Analysis of a group-sequential trial with a survival endpoint

Marcel Wolbers, Gernot Wassmer, and Friedrich Pahlke

Last change: 20 Juni, 2019

This tutorial provides two examples:

- The **first example** illustrates **how to get inference (point estimate, confidence interval, and p-value) which respect the group-sequential design after rejecting the null hypothesis at the interim or final analysis**.
- The **second example** illustrates tools which are relevant for **monitoring interim results of an ongoing trial**: **repeated confidence intervals** and **conditional power**.

For a general introduction to "Inference in group-sequential designs", please refer to the book ["Group Sequential and Confirmatory Adaptive Designs in Clinical Trials"]{http://monograph.wassmer.brannath.rpact.net/} by Gernot Wassmer & Werner Brannath.

This tutorial only covers survival endpoints. Code for other endpoints is similar but the dataset needs to be provided in a different format (see `?getDataset` for details).

Inference for the Gallium trial which was stopped at an interim analysis

For details about the Gallium trial, we refer to the primary study publication: Marcus et al, N Engl J Med 2017; 377:1331-1344.

Gallium design background

Trial characteristics:

- Population: Treatment-naive follicular lymphoma patients.
- Comparison: Rituximab + chemotherapy vs. Obinutuzumab + chemotherapy. Rituximab: Rituxan, Mabthera. Obinutuzumab: Gazyva(ro).
- Phase III, 1:1 randomized, open-label clinical trial.
- Primary endpoint: investigator-assessed progression-free survival (PFS).

Group-sequential design:

- O'Brien-Fleming boundary with interim analyses after 30% and 67% of PFS events.
- Non-binding futility after 30% of PFS events (if estimated HR > 1).
- Target HR 0.74, 80% power at two-sided 5% significance level $\Rightarrow$ final analysis at 370 PFS events.
- Target sample size is 1202 subjects.

Conventional analyses at the first and second interim analysis

Results from standard inference at the **futility interim analysis after 113 events**:

- Stratified HR 0.69 (95% CI 0.47 to 1.01). $\Rightarrow$ **Trial continues.**
- log(HR) = log(0.69) with standard error 0.20.
- Corresponding Z-score: log(0.69)/0.20 = -1.86.

Results from standard inference at the **efficacy interim analysis after 245 events**:

- Stratified HR 0.66 (95% CI 0.51 to 0.85).

- log(HR) = log(0.66) with standard error 0.13.
- Corresponding Z-score: log(0.66)/0.13 = -3.225.
- Two-sided p-value is 0.0012 which was smaller than the critical value from the O'Brien-Fleming boundary of 0.012. ⇒ **Trial stopped early for efficacy**.

Analysis accounting for the group-sequential design

First, load the rpact package and **define the group-sequential boundaries** using the function `getDesignGroupSequential`. Note that while the Gallium protocol specified a two-sided significance level of 5%, we implement this via a one-sided significance level of 2.5% as rpact (sensibly) only supports one-sided designs if futility interim analyses are specified.

```
# Load rpact
library(rpact)
packageVersion("rpact") # version should be version 2.0.1 or later
```

```
## [1] '2.0.1'
```

```
### Define design
# FutilityBounds = c(0,-6) are on the Z-scale; a value of Z = 0 implies futility
# if the interim estimate is "in the wrong direction" (i.e., HR >= 1 here),
# a value of Z = -6 is essentially the same as Z = -Inf and implies no futility
# boundary for the second interim as per the Gallium design
design <- getDesignGroupSequential(informationRates = c(113,245,370)/370,
    typeOfDesign = "asOF", sided = 1, alpha = 0.025,
    futilityBounds = c(0,-6), bindingFutility = FALSE)
```

Note that `bindingFutility = FALSE` has no impact because it is the default, so actually this could be omitted (same holds for `sided = 1` and `alpha = 0.025`).

