Sample Size Recalculation in Internal Pilot Study Designs: A Review

Tim Friede
Warwick Medical School
The University of Warwick
t.friede@warwick.ac.uk
Acknowledgement

This presentation includes joint work with

- Meinhard Kieser (Heidelberg)
- Günther Müller-Velten (Novartis)
- Charles Mitchell (ETH Zürich)
Outline

- background and motivating examples
- designs with sample size re-estimation
- internal pilot study designs
  - blinded vs. unblinded sample size reviews
  - continuous and binary outcomes
  - superiority and non-inferiority
- further issues and conclusions
Background

• adequate sample sizes - why?
  – ethics, budget, and time
  – power for testing, precision for estimation

• uncertainty in planning $\rightarrow$ high risk of inadequate sample sizes

• solution: mid-course re-estimation of sample size
Example: St John’s Wort in Depression

- **objective**: to assess the efficacy and safety of St John’s wort in mild to moderate depression

- **design**: randomised, double-blind, placebo-controlled

- **endpoint**: change in HAMD from baseline to day 42

- **initial sample size estimate**: 128 (= $2 \times 64$) patients
  
  - power $1 - \beta = 0.80$, relevant difference $\Delta^* = 4$, SD $\sigma_0 = 8$
Example: St John’s Wort in Depression (cont.)

- **uncertainty in the planning phase**
  - SD of HAMD at end of therapy 4-14.5 (Linde & Mulrow 2000)
  - placebo effect: very variable in depression

- **design**: two-stage Bauer/Köhne design (IA with 60 patients)
  - sample size reestimation to address uncertainty regarding SD
  - early stopping to address variability regarding placebo effect
Example: Anti-hypertensive Trial

- **design**: randomized, double-blind, parallel group, active-controlled

- **patients** with hypertension and non-insulin dependent diabetes

- **primary endpoint**: proportion of patients who . . .
  - completed study on treatment (tolerability, safety)
  - with mean 24h blood pressure < 130/80 mmHg (syst./diast.) (efficacy)

- **non-inferiority margin**
  - defined in terms of risk differences: 10 percentage points
Example: Anti-hypertensive Trial (cont.)

• **sample size**
  
  – assumed overall response rate 70%
  
  – target power 80% \(\Rightarrow\) 330 patients per group

• **results**: overall response 42%
  
  – experimental treatment 133/327
  
  – control treatment 141/326

• **problem**: power 75% (assuming response 42%) rather than 80%
Designs with Sample Size Re-estimation

• interim analysis

  – estimation of **treatment effect**
  – hypothesis test (offers opportunity for early stopping)
  – basically two types
    * classical group sequential designs (e.g. Jennison & Turnbull 1999)
    * designs based on combination of $p$-values (e.g. Bauer & Köhne 1994)
  – sample size re-estimation could be based on observed treatment effect

• sample size review

  – estimation of **nuisance parameters** (e.g. variance), no hypothesis test
  – design with internal pilot study (e.g. Wittes & Brittain 1990)
Internal Pilot Study Design (Wittes & Brittain 1990)

• initial sample size estimation \( n_0 = n(\alpha, 1 - \beta, \Delta^*, \hat{\sigma}_0^2) \)
  – significance level \( \alpha \), desired power \( 1 - \beta \), clinically relevant effect \( \Delta^* \)
  – initial estimate \( \hat{\sigma}_0^2 \) of the nuisance parameter \( \sigma^2 \) (from other studies)

• sample size review:
  – after recruitment of \( n_1 = \pi n_0 \) patients (e.g., \( \pi = 1/2 \))
  – estimation of nuisance parameter \( \rightarrow \hat{\sigma}^2 \)
  – sample size re-estimation \( \tilde{N} = n(\alpha, 1 - \beta, \Delta^*, \hat{\sigma}^2) \)
    * ”restricted”: \( n_2 = \max(n_0, \tilde{N}) - n_1 \)
    * ”unrestricted”: \( n_2 = \max(n_1, \tilde{N}) - n_1 \) (Birkett & Day 1994)

• final analysis
  – estimation of treatment effect and hypothesis test
  – with all \( n_1 + n_2 \) patients
Sample Size Re-estimation and International Guidelines

- **ICH Guideline E9 (1998), Section 4.4 Sample size adjustment:**

  > The steps taken to preserve blindness and consequences, if any, for the type I error […] should be explained.

- **CHMP Reflection Paper on Adaptive Designs (2007), Section 4.2.2 Sample size reassessment:**

  > Whenever possible, methods for blinded sample size reassessment […] that properly control the type I error should be used.

