# 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)
  - Based on Armitage et al. (1969), McPherson (1974), Pocock (1977), O'Brien/Fleming (1979), Lan/DeMets (1983), Jennison/Turnbull (1999), …
- 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

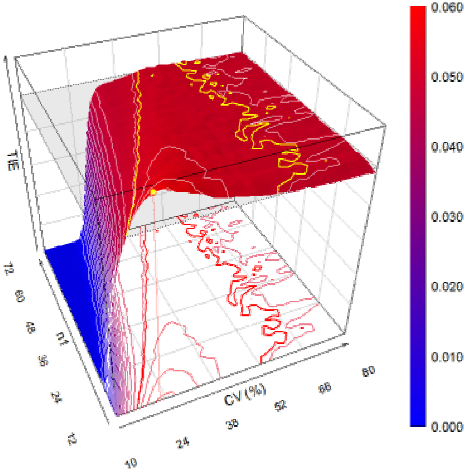
- Fixed total sample size (*N*) and – in BE – one interim analysis
  - First proposal by Gould (1995) in the field of BE did not get regulatory acceptance in Europe
  - Asymmetric split of $\alpha$ is possible, *i.e.*,
    - a small $\alpha$ in the interim (*i.e.*, stopping for futility) and
    - a large one in the final analysis (*i.e.*, only small sample size penality)
    - 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

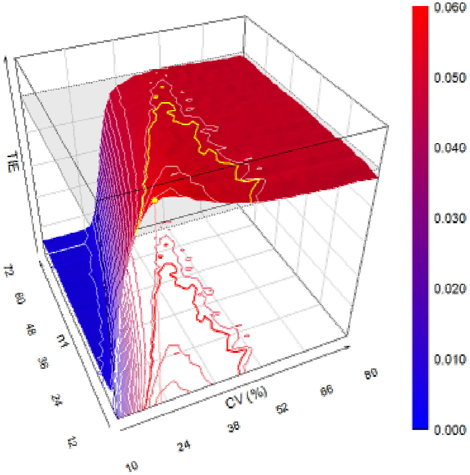


Maximum 0.05849

$\alpha_2$ 0.0413 needed to control the TIE

### O'Brien/Fleming

$\alpha_1$ 0.005, $\alpha_2$ 0.048

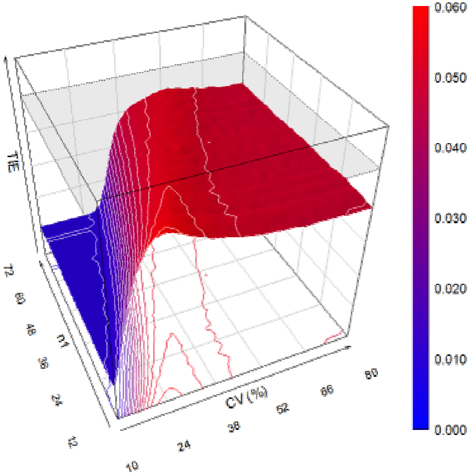


Maximum 0.05700

$\alpha_2$ 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 *CV*s 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, *CV*s, *GMR*s, 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