Second, the **results after the first and second interim** are specified using the function `getDataset`:

```
# overallLogRanks: One-sided logrank statistic or Z-score (=log(HR)/SE) from Cox regression
results <- getDataset(
    overallEvents = c(113,245),
    overallLogRanks = c(-1.86,-3.225),
    overallAllocationRatio = c(1, 1))
```

Finally, this is used for creating the **adjusted inference** using the function `getAnalysisResults` (`directionUpper = FALSE` is specified because the power is directed towards negative values of the logrank statistics):

```
adj_result <- getAnalysisResults(design = design,
    dataInput = results,stage=2,directionUpper = FALSE)
```

```
## [PROGRESS] Stage results calculated [0.009 secs]
## [PROGRESS] Conditional power calculated [0.001 secs]
## [PROGRESS] Conditional rejection probabilities (CRP) calculated [0 secs]
## [PROGRESS] Repeated confidence interval of stage 1 calculated [0.2284 secs]
## [PROGRESS] Repeated confidence interval of stage 2 calculated [0.2394 secs]
## [PROGRESS] Repeated confidence interval calculated [0.4677 secs]
## [PROGRESS] Overall repeated p-values of stage 1 calculated [0.4289 secs]
## [PROGRESS] Overall repeated p-values of stage 2 calculated [0.4468 secs]
## [PROGRESS] Repeated p-values calculated [0.8767 secs]
## [PROGRESS] Final p-value calculated [0.001 secs]
## [PROGRESS] Final confidence interval calculated [0.0379 secs]
```

```
adj_result
```

```
## Analysis results (group sequential design):
##   Stages                                  : 1, 2, 3
##   Information rates                       : 0.305, 0.662, 1.000
##   Critical values                         : 3.891, 2.520, 1.992
##   Futility bounds (non-binding)           : 0.000, -Inf
##   Cumulative alpha spending               : 4.995e-05, 5.879e-03, 2.500e-02
##   Stage levels                            : 4.995e-05, 5.861e-03, 2.318e-02
##   Effect sizes                            : 0.7047, 0.6623, NA
##   Test statistics                         : -1.860, -2.673, NA
##   p-values                                : 0.031443, 0.003762, NA
##   Overall test statistics                 : -1.860, -3.225, NA
##   Overall p-values                        : 0.0314428, 0.0006299, NA
##   Actions                                 : continue, reject and stop, NA
##   Theta H0                                : 1
##   CRP                                     : 0.1373, 0.8616, NA
##   Planned sample size                     : NA, NA, NA
##   Planned allocation ratio                : 1
##   Assumed effect                          : NA
##   Conditional power                       : NA, NA, NA
##   RCIs (lower)                            : 0.339, 0.480, NA
##   RCIs (upper)                            : 1.465, 0.914, NA
##   Repeated p-values                       : 0.234459, 0.005409, NA
##   Final stage                             : 2
##   Final p-value                           : NA, 0.0006656, NA
##   Final CIs (lower)                       : NA, 0.516, NA
##   Final CIs (upper)                       : NA, 0.852, NA
##   Median unbiased estimate                : NA, 0.663, NA
```

The output is explained as follows:

- `Critical values` are group-sequential efficacy boundary values on the Z-scale, `stage levels` are the corresponding one-sided local significance levels.
- `Effect sizes`, `Test statistics`, and `p-values` refer to hazard ratio estimates, Z-scores, and p-values obtained from the first interim analysis and results which would have been obtained after the second interim analysis if not all data up to the second interim analysis but only new data since the first interim had been included (i.e., per-stage results).
- `Overall test statistics` are the given (overall, not per-stage) Z-scores from each interim and `Overall p-value` the corresponding one-sided p-values.
- `RCIs` are repeated confidence intervals which provide valid (but conservative) inference at any stage of an ongoing or stopped group-sequential trial. `Repeated p-values` are the corresponding p-values.
- `Final p-value` is the **final one-sided adjusted p-value** based on the stagewise ordering of the sample space.
- `Median unbiased estimate` and `Final CIs` are the corresponding **adjusted treatment effect estimate** and the **confidence interval** for the hazard ratio at the interim analysis where the trial was stopped.

Note that for this example, the **adjusted final hazard ratio of 0.66 and the adjusted confidence interval of (0.52, 0.85) match the results from the conventional analysis almost exactly for the first two decimals.** This is consistent with the finding that stopping a trial after 50% or more of the events had been collected has a negligible impact on estimation.

Tools for monitoring an ongoing trial

Monitoring ongoing trials is also possible with the function `getAnalysisResults` introduced above. **Repeated confidence intervals** which provide valid (but conservative) inference at any stage of an ongoing or stopped group-sequential trial can be obtained using the same code as introduced in the previous example. **Conditional power calculations** require additional specification of the following arguments:

- The assumed true hazard ratio `thetaH1`.
- The planned number of additional events for future interim stages (`nPlanned`).
- The planned allocation ratio `allocationRatioPlanned` for future interim stages (default is 1).

