Group-Sequential and Two-Stage Designs

Helmut Schütz
Group-Sequential Designs

Dealing with Uncertainty: Group-Sequential Designs

• Long and accepted tradition in clinical research (phase III)

• Fixed total sample size ($N$) and – in BE – one interim analysis
  – Requires two assumptions
    – A ‘worst case’ CV for the total sample size and
    – A ‘realistic’ CV for the interim
  – All published methods were derived for superiority testing, parallel groups, normal distributed data with known variance, and the interim analysis at exactly $N/2$
    – That’s not what we have in BE
      » Equivalence (generally crossover), lognormal data with unknown variance
      » Due to drop-outs, the interim might not be exactly at $N/2$
        (might inflate the Type I Error)
Group-Sequential Designs

Dealing with Uncertainty: Group-Sequential Designs

- Fixed total sample size ($N$) and – in BE – one interim analysis
  - First proposal by Gould (1995) in the field of BE did not get regulatory acceptance in Europe
  - Asymmetric split of $\alpha$ is possible, i.e.,
  - a small $\alpha$ in the interim (i.e., stopping for futility) and
  - a large one in the final analysis (i.e., only small sample size penalty)
  - Examples
    - Haybittle/Peto ($\alpha_1$ 0.001, $\alpha_2$ 0.049)
    - O’Brien/Fleming ($\alpha_1$ 0.005, $\alpha_2$ 0.048)
  - Not developed for crossover designs and sample size re-estimation (fixed $n_1$ and variable $N$): Lower $\alpha_2$ or $\alpha$-spending functions (Lan/DeMets, Jennison/Turnbull) are needed in order to control the Type I Error
  - Zheng et al. (2015) for BE in crossovers ($\alpha_1$ 0.01, $\alpha_2$ 0.04) controls the TIE
Excursion 1

Type I Error

Haybittle/Peto
\( \alpha_1 0.001, \alpha_2 0.049 \)

Maximum 0.05849
\( \alpha_2 0.0413 \) needed to control the TIE

O’Brien/Fleming
\( \alpha_1 0.005, \alpha_2 0.048 \)

Maximum 0.05700
\( \alpha_2 0.0415 \) needed to control the TIE

Zheng et al.
\( \alpha_1 0.01, \alpha_2 0.04 \)

Maximum 0.04878
Group-Sequential Designs

Review of Guidelines

• Australia (2004), Canada (Draft 2009)
  – Application of Bonferroni’s correction ($\alpha_{adj} 0.025$)
  – Theoretical Type I Error $\leq 0.0494$
  – For CVs and samples sizes common in BE the TIE generally is $\leq 0.04$

• Canada (2012)
  – Pocock’s $\alpha_{adj} 0.0294$
  – $n_1$ based on ‘most likely variance’ + additional subjects in order to compensate for expected dropout-rate
  – $N$ based on ‘worst-case scenario’
  – If $n_1 \neq N/2$ relevant inflation of the Type I Error is possible!
    – $\alpha$-spending functions can control the TIE
    – Are not mentioned in the guidance...
(Adaptive) Sequential Two-Stage Designs

Dealing with Uncertainty:
(Adaptive) Sequential Two-Stage Designs

- Fixed stage 1 sample size \( n_1 \), sample size re-estimation in the interim analysis
  - Generally a fixed GMR is assumed
  - All published methods are valid only for a range of combinations of stage 1 sample sizes, CVs, GMRs, and desired power
  - Fully adaptive methods (i.e., taking also the GMR of stage 1 into account) are problematic
    - May deteriorate power and require a futility criterion
    - Simulations mandatory
  - With one exception (inverse normal method) no analytical proof of controlling the TIE exists
    - It is the responsibility of the sponsor to demonstrate (e.g., by simulations) that the consumer risk is preserved
(Adaptive) Sequential Two-Stage Designs

Dealing with Uncertainty: (Adaptive) Sequential Two-Stage Designs

• Fixed stage 1 sample size \( n_1 \), sample size re-estimation in the interim analysis
  – Two ‘Types’ (Schütz 2015)
    1. The same adjusted \( \alpha \) is applied in both stages – regardless whether a study stops in the first stage or proceeds to the second stage
    2. An unadjusted \( \alpha \) may be used in the first stage, dependent on interim power
Type 1 and Type 2

100(1 – 2\(\alpha_{\text{adj}}\)) CI

BE?

yes / stop

Interim power based on GMR, \(\alpha_{\text{adj}}\), and observed CV

100(1 – 2×0.05) CI

Interim power based on GMR, \(\alpha_{\text{adj}}\), and observed CV

100(1 – 2\(\alpha_{\text{adj}}\)) CI

\(\geq \pi\)

yes / stop

Total sample size \(N\) based on GMR, \(\alpha_{\text{adj}}\), \(\pi\), and observed CV

