Interim Analyses in Clinical Trials
Learning During the Course of a Trial, Design Adaptations

Prof. Dr. Hans-Helge Müller
Hans-Helge.Mueller@ibe.med.uni-muenchen.de
IBE – Institut für Medizinische Informationsverarbeitung,
Biometrie und Epidemiologie
Introduction:

General Considerations

Approaches
Learning During the Course of a Trial: Why?

Uncertainty at the planning stage regarding

- effect size and variance
  - sample size
  - interim analyses
- target population
- dosing
- etc.

despite of careful planning

Long-term trials are rarely running as planned!

Corrections possible without violation of integrity?

Ethics and costs: correction versus pilot studies
Approaches For Design Adaptations With Control of the Type I Error Rate

- **Internal Pilot Design**
- **Adaptive Designs**
- **CRP-Approach**
Internal pilot design of Kieser & Friede:
• pre-specification of the interim time point
• only the estimator of the nuisance parameter is used for recalculation
• pre-specification of a recalculation rule

Methods for flexible design adaptations:
• all data (grouped, internal and external) can be used for recalculation
• pre-specification of a rec. rule not necessary
Adaptive designs
(Bauer & Köhne, Lehmacher & Wassmer, P & H, …)
• pre-specification of interim time point(s)
• pre-specification of a combination rule:
  combination test or conditional error function (CEF) or …

CRP-approach (Müller & Schäfer)
• pre-specification of a conventional (group sequential) design
• design adaptations at any time point possible
  (e.g. immediately following an emergency)
• more efficient if no adaptations are necessary
Most Popular Adaptation:
Sample Size Recalculation
Statistical Model and Setting:
One-sided t-test for comparison of two groups in the case of normal distributions with unknown common variance $\sigma^2$.
$\Delta$ or $\delta$ is the parameter for the difference in the location parameter.
Uncertainty only (!) in the specification of the variance at sample size calculation. Two-stage procedure with the interim analysis for estimation of the variance as the result of a pilot study. Accordingly, recalculation of the sample size with a pre-specified sample size formula (usually the standard formula, maybe with truncation).
Estimation of $\sigma^2$ blinded for comparison groups (?!), thus by the pooled one-sample variance estimator: $S_{os}^2$ Depending on the true difference $\delta$, this estimator is biased: 

$$E(S_{os}^2) - \sigma^2 = \frac{n_1}{2 \cdot (2 \cdot n_1 - 1)} \cdot \delta^2$$

Adjusted estimator:

$$S_{adj}^2 = S_{os}^2 - \frac{n_1}{2(2n_1 - 1)} \delta^2$$

$$S_{adj}^2 = \frac{2(n_1 - 1)}{2n_1 - 1} S_{1,\text{pool}}^2 + \frac{n_1}{2(2n_1 - 1)} (\bar{X}_{11} - \bar{X}_{12})^2 - \frac{n_1}{2(2n_1 - 1)} \delta^2$$

Sample size recalculation rule:

$$\hat{N}_{\text{recalc}}(S_{adj}^2) = 2 \left( \frac{z_{1-\alpha} + z_{1-\beta}}{\delta^2} \right) S_{adj}^2$$

$$= \frac{N}{2n_1 - 1} \left( \frac{2(n_1 - 1)S_{1,\text{pool}}^2}{\sigma^2} + \frac{n_1}{2} \left( \frac{\bar{X}_{11} - \bar{X}_{12}}{\sigma} \right)^2 - \frac{n_1}{2} \left( \frac{\delta}{\sigma} \right)^2 \right)$$
Representation of the t-test statistic and the recalculated sample size by (conditionally) independent variables with distributions having explicit density functions:

\[ T = \sqrt{\left( \frac{n_1 + n_2}{2} \right)} \frac{\bar{X}_1 - \bar{X}_2}{S_{\text{pool}}} \]

\[ T = \frac{\sqrt{\{(n_1)/(n_1 + n_2)\}D_1 + \sqrt{\{(n_2)/(n_1 + n_2)\}D_2}}}{\sqrt{\{(V_1 + V_2^*)/(2(n_1 + n_2 - 1))\}}} \]

\[ \hat{N}_{\text{reproc}}(S_{\text{ad}}^2) = \frac{N}{2n_1 - 1} (V_1 + D_1^2) - \frac{n_1}{2n_1 - 1} (z_{1-\alpha} + z_{1-\beta})^2 \]

\[ D_i = \sqrt{\left( \frac{n_i}{2} \right)} \frac{\bar{X}_{i1} - \bar{X}_{i2}}{\sigma^2}, \quad V_i = \frac{2(n_i - 1)S_{i,\text{pool}}^2}{\sigma^2}, \quad i = 1, 2 \]

\[ V_2^* = V_2 + n_1 n_2/((n_1 + n_2)\sigma^2)((\bar{X}_{11} - \bar{X}_{21})^2 + (\bar{X}_{12} - \bar{X}_{22})^2) \]
Joint density function:

\[
f(d_1, v_1, d_2, v^*_2 \mid H_1) = g_{N(\theta_1, 1)}(d_1)g_{\chi^2_{2(n_1-1)}}(v_1)g_{N(\theta_2, 1)}(d_2)g_{\chi^2_{2n_2}}(v^*_2)
\]

\[
\theta_i = \sqrt{(n_i/2)} \frac{\Delta}{\sigma}, \quad i = 1, 2
\]

