibrated after the first interim analysis



Designing a study with a time-to-event endpoint

Designing survival studies

This heavily borrows from the vignette "Basic time-to-event group sequential design using `gsSurv`".

- The strategy entails the following two steps

Designing survival studies

This heavily borrows from the vignette "Basic time-to-event group sequential design using `gsSurv`".

- The strategy entails the following two steps
  - Design the survival study based on no interim analysis using the package `nSurv`

Designing survival studies

This heavily borrows from the vignette "Basic time-to-event group sequential design using `gsSurv`".

- The strategy entails the following two steps
  - Design the survival study based on no interim analysis using the package `nSurv`
  - Add the interim monitoring layer on top of the fixed design using the package `gsSurv`

Trial parameters

- Suppose we are given the following information about trial we are about to design:

Trial parameters

- Suppose we are given the following information about trial we are about to design:
  - The median time-to-event in the control group ($m = 12$ months)

Trial parameters

- Suppose we are given the following information about trial we are about to design:
  - The median time-to-event in the control group ($m = 12$ months)
  - The dropout rate ($\eta = 0.001$)

Trial parameters

- Suppose we are given the following information about trial we are about to design:
  - The median time-to-event in the control group ($m = 12$ months)
  - The dropout rate ($\eta = 0.001$)
  - The hazard ratios under the null ($HR_0 = 1$)

Trial parameters

- Suppose we are given the following information about trial we are about to design:
  - The median time-to-event in the control group ($m = 12$ months)
  - The dropout rate ($\eta = 0.001$)
  - The hazard ratios under the null ($HR_0 = 1$)
  - THe hazard ratio under the alternative ($HR = 0.75$)

Trial parameters

- Suppose we are given the following information about trial we are about to design:
  - The median time-to-event in the control group ($m = 12$ months)
  - The dropout rate ($\eta = 0.001$)
  - The hazard ratios under the null ($HR_0 = 1$)
  - THe hazard ratio under the alternative ($HR = 0.75$)
  - The desired Type-I error rate ($\alpha = 0.025$)

Trial parameters

- Suppose we are given the following information about trial we are about to design:
  - The median time-to-event in the control group ($m = 12$ months)
  - The dropout rate ($\eta = 0.001$)
  - The hazard ratios under the null ($HR_0 = 1$)
  - THe hazard ratio under the alternative ($HR = 0.75$)
  - The desired Type-I error rate ($\alpha = 0.025$)
  - THe desired Type-II error rate ($\beta = 0.1$, so power is 90%)

Enrollment and trial duration

To determine the enrollment and trial duration we follow here the method of Lachin & Foulkes (1986).

Their method fixes the accrual duration plus the total trial duration and calibrates the accrual *rate* to obtain desired power.

An alternative, which is not shown here, is the method by Kim & Tsiatis (1990), which fixes the accrual rate and follow-up (post-accrual) duration. Then the total trial duration is calibrated to generate the desired power.

- Here we set these parameters as follows:

Enrollment and trial duration

To determine the enrollment and trial duration we follow here the method of Lachin & Foulkes (1986).

Their method fixes the accrual duration plus the total trial duration and calibrates the accrual *rate* to obtain desired power.

An alternative, which is not shown here, is the method by Kim & Tsiatis (1990), which fixes the accrual rate and follow-up (post-accrual) duration. Then the total trial duration is calibrated to generate the desired power.

- Here we set these parameters as follows:
  - The study duration is 36 months

Enrollment and trial duration

To determine the enrollment and trial duration we follow here the method of Lachin & Foulkes (1986).

Their method fixes the accrual duration plus the total trial duration and calibrates the accrual *rate* to obtain desired power.

An alternative, which is not shown here, is the method by Kim & Tsiatis (1990), which fixes the accrual rate and follow-up (post-accrual) duration. Then the total trial duration is calibrated to generate the desired power.

- Here we set these parameters as follows:
  - The study duration is 36 months
  - The minimum follow-up after accrual is 12 months

Enrollment and trial duration

To determine the enrollment and trial duration we follow here the method of Lachin & Foulkes (1986).