- **requirements:** **blinding** and **control of type I error rate**
Continuous Data: t-Test

- **data**: normally distributed with equal within-group variances $\sigma^2$

- **hypotheses**: $H_0 : \mu_T \leq \mu_C$  vs.  $H_1 : \mu_T > \mu_C$

- **approximate sample size**: $N = 4 \frac{\left( \Phi^{-1}(\alpha) + \Phi^{-1}(\beta) \right)^2}{\Delta^*^2} \sigma^2$

- **sample size adjustment**
  - re-estimating $\sigma^2$ by $S^2 = \frac{1}{n_1 - 2} \sum_{i,j}(X_{ij} - \bar{X}_i)^2$
  - partial **unblinding!**, requires Independent Data Monitoring Committee (IDMC)
Unblinded Sample Size Review: Actual Type I Error Rate (Kieser & Friede 2000)

- nominal level $\alpha = 0.025$
- unrestricted design
- actual type I error rate $\alpha_{act}$ depending on
  - size of the internal pilot study $n_1$
  - required, but unknown sample size $N$
Unblinded Sample Size Review: Control of Type I Error Rate

- search for adjusted level $\alpha_{adj}$ that fulfills
  \[
  \max_N \alpha_{act}(\alpha_{adj}, n_1, N) \leq \alpha
  \]

- table below gives $\alpha_{adj}$ for $\alpha = 0.025$ and unrestricted design

- slightly conservative, but adjusted level reasonably close to nominal level for say $n_1 \geq 50$

<table>
<thead>
<tr>
<th>$n_1$</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>50</th>
<th>100</th>
<th>180</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{adj}$</td>
<td>0.0178</td>
<td>0.0210</td>
<td>0.0223</td>
<td>0.0233</td>
<td>0.0241</td>
<td>0.0245</td>
</tr>
</tbody>
</table>
Alternative Approach for Type I Error Rate Control

- **cause of type I error rate inflation**: biased variance estimator (variance underestimated)

- **idea**: add correction term to variance (in test statistic)

- **result**: actual level close to nominal level

- **reference**: Miller (2005)
Example: St John’s wort in patients with depression

- **interim analysis** with 65 patients (31 St John’s wort, 34 placebo)
  - $\hat{\Delta} = 4.9$, $s_1 = 5.8 \rightarrow p_1 < 0.001$
  - early rejection of null hypothesis

- **sample size review**
  - *imagine*: same study as above, but with IPS
  - $s_1 = 5.8 \rightarrow n = 68$ (unrestricted), $n = 128$ (restricted)
Variance Estimators for Blinded Sample Size Reviews

- **idea**: total variance = within-group + between-group variance

- **one-sample variance** $S_{OS}^2 = \frac{1}{n_1 - 1} \sum_{i,j} (X_{ij} - \bar{X})^2$
  
  - in typical clinical trials, between-group variance relatively small compared to within-group variance

- **adjusted one-sample variance** (Zucker et al. 1999)
  
  - idea: $S_{adj}^2$ unbiased under alternative $\Delta = \Delta^*$
  
  - $S_{adj}^2 = S_{OS}^2 - \frac{1}{4n_1 - 1} \Delta^*^2$
Blinded Sample Size Review: Actual Type I Error Rate (Kieser & Friede 2003)

- **situations considered:** $N = 20, 40, \ldots, 200$, $n_1 = 20, 30, \ldots, 100$

- **conclusion:** no relevant excess of the nominal level observed!

### one-sample variance $S_{OS}^2$

<table>
<thead>
<tr>
<th>Situation</th>
<th>$\alpha_{act} - \alpha$</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 0.025$</td>
<td>-0.0001</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>$\alpha = 0.05$</td>
<td>-0.0001</td>
<td>0.0001</td>
<td></td>
</tr>
</tbody>
</table>

### adjusted variance $S_{adj}^2$

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$1 - \beta$</th>
<th>$\alpha_{act} - \alpha$</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.025</td>
<td>0.80</td>
<td>-0.0001</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>0.025</td>
<td>0.90</td>
<td>-0.0001</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.80</td>
<td>-0.0001</td>
<td>0.0002</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.90</td>
<td>-0.0001</td>
<td>0.0001</td>
<td></td>
</tr>
</tbody>
</table>
Power of Blinded Sample Size Adjustment Procedures

\[ 1 - \beta = 0.80, \ \alpha = 0.025 \]