Rejection Region:

\[
C = \{(d_1, d_2, v_1, v^*_2): T(d_1, d_2, v_1, v^*_2) > t_{1-\alpha, 2} \cdot (n_1+n_2-1)\}
\]

Type I error rate accounting for the sample size recalculation:

\[
\alpha_{\text{act}} = \int_{\mathcal{C}} f(d_1, d_2, v_1, v^*_2 \mid H_0) \, dd_1 dd_2 dv_1 dv^*_2
\]
Some Major References:


Adaptive Designs
Some Milestones


Lehmacher W, Wassmer G (1999): Construction of an adaptive design with the inverse normal method
P-Value Combination Tests (Bauer & Köhne, 1994)

Designs were proposed which allow for an adaptive interim analysis. Flexibility of design adaptations include: sample size adjustment, change of statistical test, change of endpoint, change of comparison, and more.

The time of an adaptive interim analysis will be planned. Interim analyses divide the trial into independent stages/sub-samples.

One-sided p-values of the stages will be calculated, and the p-values will be combined using a combination rule/test which maintains the overall type I error level.
Observations: independent one-sided p-values $p_1$ and $p_2$ of the two stages

Critical values: upper boundaries $\alpha_1$ and $\alpha_2$

Sequential rule: Reject $H_0$ if $p_1 \leq \alpha_1$.
If $\alpha_1 < p_1$, plan the second stage using Fisher’s product criterion $\alpha_2 = c/p_1$ where $\alpha_1 = c = c_\alpha = \exp(-\frac{1}{2} \chi^2_{4,1-\alpha})$. 

Bauer&Köhne Design, One-Sided, with Nonstochastic Curtailing

Illustration of the rejection region with one-sided p-values
one-sided level alpha = 0.025
Observations:
as before

Critical values:
as before

Futility stopping:
additional lower boundary $\alpha_0$

Modified sequential rule:
Reject $H_0$ if $p_1 \leq \alpha_1$, accept $H_0$ if $p_1 > \alpha_0$.
If $\alpha_1 < p_1 \leq \alpha_0$, plan the second stage using $\alpha_2 = c/p_1$
where $\alpha = \alpha_1 + c \left( \log \alpha_0 - \log \alpha_1 \right)$ and $c$ as before.
Illustration of the rejection region with one-sided p-values
one-sided level alpha = 0.025
Remark on determination of $c$ for the product criterion:

$p$-values $p_1, p_2$ i.i.d. $U([0,1])$;
$-2 \log p_1, -2 \log p_2$ i.i.d. $\exp(\frac{1}{2})$; $\exp(\lambda) = \gamma(1,\lambda)$;
$\gamma(x,z) \times \gamma(y,z) = \gamma(x+y,z)$; and $\chi^2_n = \gamma(\frac{n}{2}, \frac{1}{2})$.

Examples:

$\alpha=0.05$: $c=0.00870,$
$\alpha_0=1$: $\alpha_1=c,$
$\alpha_0=0.7$: $\alpha_1=0.0183$, $\alpha_0=0.5$: $\alpha_1=0.0233$, $\alpha_0=0.3$: $\alpha_1=0.0299$

$\alpha=0.025$: $c=0.00380,$
$\alpha_0=1$: $\alpha_1=c,$
$\alpha_0=0.7$: $\alpha_1=0.0080$, $\alpha_0=0.5$: $\alpha_1=0.0102$, $\alpha_0=0.3$: $\alpha_1=0.0131$
Illustration of the rejection region with one-sided p-values extension to two-sided level alpha=0.05
Bauer&Köhne Design versus an Example of a Group Sequential Design
Statistical Model:
One-sided test with a two-stage design in the case of normal distributions with essentially known variance, in our notation the Brownian motion model.

The time of the interim analysis will be planned, however, the sample size of the second stage will be calculated on the basis of the observations at the time of the interim analysis.
Comparison of two groups, variables X and Y normally distributed with E of $\mu_X$ and $\mu_Y$, known equal SD of $\sigma$, $\delta = (\mu_Y - \mu_X)/\sqrt{2}\sigma$, $n_1$ per group at interim, standardized statistic $T_1 = TObs$, extended by $n_2(T_\text{obs})$ per group to $n$ per group, standardized statistic $T$.