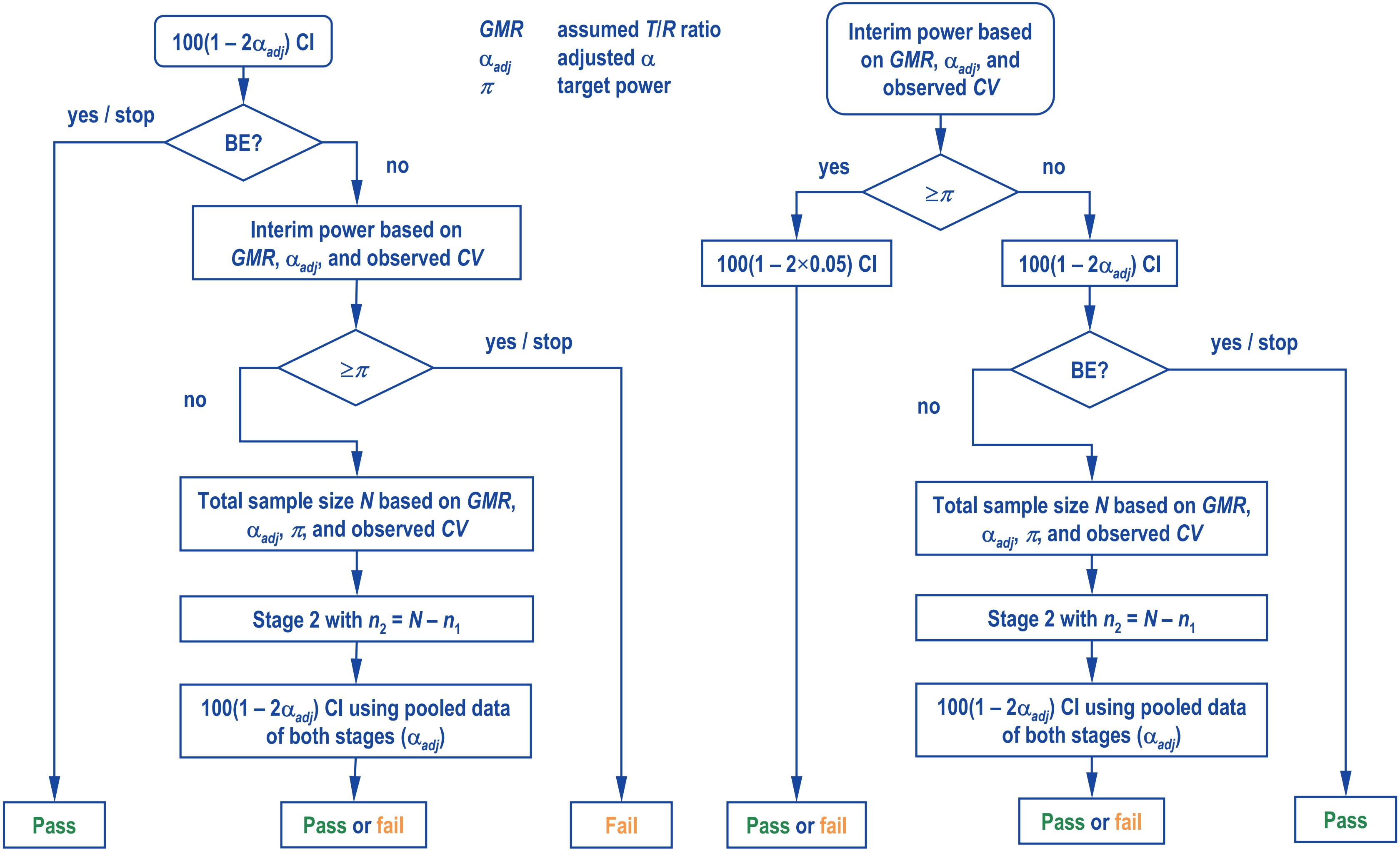
# (Adaptive) Sequential Two-Stage Designs

## 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

| | |
|---|---|
| *GMR* | assumed *T*/*R* ratio |
| $\alpha_{adj}$ | adjusted $\alpha$ |
| $\pi$ | target power |



100(1 – 2α_adj) CI
BE?
yes / stop
no
Interim power based on GMR, α_adj, and observed CV
≥π
yes / stop
no
Total sample size N based on GMR, α_adj, π, and observed CV
Stage 2 with n2 = N – n1
100(1 – 2α_adj) CI using pooled data of both stages (α_adj)
Pass
Pass or fail
Fail
Interim power based on GMR, α_adj, and observed CV
≥π
yes
no
100(1 – 2×0.05) CI
100(1 – 2α_adj) CI
BE?
yes / stop
no
Total sample size N based on GMR, α_adj, π, and observed CV
Stage 2 with n2 = N – n1
100(1 – 2α_adj) CI using pooled data of both stages (α_adj)
Pass or 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.

| Reference | Type | Method | *GMR* | Target power | $CV_w$ | $\alpha_{adj}$ | $TIE_{max}$ |
|---|---|---|---|---|---|---|---|
| Potvin *et al.* (2008) | 1 | B | 0.95 | 80% | 10 – 100% | 0.0294 | 0.0485 |
| | 2 | C | | | | | 0.0510 |
| Montague *et al.* (2012) | 2 | D | 0.90 | | | 0.0280 | 0.0518 |
| Fuglsang (2013) | 1 | B | 0.95 | 90% | 10 – 80% | 0.0284 | 0.0501 |
| | 2 | C/D | | | | 0.0274 | 0.0503 |
| | 2 | C/D | 0.90 | | | 0.0269 | 0.0501 |

- Xu *et al.* (2015). *GMR* 0.95, target power 80%, futility for the (1–2$\alpha_1$) CI.