A hypothetical example

Assume the same design as for the Gallium trial introduced above and the following **hypothetical interim results**:

Hypothetical results from standard inference at the **futility interim analysis** after 113 events:

- Stratified HR 0.69 (95% CI 0.47 to 1.01). ⇒ **Trial would continue.**
- log(HR) = log(0.69) with standard error 0.20.
- Corresponding Z-score: log(0.69)/0.20 = -1.86.

Hypothetical results from standard inference at the **efficacy interim analysis** after 245 events:

- Stratified HR 0.80 (95% CI 0.62 to 1.03). ⇒ **Trial would continue.**
- log(HR) = log(0.80) with standard error 0.13.
- Corresponding Z-score: log(0.80)/0.13 = -1.716.

Calculation of repeated confidence intervals and **conditional power**:

```
# 1) Specify results so far using function getDataset as before
results <- getDataset(
    overallEvents = c(113,245),
    overallLogRanks = c(-1.86,-1.716),
    overallAllocationRatio = c(1, 1))

# 2) Calculate repeated confidence intervals and conditional power using
# the function getAnalysisResults as before
# Additional arguments for the conditional power calculation are
# - nPlanned: additional events from second interim until final analysis
#   (370-245 for this trial)
# - thetaH1: True hazard ratio governing future stages
#   (set to 0.74 here as per the original protocol assumptions)
interim_results <- getAnalysisResults(design = design,
    dataInput = results,directionUpper = FALSE,
    nPlanned=370-245,thetaH1=0.74)
```

```
## [PROGRESS] Stage results calculated [0.009 secs]
## [PROGRESS] Conditional power calculated [0 secs]
## [PROGRESS] Conditional rejection probabilities (CRP) calculated [0.001 secs]
## [PROGRESS] Repeated confidence interval of stage 1 calculated [0.2364 secs]
## [PROGRESS] Repeated confidence interval of stage 2 calculated [0.4787 secs]
## [PROGRESS] Repeated confidence interval calculated [0.7151 secs]
## [PROGRESS] Overall repeated p-values of stage 1 calculated [0.392 secs]
## [PROGRESS] Overall repeated p-values of stage 2 calculated [0.3989 secs]
## [PROGRESS] Repeated p-values calculated [0.7919 secs]
## [PROGRESS] Final p-value calculated [0.001 secs]
## [PROGRESS] Final confidence interval calculated [0.009 secs]
```