Stage 2 with \(n_2 = N – n_1\)

100(1 – 2\(\alpha_{\text{adj}}\)) CI using pooled data of both stages (\(\alpha_{\text{adj}}\))

Pass

Pass or fail

Fail

Pass or fail

Pass
(Adaptive) Sequential Two-Stage Designs

Methods by Potvin et al. (2008) first validated framework in the context of BE

- Supported by the ‘Product Quality Research Institute’ (FDA/CDER, Health Canada, USP, AAPS, PhRMA, …)
- Inspired by conventional BE testing and Pocock’s $\alpha_{adj}$ 0.0294 for GSDs
  - A fixed GMR is assumed (only the CV in the interim is taken into account for sample size re-estimation)
    GMR in the first publication was 0.95; later extended to 0.90 by other authors
  - Target power 80% (later extended to 90%)
(Adaptive) Sequential Two-Stage Designs

Frameworks for crossover TSDs

- **Stage 1 sample sizes 12 – 60, no futility rules.**

<table>
<thead>
<tr>
<th>Reference</th>
<th>Type</th>
<th>Method</th>
<th>GMR</th>
<th>Target power</th>
<th>CV_w</th>
<th>α_adj</th>
<th>TIE_max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potvin et al. (2008)</td>
<td>1</td>
<td>B</td>
<td>0.95</td>
<td>80%</td>
<td>10 – 100%</td>
<td>0.0294</td>
<td>0.0485</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Montague et al. (2012)</td>
<td>2</td>
<td>D</td>
<td>0.90</td>
<td></td>
<td>0.0280</td>
<td>0.0518</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>B</td>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fuglsang (2013)</td>
<td>2</td>
<td>C/D</td>
<td>0.95</td>
<td>90%</td>
<td>10 – 80%</td>
<td>0.0274</td>
<td>0.0503</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>C/D</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Xu et al. (2015). GMR 0.95, target power 80%, futility for the (1–2α₁) CI.**

<table>
<thead>
<tr>
<th>Type</th>
<th>Method</th>
<th>CV_w</th>
<th>Futility region</th>
<th>α₁</th>
<th>α₂</th>
<th>TIE_max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E</td>
<td>10 – 30%</td>
<td>0.9374 – 1.0667</td>
<td>0.0249</td>
<td>0.0363</td>
<td>0.050</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>0.9492 – 1.0535</td>
<td>0.0248</td>
<td>0.0364</td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>E</td>
<td>30 – 55%</td>
<td>0.9305 – 1.0747</td>
<td>0.0254</td>
<td>0.0357</td>
<td>0.050</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>0.9350 – 1.0695</td>
<td>0.0259</td>
<td>0.0349</td>
<td>0.050</td>
<td></td>
</tr>
</tbody>
</table>
Excursion 2

Type I Error and power
- Fixed sample $2 \times 2 \times 2$ design ($\alpha$ 0.05). $GMR$ 0.95, $CV$ 10 – 80%, $n$ 12 – 72
Excursion 3

Type I Error and power

- ‘Type 1’ TSD (Potvin Method B, $\alpha_{adj}$ 0.0294). GMR 0.95, CV 10 – 80%, $n_1$ 12 – 72
(Adaptive) Sequential Two-Stage Designs

Review of Guidelines

• EMA (Jan 2010)
  – Acceptable
  – $\alpha_{adj} = 0.0294 = 94.12\%$ CI in both stages given as an example (i.e., Potvin Method B preferred?)
  – ‘... there are many acceptable alternatives and the choice of how much alpha to spend at the interim analysis is at the company’s discretion.’
  – ‘... pre-specified ... adjusted significance levels to be used for each of the analyses.’
  – Personal remarks
    – The TIE must be preserved. Especially important if ‘exotic’ methods are applied.
    – Does the requirement of pre-specifying both alphas imply that $\alpha$-spending functions or adaptive methods (where $\alpha_2$ is based on the interim and/or the final sample size) are not acceptable?
    – TSDs are on the workplan of the EMA’s Biostatistics Working Party for 2018...
(Adaptive) Sequential Two-Stage Designs

Review of Guidelines

- EMA Q&A Document Rev. 7 (Feb 2013)
  - The model for the combined analysis is (all effects fixed):
    \[
    \text{stage} + \text{sequence} + \text{sequence(stage)} + \text{subject(sequence \times stage)} + \text{period(stage)} + \text{formulation}
    \]
  - At least two subjects in the second stage
  - Personal remarks
    - *None* of the publications used `sequence(stage)`;
    - no poolability criterion – combining is always allowed, even if a significant difference between stages is observed
    - Simulations performed by the BSWP or out of the blue sky?
    - Modification shown to be irrelevant (Karalis/Macheras 2014). Furthermore, no difference whether subjects are treated as a fixed or random term (unless PE >1.20). Requiring two subjects in the second stage is unnecessary.

```
library(Power2Stage)
power.2stage(method="B", CV=0.2, n1=12, theta0=1.25)$pBE
[1] 0.046262
power.2stage(method="B", CV=0.2, n1=12, theta0=1.25, min.n2=2)$pBE
[1] 0.046262
```
(Adaptive) Sequential Two-Stage Designs