Conditional Power function:

$$CP_\delta (T_\text{obs}) = P_\delta (T > k(\alpha) \mid T_1 = T_\text{obs}) = 1 - \Phi\left(\frac{k(\alpha)\sqrt{n_1 + n_2} - T_\text{obs} \sqrt{n_1 + n_2}\delta}{\sqrt{n_2}}\right)$$

For $\delta = 0$ and $k(\alpha) = z_{1-\alpha}$ maximal (where $\eta = n_2/n_1$):

$T_\text{obs} > k(\alpha)$: $\rightarrow 1$ for $\eta \rightarrow 0$ ($n_2 = 0$)

$T_\text{obs} < k(\alpha)$: $1 - \Phi\left(\sqrt{k^2(\alpha) - T^2_\text{obs}}\right)$ at $\eta = \left(\frac{k(\alpha)}{T_\text{obs}}\right)^2 - 1$
Indiscriminately Extending:
The type I error level $\alpha$ may be inflated up to

$$\alpha_{\text{max}} = \alpha + \frac{1}{4} \exp\left(-\frac{1}{2} (\Phi^{-1}(1-\alpha))^2\right),$$

where $\Phi$ is the distribution function of the standard normal distribution,
e.g. from $\alpha=0.05$ to $\alpha_{\text{max}}=0.1146$.

(Note: $\alpha_{\text{max}} = \int_{-\infty}^{\infty} CP_0(z) \phi(z) \mathrm{d}z$ with $CP_0(z)$ the maximum,
where $\phi$ denotes the density function of the standard normal distribution.)
Extension procedure maintaining the type I error level:
Choose an increasing function $A$ with range $[0, 1]$, denoted as conditional error function, satisfying
\[ \int_{-\infty}^{\infty} A(z) \phi(z) \, dz = \alpha \]
where $\phi$ denotes the density function of the standard normal distribution.
Assume that $T_{obs}$ is the value of the statistic at the time of the interim analysis based on $n_1$ observations per group (first stage) and that at this time $n_2$ observations per group (second stage) will be planned for the remaining trial (e.g. on the basis of conditional power), then calculate

$$c = \frac{\sqrt{n_1} T_{obs} + \sqrt{n_2} \Phi^{-1}(1-A(T_{obs}))}{\sqrt{n_1+n_2}}$$

as the critical value for the final test statistic based on $n_1+n_2$ observations per group.
Circular conditional error functions defined by

\[
A_{\text{cir}}(z) = \begin{cases} 
0 & \text{if } z < \Phi^{-1}(1-p_{\text{nfu}}) \quad \text{futility} \\
1 - \Phi\left(\sqrt{k^2 - z^2}\right) & \Phi^{-1}(1-p_{\text{nfu}}) \leq z < k \quad \text{continue} \\
1 & z \geq k \quad \text{rejection}
\end{cases}
\]

with \( \alpha_{\text{max}} = 1 - \Phi(k) + \frac{1}{4}\exp\left(-\frac{1}{2}k^2\right) \) in the case of \( p_{\text{nfu}} = 1 \)

and linear conditional error functions (not considered here) were proposed as conditional error functions.

Examples (circular) to achieve \( \alpha_{\text{max}} = \alpha \) for \( \alpha = 0.05 \):

- \( p_{\text{nfu}} = 0.15 \): \( k = 1.82 \)
- \( p_{\text{nfu}} = 0.25 \): \( k = 1.875 \)
- \( p_{\text{nfu}} = 0.5 \): \( k = 1.95 \)
Circular conditional error function
one-sided level alpha = 0.05, futility probability = 0.5
The idea behind the proposed method is to construct a multi-stage p-value combination test of the Bauer&Köhne type using the inverse normal method of combining independent p-values by Hedges&Olkin and classical group sequential designs with equidistant time schedule in the Brownian motion model.

The number of stages have to be fixed in advance. After each stage design adaptations may be performed as flexible as in the Bauer&Köhne designs.
The m-stage procedure:

Selecting the design:

Choose a classical group sequential design with m stages and equidistant time schedule in the Brownian motion model, and calculate the critical values for the m stages, e.g. choose a design of the $\Delta$ class (Wang&Tsiatis) where the critical boundaries are proportional $t^{\Delta-0.5}$, particularly the Pocock design ($\Delta=0.5$) or the O’Brien&Fleming design ($\Delta=0$).
The m-stage procedure:

Decision making:

Perform m stages of the trial sequentially resulting in independent one-sided p-values $p_1$, $p_2$, ..., $p_m$.

At the end of stage $k$, use the group sequential decision rule of the chosen design for the statistic given by

$$
\frac{1}{\sqrt{k}} \sum_{i=1}^{k} \Phi^{-1}(1-p_i).
$$
Another approach of adaptively designing clinical trials is based on (self)-designing the statistic for testing the null-hypothesis:

The observed units are assumed to be i.i.d. normally distributed with known variance. One statistical test will be performed where the statistic may assumed to be standard normal distributed under the null-hypothesis. The statistic $Z$ consists of $m$ independent under the null-hypothesis standard normal distributed sub-statistics $Z_1, Z_2, ..., Z_m$ representing the results of the sub-samples of the $m$-stage procedure. $m$ needs not to be specified in advance.
The Variance Spending Method and Self-Designing (LD Fisher, 1998)

\[ Z = \sum_{i=1}^{m} a_i Z_i \] is a weighted statistic with positive weights \( a_i \) necessarily satisfying \( \sum_{i=1}^{m} a_i^2 = 1 \).