Their method fixes the accrual duration plus the total trial duration and calibrates the accrual *rate* to obtain desired power.

An alternative, which is not shown here, is the method by Kim & Tsiatis (1990), which fixes the accrual rate and follow-up (post-accrual) duration. Then the total trial duration is calibrated to generate the desired power.

- Here we set these parameters as follows:
  - The study duration is 36 months
  - The minimum follow-up after accrual is 12 months
  - Instead of assuming uniform accrual, we split the accrual period into four subperiods $R = 1, 2, 3, 4$

Enrollment and trial duration

To determine the enrollment and trial duration we follow here the method of Lachin & Foulkes (1986).

Their method fixes the accrual duration plus the total trial duration and calibrates the accrual *rate* to obtain desired power.

An alternative, which is not shown here, is the method by Kim & Tsiatis (1990), which fixes the accrual rate and follow-up (post-accrual) duration. Then the total trial duration is calibrated to generate the desired power.

- Here we set these parameters as follows:
  - The study duration is 36 months
  - The minimum follow-up after accrual is 12 months
  - Instead of assuming uniform accrual, we split the accrual period into four subperiods $R = 1, 2, 3, 4$
  - We then assume a piece-wise uniform accrual rate in these four periods $\gamma = 1, 1.5, 2.5, 4$ patients per month respectively

Deriving design with no interim analyses

This information is sufficient to design a trial with no interim analyses using the package *nSurv*.

Note that the median time to event in the control group must be transformed to a hazard rate. This is done by assuming an exponential survival, which results in solving the integral

$$\int_0^m \lambda e^{-\lambda t} dt = 0.5$$

for λ. The solution is $\lambda = log(2)/m$, where m is the median time to event.

In the previous example, $m = 12$ so $\lambda = 0.0578$.

Note that, by similar arguments, the probability of annual dropout (assuming exponential dropout) is $\int_0^{12} \eta e^{-\eta t} dt = 0.99$, so we expect 1% dropout per year in the study.

Implementation of the no monitoring design

Given all of the above, we proceed as follows:

```
x <- nSurv(R = R, gamma = gamma, eta = eta, minfup = minfup, T = T,
           lambdaC = log(2) / median, hr = hr, hr0 = hr0,
           beta = beta, alpha = alpha)
```

THe design with no monitoring

Printing the design we get

```
x
```

```
## Fixed design, two-arm trial with time-to-event
## outcome (Lachin and Foulkes, 1986).
## Solving for:  Accrual rate
## Hazard ratio                  H1/H0=0.75/1
## Study duration:                   T=36
## Accrual duration:                   24
## Min. end-of-study follow-up: minfup=12
## Expected events (total, H1):        507.1519
## Expected sample size (total):       775.0306
## Accrual rates:
##      Stratum 1
## 0-1     9.2818
## 1-3    13.9227
## 3-6    23.2045
## 6-24   37.1272
## Control event rates (H1):
##       Stratum 1
## 0-Inf    0.0578
## Censoring rates:
##       Stratum 1
## 0-Inf     0.001
## Power:                 100*(1-beta)=90%
## Type I error (1-sided):   100*alpha=2.5%
## Equal randomization:          ratio=1
```

Comments

We note that the enrollment rates required to power the study within our required total study duration of $T = 36$ months are dramatically higher than the ones we entered (but remain in the proportion suggested by our input).

Had we not fixed the trial duration and considered only those rates, the trial duration would have been very long indeed.

Try it!

Adding the monitoring layer

Now we add the group sequential design.

We keep all parameters used previously.

- The following additional information is needed:

Adding the monitoring layer

Now we add the group sequential design.

We keep all parameters used previously.

- The following additional information is needed:
  - The total number of analyses $k = 3$

Adding the monitoring layer

Now we add the group sequential design.

We keep all parameters used previously.

- The following additional information is needed:
  - The total number of analyses $k = 3$
  - The timing of these analyses $(t = 0.25, 0.75, 1)$

Adding the monitoring layer

Now we add the group sequential design.