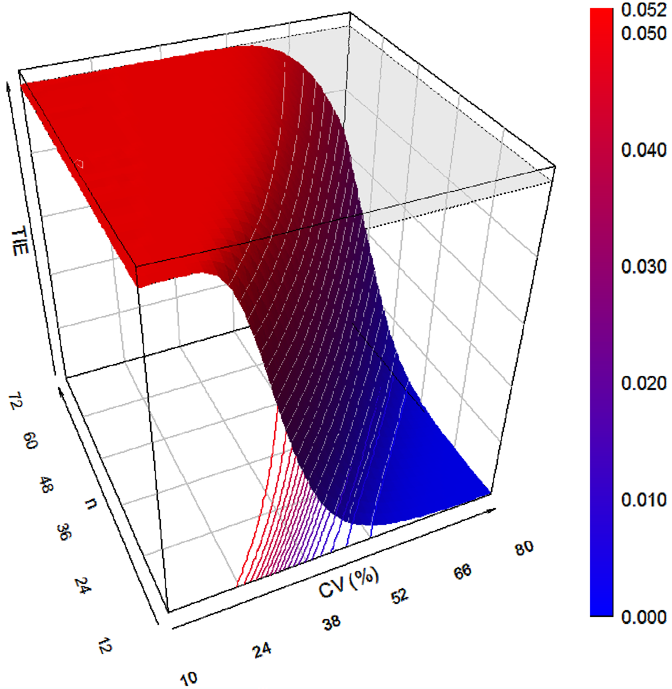
| Type | Method | $CV_w$ | Futility region | $\alpha_1$ | $\alpha_2$ | $TIE_{max}$ |
|---|---|---|---|---|---|---|
| 1 | E | 10 – 30% | 0.9374 – 1.0667 | 0.0249 | 0.0363 | 0.050 |
| 2 | F | | 0.9492 – 1.0535 | 0.0248 | 0.0364 | 0.050 |
| 1 | E | 30 – 55% | 0.9305 – 1.0747 | 0.0254 | 0.0357 | 0.050 |
| 2 | F | | 0.9350 – 1.0695 | 0.0259 | 0.0349 | 0.050 |