At stage \( k-1 \) the next stage \( k \) will be specified: \( a_k \), the sample size of sub-sample \( k \) and the choice of the sub-statistic may depend on the data of the \( k-1 \) observed sub-samples.

The procedure is self-designing if a rule for these specifications is fixed in advance as done in Shen&Fisher (1999).
The Variance Spending Method and Self-Designing (LD Fisher, 1998)

At stage $k$ the partial sum $\sum_{i=1}^{k} a_i^2 \leq 1$ is the variance of
the partial statistic $\sum_{i=1}^{k} a_i Z_i$ constructed so far.
Thus the variance is spent during the course of the trial.

$k$ is the final stage $m$ if the variance is summed up to be 1 by specification of $a_k$.

At the end of the final stage, the procedure completes with testing the null-hypothesis using the statistic $Z$. 
Arbitrary response variables: Use the inverse normal transformation for the one-sided p-values of the sub-samples to achieve standard normal distributed sub-statistics under the null-hypothesis.

Significance testing at the times of the interim analyses: Use $\chi^2$ distributed sub-statistics with degrees of freedom instead of standard normal sub-statistics and weights. Spent degrees of freedom instead of variance. Apply the inverse $\chi^2$ transformation instead of the inverse normal transformation. As in the Bauer&Köhne approach non-stochastic curtailing can be used for early rejection of the null-hypothesis.
Adaptive Designs
In a clinical trial testing differences (two-sided at a significance level of 0.05) two two-stage Bauer&Köhne designs for simultaneously testing each of the two sides of the alternative (superiority and inferiority) at a level of $\alpha=0.025$ were fixed in the study protocol. The tests were based on Fisher‘s combination rule without stopping for futility ($\alpha_0=1$). The one-sided p-values $p_1=0.00381$ and $p_2=0.997$ for testing superiority were calculated from the results of the two stages. The corresponding p-values for testing inferiority were $1-p_1$ and $1-p_2$.

Task: Conduct the test procedure. Thereby state the reasons for statistical decision making. Discuss the decisions for the purpose of consistent recommendations from the analysis.
CRP-Approach
Müller H-H, Schäfer H (2001, 2004): Design adaptation during the course of a trial, the conditional rejection probability principle and approach
Disadvantages of Proposed Adaptive Methods

Only two stages: Proschan & Hunsberger
Fixed number of stages: all adaptive designs
One confirmatory final analysis, iteratively scheduling interim analyses for flexibly fixing the final analysis: LD Fisher, Hartung

Redesign of confirmatory interim analyses including number, schedule and alpha-spending of analyses: Despite of non-stochastic curtailment not possible so far
Disadvantages of Proposed Adaptive Methods

Decisions possibly/usually not based on the sufficient statistic: Bauer&Köhne, Lehmacher&Wassmer, LD Fisher, Hartung

Losses in efficiency compared to (optimised) group sequential designs: all proposed adaptive procedures so far

Design adaptations only after planned interim analyses: all proposed adaptive procedures so far
• carefully plan a trial with an efficient (conventional) (group sequential) design controlling the type I error rate
• the option of a design modification is available at any time during the course of the trial
• the option of a modification has no price
• design modifications can be considered and may be performed conditionally efficient
• using the CRP-principle leads to the most flexible type I error rate preserving methods
The principle:
For the control of the type I error rate it is „necessary“ and sufficient to always control the conditional type I error rate(s) during the course of the trial where conditional means conditionally on all so far observed data.

A formula for application:
\[ E_0(\psi \circ Y \mid Z = z) \leq \text{most possible} = E_0(\varphi \circ X \mid Z = z) \]

This principle offers most flexible design modifications while statistically monitoring a trial.
### Adaptive Designs

<table>
<thead>
<tr>
<th>Start of study ( (n_1+?) )</th>
<th>Start of study ( (n) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Learning phase ( (n_1) )</td>
<td>- Planning of whole study ( (n) )</td>
</tr>
<tr>
<td>- UMPU-Test</td>
<td>- UMPU-test or GSD</td>
</tr>
<tr>
<td>- Predefinition of a combination rule ( c(p_1, p_2) )</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>End of learning phase</th>
<th>At any time</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Calculation of ( p_1 ) ( (n_1) )</td>
<td>- Calculation of CRP ( (n_1) )</td>
</tr>
<tr>
<td>- Planning of test phase ( (n_2) )</td>
<td>- Tentative re-design ( (n^*) )</td>
</tr>
<tr>
<td>UMPU-test (if so an AD)</td>
<td>maintaining the CRP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>End of test phase ( (n_1+n_2) )</th>
<th>End of study ( (n^*=n_1+n_2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Calculation of ( p_2 ) ( (n_2) )</td>
<td>- Test decision according to modified design</td>
</tr>
<tr>
<td>- Test decision using ( c(p_1, p_2) )</td>
<td></td>
</tr>
</tbody>
</table>

### CRP-Principle

- Planning of whole study \( (n) \)
- UMPU-test or GSD
- Tentative re-design \( (n^*) \) maintaining the CRP
- Test decision according to modified design
Advantages of the CRP-Approach Compared with Adaptive Designs