We keep all parameters used previously.

- The following additional information is needed:
  - The total number of analyses $k = 3$
  - The timing of these analyses ($t = 0.25, 0.75, 1$)
  - The spending function parameters. The vignette uses the Lan & DeMets approximation of the O'Brien-Fleming bounds `sfLDOF`

Generating the design

Now we are prepared to generate the design[3].

```
x <- gsSurv(
  k = k, test.type=2, timing = timing, R = R, gamma = gamma, eta = eta,
  minfup = minfup, T = T, lambdaC = log(2) / median,
  hr = hr, hr0 = hr0, beta = beta, alpha = alpha,
  sfu = sfu, sfupar = sfupar
)
```

[3]The vignette also adds a beta spending function. We will skip this for now and address this nuance later in this course

Design summary

```
x
```

```
## Time to event group sequential design with HR= 0.75
## Equal randomization:          ratio=1
## Symmetric two-sided group sequential design with
## 90 % power and 2.5 % Type I Error.
## Spending computations assume trial stops
## if a bound is crossed.
##
##
##   Analysis  N   Z   Nominal p  Spend
##          1 129 4.33    0.0000 0.0000
##          2 387 2.34    0.0096 0.0096
##          3 516 2.01    0.0221 0.0154
##      Total                    0.0250
##
## ++ alpha spending:
##  Lan-DeMets O'Brien-Fleming approximation spending function with none = 1.
##
## Boundary crossing probabilities and expected sample size
## assume any cross stops the trial
##
## Upper boundary (power or Type I Error)
##           Analysis
##    Theta      1      2      3 Total  E{N}
##   0.0000 0.0000 0.0096 0.0154 0.025 513.3
##   0.1439 0.0035 0.6849 0.2116 0.900 426.1
##
## Lower boundary (futility or Type II Error)
##           Analysis
##    Theta 1      2      3 Total
##   0.0000 0 0.0096 0.0154 0.025
##   0.1439 0 0.0000 0.0000 0.000
##              T        n   Events HR futility HR efficacy
## IA 1  15.89222 482.0674 128.9423       2.145       0.466
## IA 2  27.97574 788.2026 386.8284       1.269       0.788
## Final 36.00000 788.2026 515.7712       1.194       0.838
## Accrual rates:
##      Stratum 1
```

Targeted summaries

As the output is chock full of information and can be overwhelming to the reader, it is instructive to break it apart.

Here are some proposals on how to do that.

```
kableExtra::kable(data.frame(Period = paste("Month", rownames(x$gamma)),
       Rate = as.numeric(x$gamma)))
```

Period	Rate
Month 0-1	9.439552
Month 1-3	14.159329
Month 3-6	23.598881
Month 6-24	37.758210

Design summaries

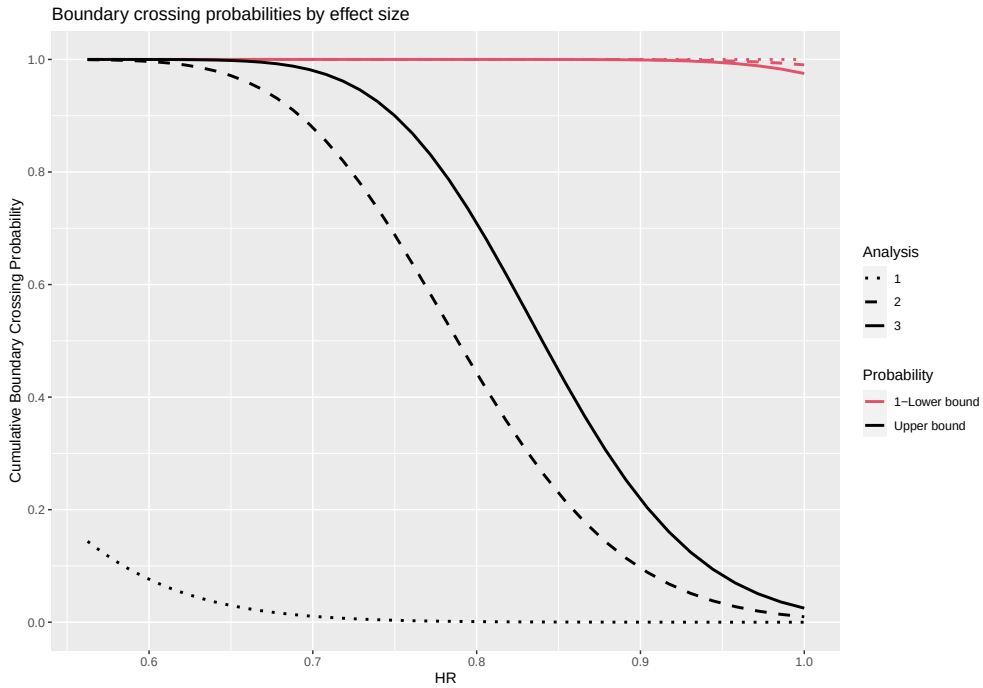
```
gsBoundSummary(x) %>% kableExtra::kable()
```