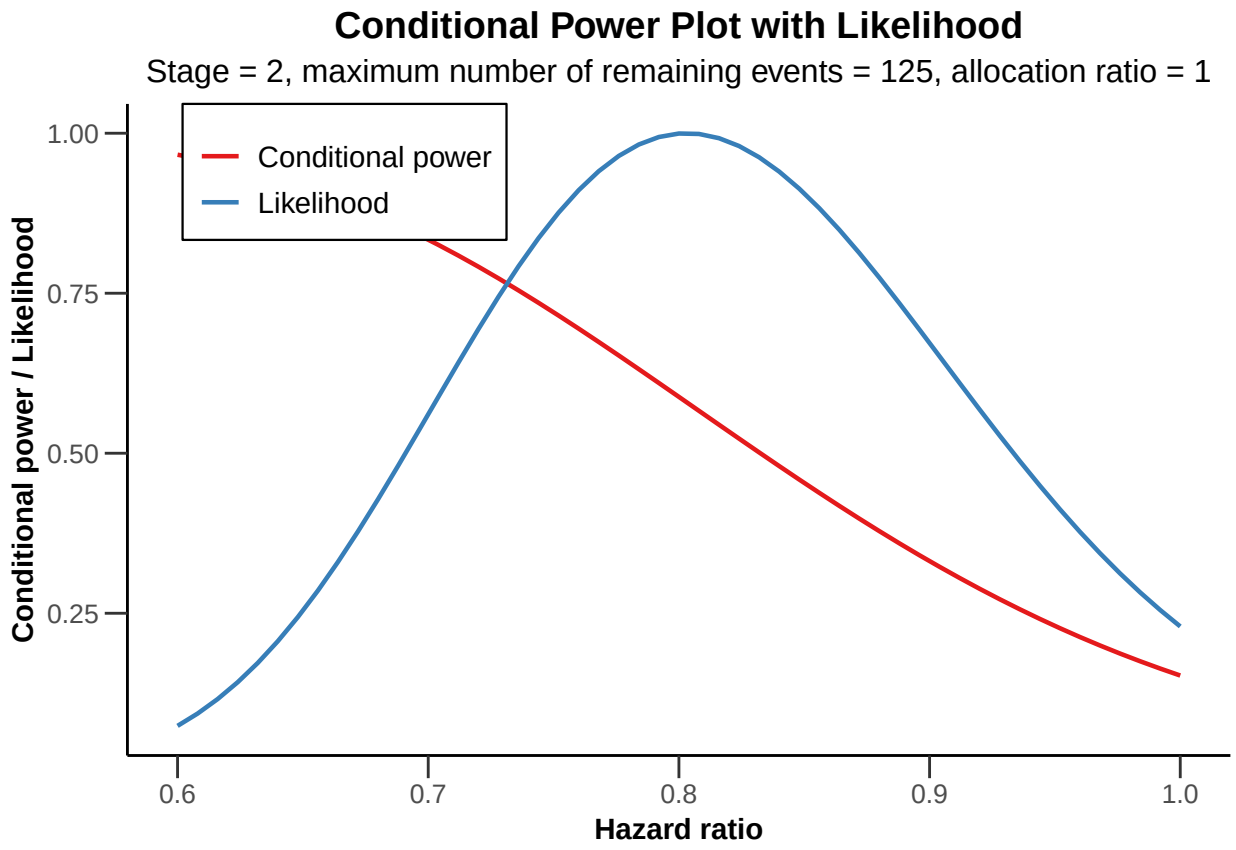
```
interim_results
```

```
## Analysis results (group sequential design):
##   Stages                                  : 1, 2, 3
##   Information rates                       : 0.305, 0.662, 1.000
##   Critical values                         : 3.891, 2.520, 1.992
##   Futility bounds (non-binding)           : 0.000, -Inf
##   Cumulative alpha spending               : 4.995e-05, 5.879e-03, 2.500e-02
##   Stage levels                            : 4.995e-05, 5.861e-03, 2.318e-02
##   Effect sizes                            : 0.7047, 0.8031, NA
##   Test statistics                         : -1.860, -0.617, NA
##   p-values                                : 0.03144, 0.26865, NA
##   Overall test statistics                 : -1.860, -1.716, NA
##   Overall p-values                        : 0.03144, 0.04308, NA
##   Actions                                 : continue, continue, NA
##   Theta H0                                : 1
##   CRP                                     : 0.1373, 0.1527, NA
##   Planned sample size                     : NA, NA, 125
##   Planned allocation ratio                : 1
##   Assumed effect                          : 0.74
##   Conditional power                       : NA, NA, 0.7448
##   RCIs (lower)                            : 0.339, 0.582, NA
##   RCIs (upper)                            : 1.47, 1.11, NA
##   Repeated p-values                       : 0.2345, 0.1013, NA
##   Final stage                             : NA
##   Final p-value                           : NA, NA, NA
##   Final CIs (lower)                       : NA, NA, NA
##   Final CIs (upper)                       : NA, NA, NA
##   Median unbiased estimate                : NA, NA, NA
```

As per the output above, the recommended action after the second interim analysis of this hypothetical trial would be to continue the trial, a repeated confidence interval for the hazard ratio is (0.58 to 1.11), and the conditional power to reach significance at the final analysis under protocol assumptions is 0.745. Final estimates and p-values are still missing as the trial has not stopped yet.

To obtain a **plot of the conditional power over a range of alternatives** you might call the rpact `plot` function and specify the range for theta with `thetaRange`. This produces the conditional power curve together with the likelihood function over the specified range:

```
plot(interim_results, thetaRange = c(0.6, 1))
```