• Options for design adaptations without price
• Interim analyses user-defined during the course of the trial
• Most flexible confirmatory method
• UMPU tests resp. optimised GSDs
• Facilitates ever the (conditionally) efficient design for the remaining part of the trial
Treatments: experimental (E) and control (C)

Number of patients: \( n = n_E + n_C, n_E = n_C = 100 \)

Allocation fraction: \( \tau = \frac{1}{2} \)

Primary endpoint: \( X_i \) independent, normally distributed with location parameters \( \mu_E \) and \( \mu_C \) and with known variance \( \sigma^2 \)

Effect parameter: \( \vartheta = \frac{\mu_E - \mu_C}{\sigma} \)
The test procedure:

two-stage design by Fleming et al., specifically

\( \alpha = 0.05 \) (two-sided)

one planned interim analysis at \( t_1 = 0.5 \) \( (n_E = n_C = 50) \)

\( \alpha_1 = 0.02 \), leaving nominal \( \alpha_2 = 0.0379 \)

\( \delta^* = 2.8799 \) for a power of \( 1 - \beta = 0.8 \) in the BM model

power \( 1 - \beta = 0.8 \) (power of 0.386 at interim) for the detectable difference of

\[
\nu^* = \frac{\delta^*}{\sqrt{n} \, \tau (1 - \tau)} = \frac{2.8799}{\sqrt{50}} = 0.4037
\]
Calculation of the CRP: An Example

Assume:

unplanned interim analysis after assessment of the first $n_E=n_C=30$ patients
(information time $s=0.3$)

- calculate the effect size estimate

\[ \hat{\theta} = \frac{\bar{X}_E - \bar{X}_C}{\sigma} \]

- based on these assessed patients

- calculate the upper and the lower conditional rejection error probability
The critical values of the two-stage Fleming et al. design at the times of the planned analyses are at $t_1=0.5$: $cv=\pm 2.3264$ and at $t_2=1$: $cv=\pm 2.0758$. Suppose an effect size of $\hat{\vartheta} = 0.37$ was observed at the time $s=0.3$ of the unplanned interim analysis, i.e. $T_{s,\text{obs}} = \sqrt{n_{\text{obs}} \tau (1-\tau)} \, \hat{\vartheta} = \sqrt{15 \cdot 0.37} = 1.4330$.

Calculate (e.g. SAS IML: CALL SEQ):
upper rejection error probability $= 0.078875$
lower rejection error probability $= 0.000314$

Plan the remaining design with these rejection error probabilities as type I error levels.
Calculation of CRPs for Group Sequential Designs in the Brownian Motion Model

$s \in [0, 1]$ information time of interim analysis (planned or unplanned), $t \in ]s, 1]$ information time of an (interim) analysis planned later

Reparameterisation (unconditional with ‘\('):$

\delta' = \delta \sqrt{1-s}, \quad t' = \frac{t-s}{1-s} \in ]0,1],

T'_{t'} = \frac{T_t \sqrt{t - T_{s,\text{obs}}}}{\sqrt{t-s}} = \frac{1}{\sqrt{t'}} B_{\delta',t'},

CV' = \frac{CV \sqrt{t - T_{s,\text{obs}}}}{\sqrt{t-s}}
PROC IML;
START CRP(time, alpha_low, alpha_upp, z, tau, delta, crp_low, crp_upp);
epsilon=0.00000001;
cv_low=SQRRT(time)#PROBIT(alpha_low);
cv_upp=SQRRT(time)#PROBIT(1-alpha_upp);
bounds=cv_low//cv_upp;
x=SQRRT(tau)#z;
time=time-\tau;
bounds=bounds-x;
bounds=bounds-delta*(time//time);
stretch=1/time[1];
time=stretch*time;
bounds=\sqrt{stretch}*bounds;
tscale=time[1,2:NCOL(time)]-time[1,1:NCOL(time)-1];
CALL SEQ(prob,bounds) TSCALE=tscale EPS=epsilon;
crp_low=SUM(prob[1,]);
size=1-(prob[2,]-prob[1,])[NCOL(prob)];
crp_upp=size-crp_low;
FINISH;
time={0.392157 0.647059 1};
bounds={2.6175 2.6783 2.0211};
alpha_low=1-PROBNORM(bound);
alpha_upp=alpha_low;
z=PROBIT(1-0.6112);
tau=0.145098;
delta=0;
CALL CRP(time, alpha_low, alpha_upp, z, tau, delta, crp_low, crp_upp);
PRINT crp_low crp_upp;
QUIT;
Example: CRP for $\delta=0$

upper rejection (objective to claim superiority)
The critical values of the two-stage Fleming et al. design at the times of the planned analyses are at $t_1=0.5$: $cv=\pm 2.3264$ and at $t_2=1$: $cv=\pm 2.0758$.