Review of Guidelines

- **Health Canada (May 2012)**
  - Potvin Method C recommended

- **FDA**
  - Potvin Method C / Montague Method D / Xu Method E/F recommended
    (Davit *et al.* 2013; 2\(^{nd}\) / 3\(^{rd}\) GBHI conferences, Rockville 2016 and Amsterdam 2018)

  - Acceptable; Potvin Method B preferred?
(Adaptive) Sequential Two-Stage Designs

Futility Criteria

- Futility rules (for early stopping) do not inflate the TIE, but may deteriorate power
  - Stopping criteria must be unambiguously stated in the protocol
  - Simulations are mandatory in order to assess whether power is sufficient:
    Introduction of [...] futility rules may severely impact power in trials with sequential designs and under some circumstances such trials might be unethical.
    [...] before using any of the methods [...], their operating characteristics should be evaluated for a range of values of $n_1$, CV and true ratio of means that are of interest, in order to decide if the Type I error rate is controlled, the power is adequate and the potential maximum total sample size is not too great.
    
    - Simulations uncomplicated with current software
      - Finding a suitable $\alpha_{adj}$ and validating for TIE and power takes ~20 minutes with the R-package Power2Stage (open source)

Fuglsang 2014

Jones/Kenward 2014
(Adaptive) Sequential Two-Stage Designs

Dropouts

- In the first stage
  - Not relevant because the actual $n_1$ is used

- In the second stage
  - A smaller total sample size translates into
    - a lower chance to show BE and hence,
    - also a lower Type I Error
  - Like in fixed sample designs the impact on power will be small
(Adaptive) Sequential Two-Stage Designs

Cost Analysis

• Consider certain questions
  – Is it possible to assume a best/worst-case scenario?
  – How large should the size of the first stage be?
  – How large is the expected average sample size in the second stage?
  – Which power can one expect in the first stage and the final analysis?
  – Will introduction of a futility criterion substantially decrease power?
  – Is there an unacceptable sample size penalty compared to a fixed sample design?
(Adaptive) Sequential Two-Stage Designs

Cost Analysis

- Example:
  - Expected CV 20%, target power is 80% for a GMR of 0.95. Comparison of a ‘Type 1’ TSD with a fixed sample design (n 20, 83.5% power).

<table>
<thead>
<tr>
<th>$n_1$</th>
<th>$E[N]$</th>
<th>Studies stopped in stage 1 (%)</th>
<th>Studies failed in stage 1 (%)</th>
<th>Power in stage 1 (%)</th>
<th>Studies in stage 2 (%)</th>
<th>Final power (%)</th>
<th>Increase of costs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>20.6</td>
<td>43.6</td>
<td>2.3</td>
<td>41.3</td>
<td>56.4</td>
<td>84.2</td>
<td>+2.9</td>
</tr>
<tr>
<td>14</td>
<td>20.0</td>
<td>55.6</td>
<td>3.0</td>
<td>52.4</td>
<td>44.5</td>
<td>85.0</td>
<td>+0.2</td>
</tr>
<tr>
<td>16</td>
<td>20.1</td>
<td>65.9</td>
<td>3.9</td>
<td>61.9</td>
<td>34.1</td>
<td>85.2</td>
<td>+0.3</td>
</tr>
<tr>
<td>18</td>
<td>20.6</td>
<td>74.3</td>
<td>5.0</td>
<td>69.3</td>
<td>25.7</td>
<td>85.5</td>
<td>+3.1</td>
</tr>
<tr>
<td>20</td>
<td>21.7</td>
<td>81.2</td>
<td>6.3</td>
<td>74.9</td>
<td>18.8</td>
<td>86.2</td>
<td>+8.4</td>
</tr>
<tr>
<td>22</td>
<td>23.0</td>
<td>87.2</td>
<td>7.3</td>
<td>79.8</td>
<td>12.8</td>
<td>87.0</td>
<td>+15.0</td>
</tr>
<tr>
<td>24</td>
<td>24.6</td>
<td>91.5</td>
<td>7.9</td>
<td>83.6</td>
<td>8.5</td>
<td>88.0</td>
<td>+22.9</td>
</tr>
</tbody>
</table>
Conclusions