# Excursion 2

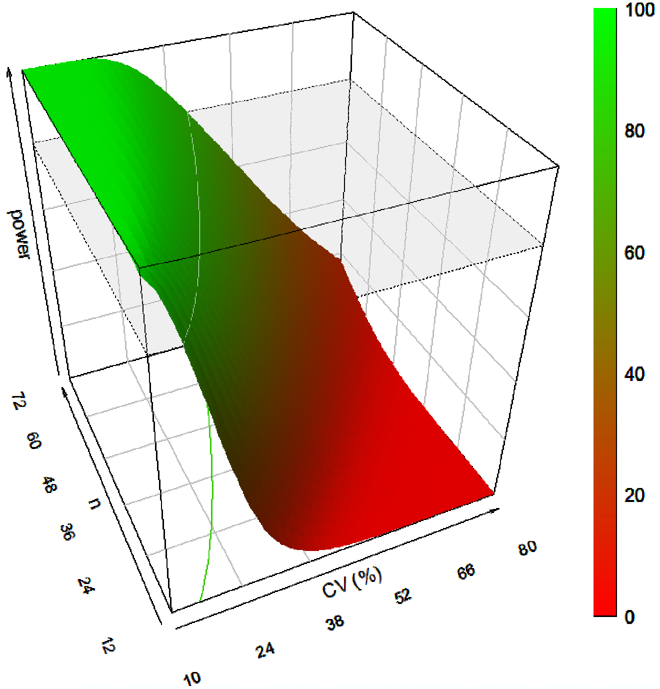
## Type I Error and power

- Fixed sample 2×2×2 design (α 0.05). *GMR* 0.95, *CV* 10 – 80%, *n* 12 –72

TIE



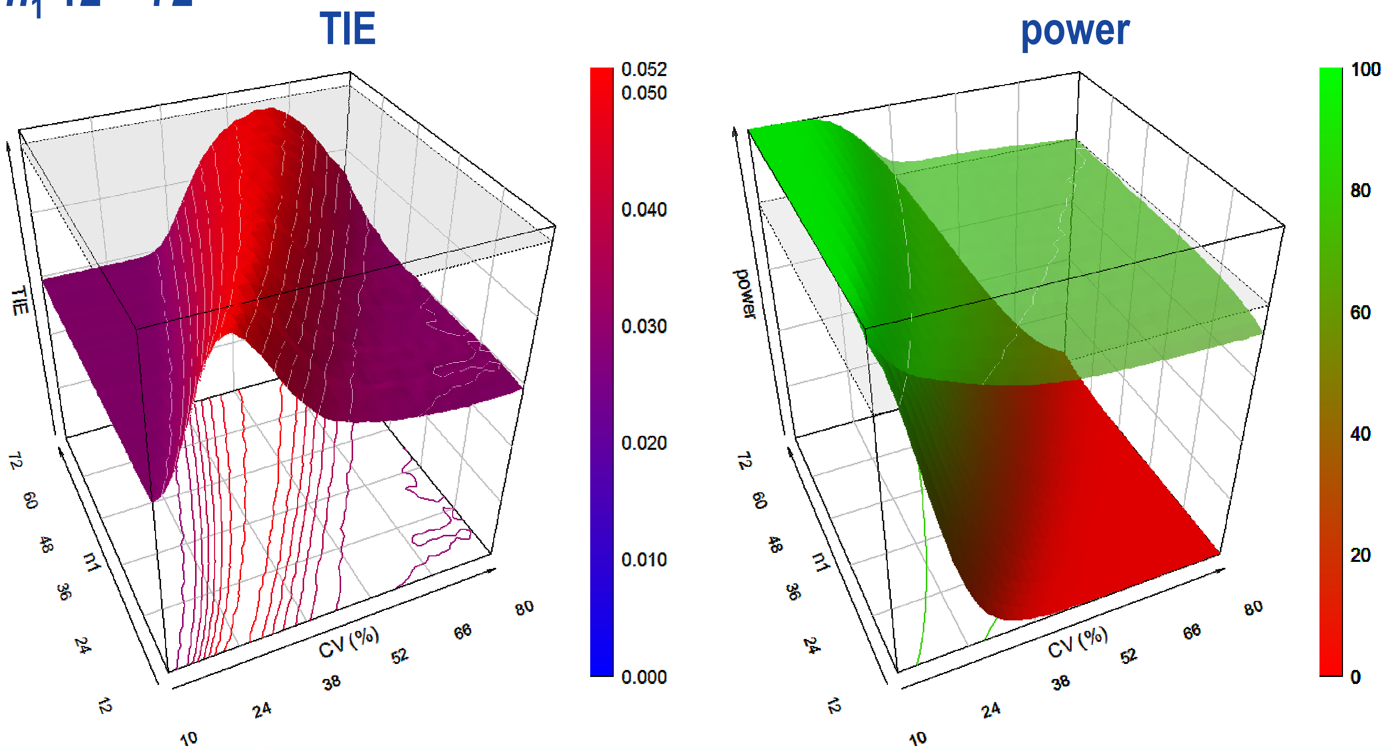
power



# 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):

    ```
    stage + sequence + sequence(stage) + subject(sequence × stage) +
    period(stage) + 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; 2nd / 3rd GBHI conferences, Rockville 2016 and Amsterdam 2018)
- Russia (2013), Eurasian Economic Union (2016)
  - 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. Fuglsang 2014

    […] 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. Jones/Kenward 2014
  - 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)

# (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).

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

# (Adaptive) Sequential Two-Stage Designs

## 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

# Rumors & Chinese Whispers (Part 2)

## 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!

# Rumors & Chinese Whispers (Part 2)

## Simulations *vs.* ‘analytical proof’

- Summer Symposium *‘To New Shores in Drug Development Implementing Statistical Innovation’*, Vienna, 27 June 2016
  - Most proofs start with …

    *Let us assume parallel groups of equal sizes*
    *and normal distributed data with $\mu$ = 0 and $\sigma$ = 1*

    … followed by some fancy formulas.

    Do these cases *ever* occur in *reality*? Peter Bauer

# Group-Sequential and Two-Stage Designs

## Thank You!

## *Open Questions?*

**Helmut Schütz**

**BEBAC**

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