A recalculation of the sample size may be performed to achieve a conditional power of $0.8=80\%$ if the effect size parameter would have the value as initially assumed at the planning stage or as estimated in the unplanned interim analysis, i.e. $\varphi = \hat{\varphi} = 0.4037$ or $\varphi = \hat{\varphi} = 0.37$, respectively:
Suggest that again a two stage design of similar type as the initial group sequential design would be applied in case of decision for a redesign. Assume a sample size inflation of 6% for such a design compared with a fixed sample design.

The upper rejection error probability was 0.078875. According to this CRP and the standard normal quantile \( z_{1-0.078875} = z_{0.921125} = 1.4127 \), the recalculated sample size for the remaining study then amounts to \( 1.06 \times 62.36 \times 2 = 134 \) or \( 1.06 \times 74.24 \times 2 = 158 \), respectively, instead of 140.
Assume:

planned interim analysis after assessment of the first $n_E=n_C=50$ patients
(information time $s=0.5$)

calculate the effect size estimate

\[ \hat{\theta} = \frac{X_E - X_C}{\sigma} \]

based on these assessed patients

calculate the upper and the lower conditional rejection error probability
The critical values of the two-stage Fleming et al. design at the times of the planned analyses are at \( t_1 = 0.5 \): \( cv = \pm 2.3264 \) and at \( t_2 = 1 \): \( cv = \pm 2.0758 \).

Suppose an effect size of \( \hat{\vartheta} = 0.32 \) was observed at the time \( s = 0.5 \) of the planned interim analysis, i.e.

\[
T_{s,\text{obs}} = \sqrt{n_{\text{obs}} \tau (1-\tau)} \hat{\vartheta} = \sqrt{25 \cdot 0.32} = 5 \cdot 0.32 = 1.6.
\]

Calculate (e.g. R: pnorm):

Reparameterisation

\[
cv' = \frac{\pm 2.0758 - 1.6 \cdot \sqrt{0.5}}{\sqrt{1-0.5}} = \begin{cases} 1.3356 \\ -4.5356 \end{cases}
\]

\[
\text{CRP}_{\text{upper}}(\vartheta=0) = 1 - \text{pnorm}(1.3356) = 0.090836
\]

\[
\text{CRP}_{\text{lower}}(\vartheta=0) = \text{pnorm}(-4.5356) = 0.00000287
\]

Redesign with these CRPs as type I error levels.
The remaining critical values of the two-stage Fleming et al. design at the time of the planned final analysis are at $t_2=1$: $cv=\pm 2.0758$.

A recalculation of the sample size may be performed to achieve a conditional power of 0.8=80% if the effect size parameter would have the value as initially assumed at the planning stage or as estimated in the planned interim analysis, i.e. $\phi = \hat{\phi} = 0.4037$ or $\phi = \hat{\phi} = 0.32$, respectively:
Suggest that again a two stage design of similar type as the initial group sequential design would be applied in case of decision for a redesign. Assume a sample size inflation of 6% for such a design compared with a fixed sample design.

The upper rejection error probability was 0.090836. According to the corresponding standard normal quantile $1.3356 = z_{1-0.090836}$ calculated before, the recalculated sample size for the remaining study then amounts to $1.06 \times 58.17 \times 2 = 124$ or $1.06 \times 92.58 \times 2 = 198$, respectively, instead of 100.
Suppose that the trial is designed as a two-armed parallel group trial with normal outcome variable and that a group sequential design has been selected with three equally sized stages and a spending function of the form $\alpha t^\rho$, with $\alpha = 0.025$, $\rho = 3$ and information time $0 < t \leq 1$. This spending function will be used for the construction of both upper and lower boundaries, resulting in a two-sided group sequential design with overall type I error risk of 5 per cent. The selected spending function is a member of a useful class of spending functions introduced by Jennison and Turnbull [24, Chapter 7.2], who remarked that the choice $\rho = 3$ will produce boundaries similar to the O’Brien and Fleming [2] boundaries. The corresponding one-sided nominal alpha levels at information times $1/3$, $2/3$ and $1$ are $0.00093$, $0.00691$ and $0.02228$, and the characteristic drift values to achieve a power of $1 - \beta = 0.8$ are $\pm 2.829$. Under the assumption of a normal outcome variable with known variance $\sigma^2$ in both groups, the required sample size per treatment group is $n = \delta^2 \sigma^2 / \Delta^2 (1 + r)^2 / r$, where $\delta$ is a characteristic drift value, $\Delta = \mu^E - \mu^C$ is the mean difference between the two treatment groups, and $r$ is the randomization ratio. If the pre-study variance estimate is $\sigma^2 = 1$ and the minimal clinically relevant difference is $\Delta = 0.5$, a sample size of $n \geq 64.04$ is required to achieve a power of $1 - \beta \geq 0.8$, which can be realized by a three-stage design with 22 patients per stage and treatment group.

the formula provides the total sample size
• A general principle for the re-design of a (group sequential) experiment at any time during its course was developed maintaining type I error control.