	Analysis	Value	Efficacy	Futility
1	IA 1: 25%	Z	4.3326	-4.3326
4	N: 484	p (1-sided)	0.0000	0.0000
7	Events: 129	~HR at bound	0.4662	2.1449
20	Month: 16	P(Cross) if HR=1	0.0000	0.0000
23		P(Cross) if HR=0.75	0.0035	0.0000
2	IA 2: 75%	Z	2.3398	-2.3398
5	N: 790	p (1-sided)	0.0096	0.0096
8	Events: 387	~HR at bound	0.7883	1.2686
21	Month: 28	P(Cross) if HR=1	0.0096	0.0096
24		P(Cross) if HR=0.75	0.6884	0.0000
3	Final	Z	2.0118	-2.0118
6	N: 790	p (1-sided)	0.0221	0.0221
9	Events: 516	~HR at bound	0.8376	1.1938
22	Month: 36	P(Cross) if HR=1	0.0250	0.0250
25		P(Cross) if HR=0.75	0.9000	0.0000

Informative plots

We have seen the basic plot with the bounds. Here is another plot, which shows how easy (or not so easy) it is to reject the null hypothesis based on values of the hazard ratio.

Figure 2: Boundary crossing probabilities for different observed hazard rates



These curves show how difficult it is to reject the null hypothesis early (see dotted black and red lines) either with in favor of the experimental or standard therapy.

Beta spending and 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

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

For all subsequent analyses analyses 2 through k these are

$$\beta(\tau_i) = \mathrm{P}\left(b_j \leq Z(\tau_j) < a_j, j = 1, \cdots, i-1, i = 2, \cdots, k, Z(\tau_j) < b_i | H_A\right)$$

with $\sum_{i=1}^{k} \beta(\tau_i) = \beta$. The beta spending probabilities are calculated as for the alpha spending case.

The trial continues while $b_i \leq Z(\tau_i) < a_i,\ i = 1, \cdots, k-1$.

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 the HSD family spending functions with $\gamma = -4$[4] and $\gamma = 1$[5] we obtain the following upper and lower boundaries:

```
x=gsDesign(k=5, test.type=4, sfu=sfHSD, sfupar=-4, sfl=sfHSD, sflpar=1)
data.frame(`futility bound`=x$lower$bound, `superiority bound`=x$upper$boun
  kableExtra::kable()
```