- Do not blindly follow guidelines!
  Some current recommendations may inflate the patient’s risk and/or deteriorate power
- Published frameworks can be applied without requiring the sponsor to perform own simulations – although they could further improve power based on additional assumptions
- GSDs and TSDs are both ethical and economical alternatives to fixed sample designs
- Recently the EMA’s BSWP – unofficially! – expressed concerns about the validity of methods based on simulations
Rumors & Chinese Whispers (Part 1)

TSDs based on simulations

- One member of the PKWP (2015):
  - I made peace with these methods and accept studies – *if* the confidence interval is not too close to the acceptance limits.
  - Personal remark: *How close is ‘not too close’?*

- Assessors of ES, AT (2016):
  - Kieser/Rauch (2015) showed that the adjusted $\alpha_{adj} 0.0294$ used by Potvin *et al.* is Pocock’s for *superiority.*
    The correct value for *equivalence* is 0.0304 (Jennison/Turnbull 1999).
  - Hence, all studies evaluated with a 94.12% CI in both stages are more conservative than necessary. At least these studies should not be problematic.
    - Personal remarks
      » One could confirm ~0.0304 for ‘Method B’ in simulations
      » However, it is a misconception that 0.0304 is ‘universally valid’ for equivalence
      » *Other* settings (GMR, power) require *other* values – even for ‘Type 1’ TSDs
Rumors & Chinese Whispers (Part 1)

TSDs based on simulations

- Another member of the PKWP asked the BSWP *which* inflation of the Type I Error would be acceptable (2015). He gave 0.0501 as an example.
  - Answer: The TIE must not exceed 0.05.
    - Personal remark: Rounding of the CI as required by the GL leads to acceptance of studies (regardless the design) with CLs of 79.995% and/or 125.004% – which inflates the TIE up to 0.0508. The BSWP should mind its own business.

- One assessor (PT) saw a study rejected by one of his colleagues – although BE was shown (2016)
  - When asked why, the answer was:
    - ‘According to the BSWP Potvin’s methods are not acceptable.’
  - He was not aware of such a statement and asked for an official document
    - ‘Such a document does not exist but all statisticians in the agencies know this statement.’
The Assessor’s Dilemma

TSDs based on simulations

- If an assessor would like to accept TSDs he/she is facing a dilemma:
  - TSDs are stated in the GL and therefore, studies are submitted
  - The BSWP does not ‘like’ methods based on simulations and prefers methods which demonstrate by an analytical proof that the patient’s risk is preserved – which seemingly don’t exist
  - According to the BSWP even a TIE of 0.0501 is not acceptable
  - With one million simulations the significance limit (>0.05) is 0.05036
    - Most methods show a TIE below this limit (and many even <0.05)
    - However, with other seeds of the random number generator (slightly) different results are possible
  - It would be desirable to assess whether a passing study (with a CI close to the AR) has a relevant impact on the patient’s risk
- I developed an R-package (AdaptiveBE), which currently is evaluated by assessors in Portugal and Spain
Simulations vs. ‘analytical proof’

- In principle regulators prefer methods where the control of the TIE can be shown analytically
  - Promising zone approach (Mehta/Pocock 2011)
    Wrong: Superiority / parallel groups / equal variances. Critized by Emerson et al. (2011).
  - Inverse normal method (Kieser/Rauch 2015)
    Wrong: Not a proof but a claim. Slight inflation of the TIE (0.05026) in the supplementary material’s simulations.
  - Inverse normal approach / maximum combination test demonstrated to control the Type I Error (Wassmer and Brannath 2016, Maurer et al. 2018)
    - For 2×2×2 designs implemented in the R-package Power2Stage available at
      https://cran.r-project.org/package=Power2Stage
Rumors & Chinese Whispers (Part 2)

Simulations vs. ‘analytical proof’

• In principle regulators prefer methods where the control of the TIE can be shown analytically
  – Repeated confidence intervals (Bretz et al. 2009)
    Adapted for BE (König et al. 2014, 2015, Maurer et al., 2018)

• Both in the inverse normal approach and with repeated CIs the final $\alpha$ is adapted based on the study’s data
  – Is this compatible with the guideline’s ‘pre-specified’ $\alpha$?
  – According to discussions at the 3rd GBHI conference (Amsterdam, April 2018) most likely yes!
Simulations vs. ‘analytical proof’

  - Most proofs start with ...

  \[
  \text{Let us assume parallel groups of equal sizes and normal distributed data with } \mu = 0 \text{ and } \sigma = 1
  \]

  … followed by some fancy formulas.

Do these cases ever occur in reality?  

Peter Bauer
Thank You!

Open Questions?

Helmut Schütz
BEBAC
Consultancy Services for Bioequivalence and Bioavailability Studies
1070 Vienna, Austria
helmut.schuetz@bebac.at