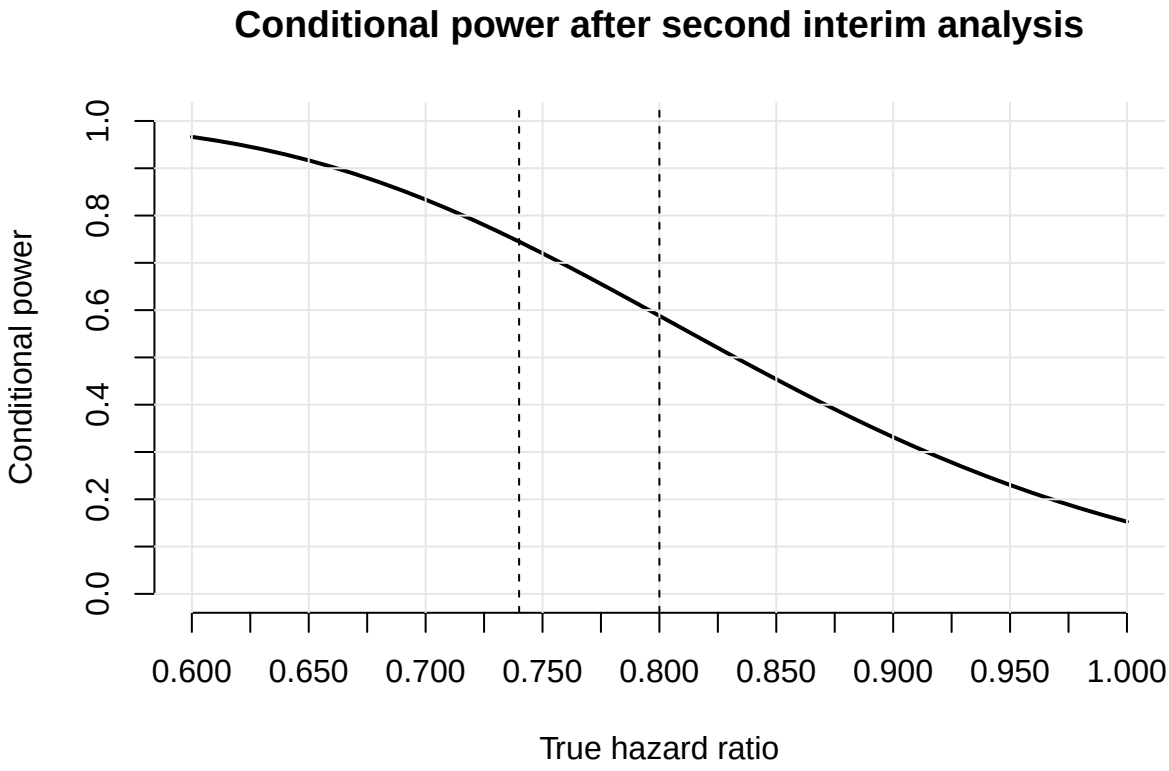


Note that `nPlanned`is from `interim_results` and can optionally be changed.

If one **only** wants to **calculate the conditional power**, then it is computationally more efficient to call the functions `getStageResults` and `getConditionalPower` instead. The code below illustrates this by plotting the conditional power curve depending on the true treatment effect. The dashed vertical lines in the plot correspond to the protocol hazard ratio of 0.74 and the observed interim hazard ratio of 0.8.

```
# get stage results so far
stageResults <- getStageResults(design, results,directionUpper = FALSE)

# calculate condition power for true HR ranging from 0.6 to 1
hr <- seq(0.6,1,by=0.01)
cpower <- rep(NA,length(hr))

for (i in 1:length(hr)){
  cpower[i] <- getConditionalPower(design,stageResults,nPlanned=370-245,
    thetaH1=hr[i])$conditionalPower[3]
}

# Plot results

plot(hr,cpower,
    type = "l",xlab = "True hazard ratio",
    ylab = "Conditional power",
    lwd = 2,ylim = c(0,1),axes = FALSE,
    main = "Conditional power after second interim analysis")
axis(1, at = seq(0.6,1,by =0.025)); axis(2,at = seq(0,1,by = 0.1))
```

```
abline(v = seq(0.6,1,by =0.05),h = seq(0,1,by = 0.1),col = gray(0.9))
abline(v=c(0.74,0.8),lty=2)
```



System: rpact 2.0.1, R version 3.5.2 (2018-12-20), platform: x86_64-w64-mingw32

To cite package 'rpact' in publications use:

Gernot Wassmer and Friedrich Pahlke (2019). rpact: Confirmatory Adaptive Clinical Trial Design and Analysis. R package version 2.0.1. https://CRAN.R-project.org/package=rpact

License