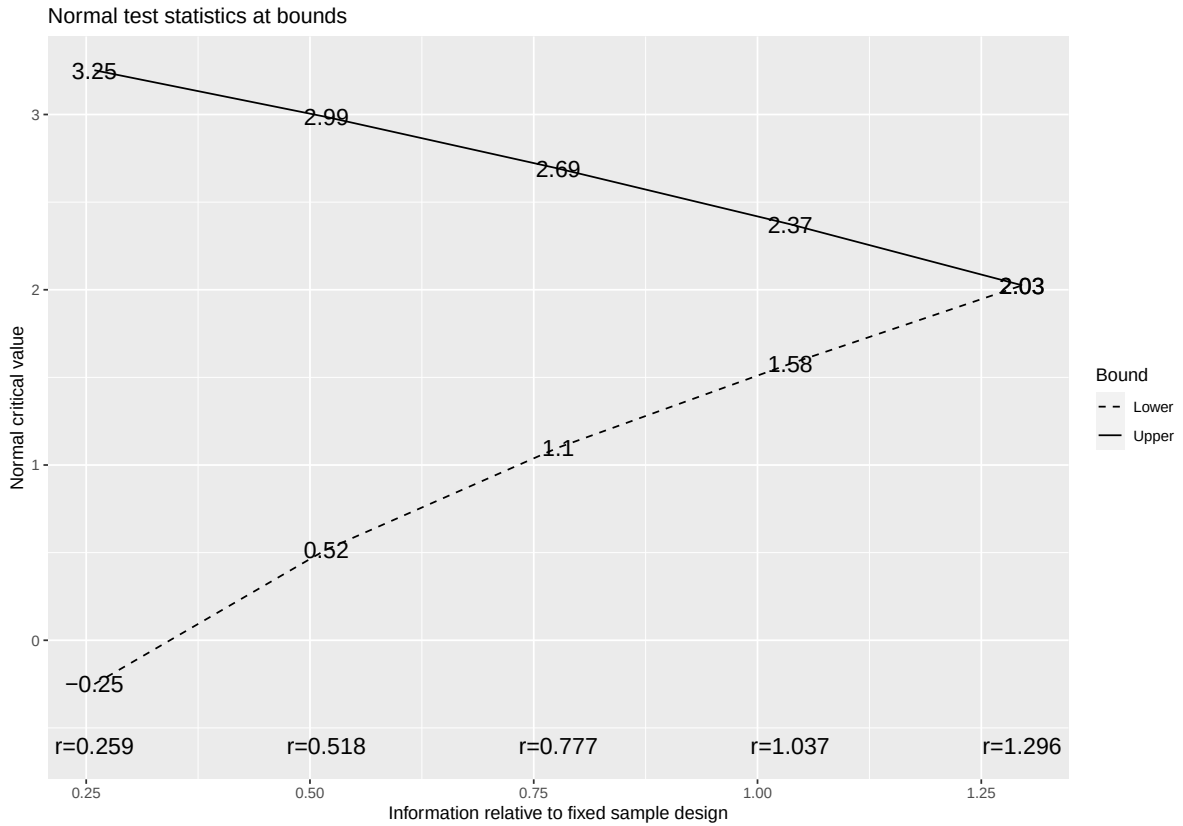
futility.bound	superiority.bound
-0.2504978	3.252668
0.5178456	2.986046
1.0995819	2.691657
1.5776135	2.373666
2.0253213	2.025321

[4] This is approximately the O'Brien-Fleming procedure.
[5] Closer to the Pocock procedure

Pictorial representation

A pictorial representation is as follows:

Figure 3: Pictorial representation of an efficacy (upper) and a futility (lower) bound



Summary of the design

The summary of the design is as follows:

```
gsBoundSummary(x)%>% kableExtra::kable()
```