• However, such a method did not replace careful statistical planning of the experiment:

Clearly, our method should never be used as an argument not to plan the statistical design of an experiment as careful as usual. Before starting the trial an efficient design for the experiment has to be set up, including interim analyses based on ethical and economic grounds. Design changes during the course of the trial always need careful considerations, especially when an unplanned interim look is performed. As for interim analyses for early stopping, the interim analysis for a decision to change the design requires high data quality, i.e. complete and valid data. We recommend that before an interim look is performed the exact procedure should be fixed in the study protocol or in an amendment. If a design modification is made, a protocol amendment should be made immediately including the reasons for changing the design and the complete description of the new design.
CRP-Approach

Outlook
CRP-Approach

Exact Procedures Incorporating Nuisance Parameters
CRP-Principle
And
t-Test

Is there an efficient method for sample size recalculation?

Aim:
Fit the CRPs of the modified test as well as possible.
$t$-test, $\alpha=0.025$ one-sided, $n=65$

at interim: $n_1=15$, $s_{n_1} = 23.2$, $z_{n_1} = 5.5$
Confidence Intervals Following a Group Sequential Design Incorporating a Sample Size Adjustment With the CRP-Approach
Confidence sets for the effect size parameter are constructed via the correspondence with a family of tests statistically testing all the single point null-hypotheses of the parameter values.

- the confidence set consists of those point values which cannot be rejected

- the test of a point value rejects the null-hypothesis iff the point value is not in the confidence set
Proposed by Müller & Schäfer 2001

Obviously, the CRP-principle can be applied to any test of a single point null-hypothesis.

Notation:
1−α: level of confidence
δ: single point value linked to the parameter of interest; refers to testing the respective single point null-hypothesis if used as index
R(α): rejection region of a level α test
Proposed by Müller & Schäfer 2001

Notation (continued):

\( z \): an interim result in the continuation region

\( ' \): remaining or modified design from interim analysis (conditional design at interim)

\( z' \): final result in the stopping region

\[ \text{CRP}(\alpha, z) := P(R(\alpha)|Z=z) \]

If \( \delta \) and \( \delta' \) are linked to the same parameter of interest, a self-evident level \( 1-\alpha \) confidence set is given by

\[ \{ \delta: z' \notin R'(\delta)(\text{CRP}_\delta(\alpha, z)) \} \]
In the following the focus is on the essential example where the parameter $\delta$ is the drift parameter of a Brownian motion:
NSCLC Trial: Consideration at the Planning Stage

Worse Scenario For a Design Extension With the CRP-Principle

Interim result: 100 of 255 deaths, $p=0.15866$ one-sided

Significance?

Predictive Power: 29.4%
Conditional Power:
66.6% for HR=14/20
34.5% for HR=16/20

Design extension desirable

Calculation of CRPs:
CRP for $H_{1+}=0.03720$
CRP for $H_{1-}=0.00034$
• initial group sequential design

• information time: 100 of 255 deaths

• observed statistic: $p = 0.15866$ one-sided

⇒ planning of a new group sequential design

with:  
\begin{align*}
\text{CRP_LOW} & \quad \text{CRP_UPP} \\
0.0003425 & \quad 0.0372049
\end{align*}
### Group Sequential Design for the Extended Part of the NSCLC Trial

<table>
<thead>
<tr>
<th>deaths</th>
<th>critical values</th>
<th>nominal α</th>
<th>cumulative α</th>
<th>detectable difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>265 (+100)</td>
<td>+2.4581</td>
<td>0.00698</td>
<td>0.00698</td>
<td>from 16 onto 24 months</td>
</tr>
<tr>
<td>565 (+100)</td>
<td>+1.8250, −3.3955</td>
<td>0.03400, 0.00034</td>
<td>0.03720, 0.00034</td>
<td>from 16 onto 20 months</td>
</tr>
</tbody>
</table>

Detectable difference for achieving the desired (conditional) power $\delta^* = 2.6516$
NSCLC Trial

Order of Initial Group Sequential Design for Delta-Shift by Tsiatis, Rosner & Mehta 1984

\[ \delta^* = 2.8424 \]

Upper Rejection

CRP for confidence sets

\[ \delta^*/ \delta \]
NSCLC Trial

Order of Initial Group Sequential Design for Delta-Shift by Tsiatis, Rosner & Mehta 1984

\[ \delta^* = 2.8424 \]

\[ \text{CRP for confidence sets} \]

\[ \text{Lower Rejection} \]

\[ \text{delta/delta}^* \]
NSCLC Trial

Tsiatis, Rosner & Mehta Type Confidence Sets After Modification With the CRP-Principle

\[ \delta^* = 2.8424 \]
NSCLC Trial

Tsiatis, Rosner & Mehta Type Confidence Sets After Modification With the CRP-Principle

\[ \delta^* = 2.8424 \]

\[ \text{delta/delta}^* \text{ limits of confidence sets} \]

\[ \text{statistic (stop at interim)} \]
Exercise

CRP-Approach
The 3 stage group sequential design of the NSCLC trial: See above.

Assume at the first interim analysis the one-sided log-rank p-value for testing superiority was 0.02 (or 0.15866). For the discussion of a re-design of the trial the upper and the lower conditional rejection probabilities should be calculated in R using the drift function of the ldbounds package. Which lower boundaries, which upper boundaries and which time series should be specified?