<table>
<thead>
<tr>
<th>\Delta/\sigma</th>
<th>N</th>
<th>( n_1 )</th>
<th>OS variance</th>
<th>Adjusted variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Power ( E(n) )</td>
<td>Power ( E(n) )</td>
</tr>
<tr>
<td>0.7</td>
<td>64</td>
<td>40</td>
<td>0.800</td>
<td>72.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>60</td>
<td>0.816</td>
<td>73.1</td>
</tr>
<tr>
<td>0.5</td>
<td>126</td>
<td>40</td>
<td>0.792</td>
<td>134.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>60</td>
<td>0.797</td>
<td>134.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80</td>
<td>0.800</td>
<td>134.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>120</td>
<td>0.811</td>
<td>135.8</td>
</tr>
<tr>
<td>0.3</td>
<td>348</td>
<td>40</td>
<td>0.787</td>
<td>356.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>60</td>
<td>0.792</td>
<td>356.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>0.796</td>
<td>355.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>150</td>
<td>0.799</td>
<td>355.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
<td>0.800</td>
<td>355.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>350</td>
<td>0.812</td>
<td>363.9</td>
</tr>
</tbody>
</table>
Example: St John’s wort in depression

- **imagine**: study as above, but with blinded sample size review

- **initial sample size estimate**: 128 patients ($\sigma_0 = 8$)

- **blinded sample size review**:
  - with 65 patients (31 St John’s wort, 34 placebo)
  - $s_{OS} = 6.3 \rightarrow n = 80$
  - $s_{adj} = 6.0 \rightarrow n = 74$
Discussion: Unblinded Review vs. Blinded Review

- unblinded estimate of within-group variance always smaller than estimate of total variance

- blinded review carried out by trial statistician and clinical trial leader, no IDMC necessary

- unblinded reviews potentially reveal information on effect size

- regulators seem to favour blinded reviews
Blinded Sample Size Reviews in Non-inferiority Trials

- **treatments similar**: blinded review even more attractive
- **Type I error rate**: small inflation observed!
- difference vs. ratio of two means
- **equivalence trials**: two one-sided tests
- **ref.**: Friede & Kieser (2003)

Std. non-inf. margin $D/\sigma = -0.3$
Blinded Sample Size Reviews with Binary Data

• nuisance parameter: overall response rate
  – can be estimated from interim data without unblinding

• effect measures
  – risk difference (RD), odds ratio (OR), relative risk (RR)
  – sample size adjustment sensitive to choice of effect measure (Gould 1995)

• Friede & Kieser (2004) investigate blinded review with RD
Blinded Reviews in Non-inferiority Trials with Binary Data (Friede et al 2007)

- motivated by anti-hypertensive study example

- nuisance parameter: overall response rate

- treatments similar: blinded review attractive

- non-inferiority margin: here defined in terms of RD
Example: Anti-hypertensive Trial

- **non-inferiority margin**
  - RD: 10 percentage points
  - in terms of OR: see plot

- fairly constant OR for mid-range risks (say 40-70 %)

- RD not suitable for risks near 0 or 1

PhARMA KOL Lecture Series 2008
Type I Error Rate for Blinded Sample Size Reestimation

Blackwelder

F&M

\[ \alpha = 0.025, \quad 1 - \beta = 0.80, \quad \theta = 1/3, 1/2, 1, \quad \delta = 0.1, \quad \delta_1^a = 0, \quad \pi = 0.30, 0.31, \ldots, 0.7 \]
Power of F&M Test: Misspecification of Overall Response

\[ \theta = 1 \]

\[ \theta = 0.33 \]

Fixed (dashed) and reest. (solid) for \( \pi^a = 0.5 \) (black), 0.7 (grey) with \( \delta_1 = \delta^a_1 = 0 \)

PhARMA KOL Lecture Series 2008
Example: Anti-hypertensive Trial

- $\alpha = 0.025$, $1 - \beta = 0.80$, $\delta = 0.10$, $\delta_1 = \delta_1^a$, $\theta = 1$

- assumed overall response 70%, actual overall response rate 42%

- **fixed design**
  - total sample size 660 $\Rightarrow$ power 74.5% (B), 74.9% (F&M)

- **blinded sample size reestimation**
  - exp. sample size 759 (B), 754 (F&M); power 79.7% (B, F&M)
Further Issues in Sample Size Reestimation

• **early readouts for sample size recalculation**
  - **problem:** at IA only small proportion of patients completed follow-up
  - **idea:** use correlation between early and final readout and gain precision in estimation
  - **references:** Marschner & Becker (2001), Wüst & Kieser (2003, 2005)

• **use of confidence bounds** rather than point estimates for sample size reestimation

• **GS procedure** (Gould & Shih 1992) inappropriate (Friede & Kieser 2002; Letter to the Editor by Gould & Shih and reply; Waksman 2007)
Conclusions

• reasons other than sample size for interim look?
  – if yes, choose design with interim analysis
  – otherwise consider blinded sample size review

• **blinded sample size review**
  – fulfills requirements according to ICH E9
  – good power and sample size properties
  – ... and it’s easy to apply!
Further Reading


References