	Analysis	Value	Efficacy	Futility
1	IA 1: 20%	Z	3.2527	-0.2505
6	N/Fixed design N: 0.26	p (1-sided)	0.0006	0.5989
11		~delta at bound	1.9712	-0.1518
34		P(Cross) if delta=0	0.0006	0.4011
39		P(Cross) if delta=1	0.0545	0.0287
2	IA 2: 40%	Z	2.9860	0.5178
7	N/Fixed design N: 0.52	p (1-sided)	0.0014	0.3023
12		~delta at bound	1.2796	0.2219
35		P(Cross) if delta=0	0.0018	0.7240
40		P(Cross) if delta=1	0.2652	0.0522
3	IA 3: 60%	Z	2.6917	1.0996
8	N/Fixed design N: 0.78	p (1-sided)	0.0036	0.1358
13		~delta at bound	0.9418	0.3847
36		P(Cross) if delta=0	0.0046	0.8867
41		P(Cross) if delta=1	0.5756	0.0714
4	IA 4: 80%	Z	2.3737	1.5776
9	N/Fixed design N: 1.04	p (1-sided)	0.0088	0.0573
14		~delta at bound	0.7192	0.4780
37		P(Cross) if delta=0	0.0104	0.9556
42		P(Cross) if delta=1	0.8128	0.0871
5	Final	Z	2.0253	2.0253
10	N/Fixed design N: 1.3	p (1-sided)	0.0214	0.0214
15		~delta at bound	0.5489	0.5489
38		P(Cross) if delta=0	0.0189	0.9811
43		P(Cross) if delta=1	0.9000	0.1000

The inflation of the sample size is about 30%.

Comments

You must distinguish between futility or beta-spending bounds and lower superiority bounds.

The latter (lower efficacy bounds) when crossed, signify the superiority of the standard therapy. By contrast, the futility bounds signify the trials inability to reject the null in either direction.

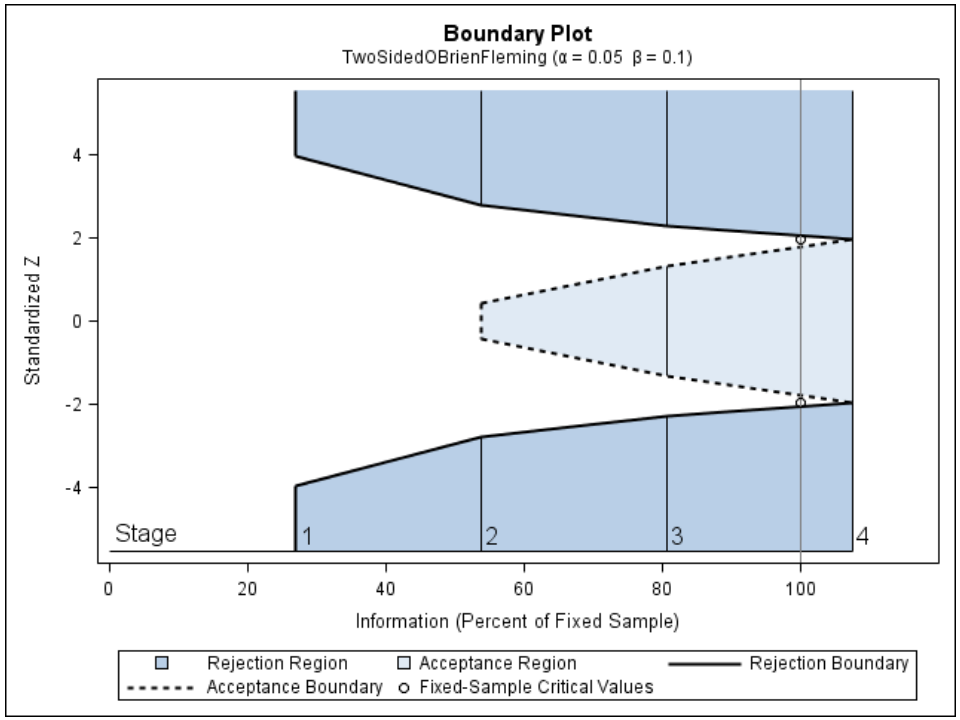
The decision to stop a study for futility is not symmetric as with rejecting the null hypothesis for superiority.

As with the situation in the example, we may choose to spend beta more quickly to explore the possibility that the study is unlikely to result in the rejection of the null hypothesis.

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 4: Two-sided boundaries with futility bounds (inner wedge) for an O-F study with 4 analyses



The study continues while $Z(t)$ is in the white region. It stops if the upper boundary (superiority of experimental treatment) or the lower boundary is crossed (superiority of the standard treatment) or if $Z(t)$ ventures into the light blue region (wedge, futility).