To calculate the rejection error probabilities (lower and upper), which value for the drift should be specified? To calculate the conditional power given the hazard ratio of 14/20, which value for the drift should be specified?
Selection of a Group Sequential Design For a Survival Trial in ACC
## Example: Survival Analysis for a Clinical Trial in Oncology with Patients Suffering from ACC

<table>
<thead>
<tr>
<th>Patients:</th>
<th>locally advanced and metastatic adrenocortical carcinoma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design:</td>
<td>multicentre, multinational, open-label, randomised, parallel groups</td>
</tr>
<tr>
<td>Experimental treatment (E):</td>
<td>chemotherapy etoposide, doxorubicin, cisplatin plus mitotane (EDP-M)</td>
</tr>
<tr>
<td>Control (C):</td>
<td>chemotherapy streptozotocin plus mitotane (Sz-M)</td>
</tr>
<tr>
<td>Objective:</td>
<td>to demonstrate superiority of E over C</td>
</tr>
<tr>
<td>Primary endpoint:</td>
<td>survival time after randomisation</td>
</tr>
</tbody>
</table>
1:1 randomisation, allocation fraction \( \tau = \frac{1}{2} \)

survival analysis, logrank-test, CI for HR \( \theta \)

\( \alpha = 5\% \) two-sided, power 80%

survival times: (nearly) exponential

detectable difference:

control: median 18 months
experimental: median 27 months

sample size: as small as possible
as large as necessary

Recruitment rate: 5 patients per month

recruitment phase: ? months

trial duration: ? months
First step of your written take home exam:

Based on the exercise (group activity with computer practical) and the discussion make your choice of a GSD for the trial in ACC.

Write a proposal on the selection of the GSD. Thereby describe the GSD for the ACC trial and the steps of its development. State reasons and motivate your selection!

With Supplementary Appendix
And Study Protocol

See also http://www.firm-act.org
• 1:1 randomisation, allocation fraction $\tau = \frac{1}{2}$

• survival analysis, logrank-test, CI for HR $\theta$

• $\alpha=5\%$ two-sided, power 80%

• survival times: (nearly) exponential

detectable difference:

  - control: median 18 months
  - experimental: median 27 months

• sample size: as small as possible
  as large as necessary

• Recruitment rate: 5 patients per month

• recruitment phase: 60 months

• trial duration: 78 months
### Group Sequential Design of the FIRM-ACT Trial

<table>
<thead>
<tr>
<th>deaths</th>
<th>critical values</th>
<th>nominal $\alpha$</th>
<th>cumulative $\alpha$</th>
<th>detectable difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>79</td>
<td>±2.5821</td>
<td>0.00982</td>
<td>0.00982</td>
<td>from 18 onto 39 months</td>
</tr>
<tr>
<td>122</td>
<td>±2.5501</td>
<td>0.01077</td>
<td>0.01704</td>
<td>from 18 onto 33 months</td>
</tr>
<tr>
<td>200</td>
<td>±2.0432</td>
<td>0.04103</td>
<td>0.05</td>
<td>from 18 onto 27 months</td>
</tr>
</tbody>
</table>

BM: detectable difference for achieving the desired power $\delta^* = 2.85692$
Two interim analyses with alpha spending after 79 and 138 observed deaths (no early stopping) and the final analysis have been performed, the latter resulted in the nominal two-sided $p=0.22$ with positive trend on the basis of 211 observed deaths (see Supplementary Appendix).

Following the conducted group sequential design of the study-protocol with adjustment by the alpha-spending approach, one may calculate the median unbiased estimate of the HR and the 95% confidence interval for the HR (unclear, if provided in the Publication / Appendix).
Second step of your written take home exam:

Assume your choice of a GSD was implemented as initial design for FIRM-ACT. Set up an alpha-spending function for your GSD, e.g. by linear interpolation up to a certain amount of information from where only the final analysis spending full $\alpha$ will be performed.

Assume the results of FIRM-ACT at 79, 138 and 211 observed deaths. Adapt your GSD according to the alpha-spending function approach and conduct primary testing and HR estimation. For these results, discuss discrepancies compared to the article.
The second interim analysis was performed after 138 observed deaths resulted in the nominal two-sided p=0.22 with positive trend. Considering also a potential redesign with sample size adjustment, the decision was to continue, suggesting to perform the final analysis after 200 observed deaths (see Supplementary Appendix).

At the second interim analysis a sample size adjustment and a redesign of the GSD was considered according to the CRP-principle.
Third step of your written take home exam:

At the second interim analysis after 138 observed deaths, consider a redesign of your GSD (so far modified using the alpha-spending function approach) according to the CRP-principle.

Please calculate the one-sided conditional rejection error probability for testing superiority, and recalculate the sample size for a fixed sample design in the remaining trial (after continuation) suggesting a conditional power of 80%

1) keeping the HR=1/1.5=0.67 to be detected and
2) suggesting HR=0.8 to be detected.