

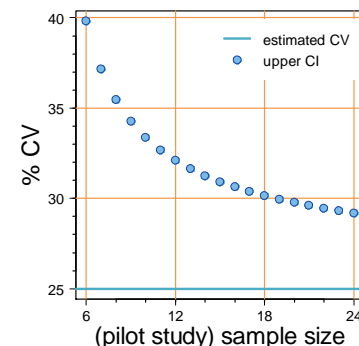
# Group-Sequential and Two-Stage Designs

Helmut Schütz

# Dealing with Uncertainty

## Nothing is 'carved in stone'.

- Never assume perfectly matching products.
  - Generally a  $\Delta$  of not better than 5% should be assumed (0.950 – 1.053).
  - For HVD(P)s do not assume a  $\Delta$  of <10% (0.900 – 1.111).
- Do not use the CV but one of its confidence limits.
  - Suggested  $\alpha$  0.2 (here: the producer's risk).
  - For ABE the upper CL.
  - For reference-scaling (generally) the lower CL.
- Better alternatives.
  - Group-Sequential Designs  
Fixed total sample size, interim analysis for early stopping.
  - (Adaptive) Sequential Two-Stage Designs  
Fixed stage 1 sample size, re-estimation of the total sample size in the interim analysis.



# Remedies?

## Group-Sequential Designs

- Fixed total sample size ( $N$ ) and – in BE – one interim analysis.
  - Requires two assumptions. One ‘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 interim at  $N/2$ .
  - That’s not what we have in BE: equivalence (generally in a crossover), lognormal data with unknown variance. Furthermore, due to drop-outs, the interim might not be exactly at  $N/2$  (might inflate the Type I Error).
  - Asymmetric split of  $\alpha$  is possible, *i.e.*, a small  $\alpha$  in the interim and a large one in the final analysis.  
 Examples: Haybittle/Peto ( $\alpha_1$  0.001,  $\alpha_2$  0.049), O’Brien/Fleming ( $\alpha_1$  0.005,  $\alpha_2$  0.048), Zheng et al. ( $\alpha_1$  0.01,  $\alpha_2$  0.04).  
 May require  $\alpha$ -spending functions (Lan/DeMets, Jennison/Turnbull) in order to control the Type I Error.

# Remedies?

## (Adaptive) Sequential Two-Stage Designs

- Fixed stage 1 sample size ( $n_1$ ), sample size re-estimation in the interim.
  - Generally a fixed *GMR* is assumed.
  - Fully adaptive methods (*i.e.*, taking also the PE of stage 1 into account) are problematic. May deteriorate power and require a futility criterion. Simulations mandatory.
  - 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.

# Group-Sequential Designs

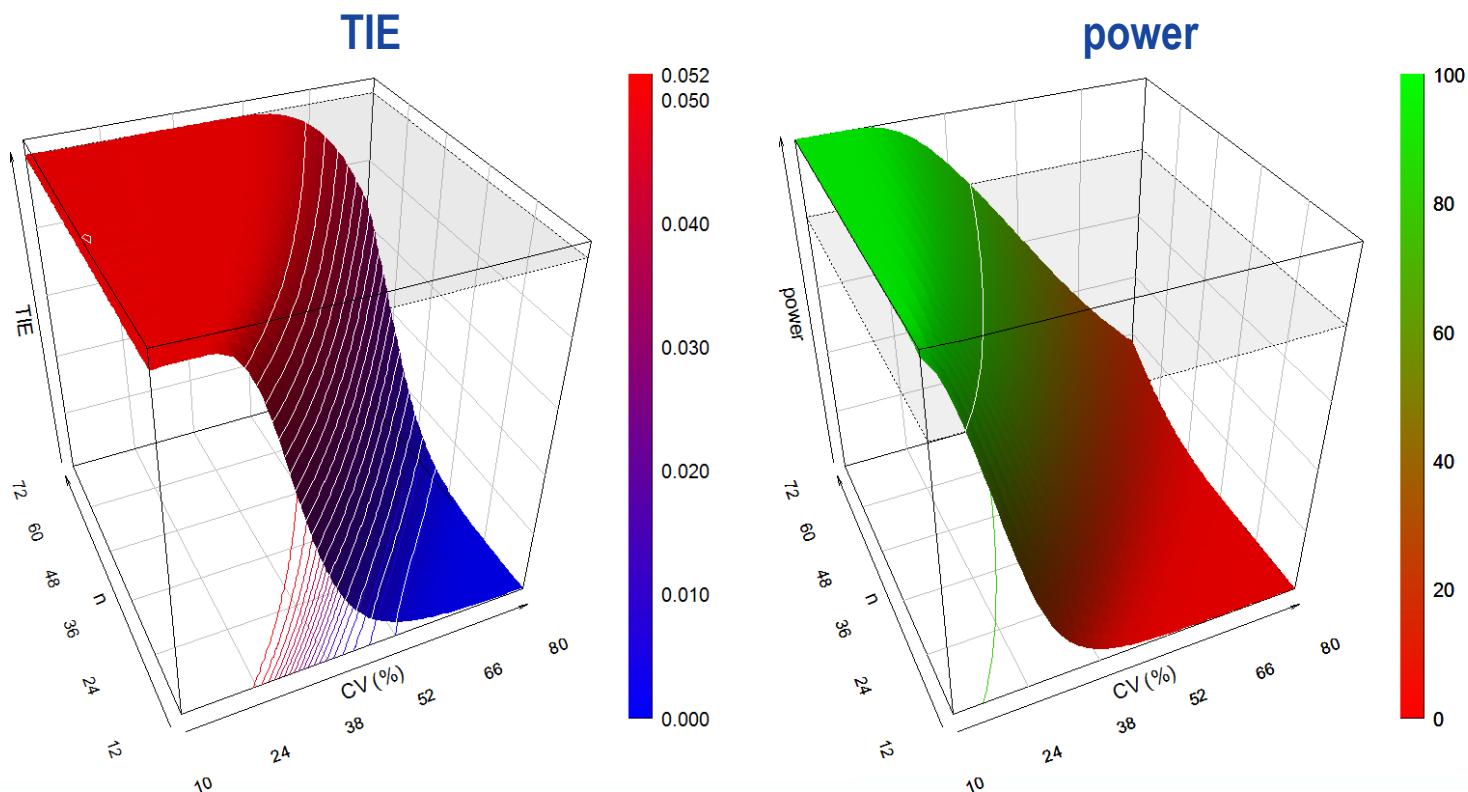
## Long and accepted tradition in clinical research (phase III)

- Based on Armitage et al. (1969), McPherson (1974), Pocock (1977), O'Brien/Fleming (1979), Lan/DeMets (1983), Jennison/Turnbull (1999), ...
  - Developed for superiority testing, parallel groups, normal distributed data with known variance, and interim at  $N/2$ .
  - 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

## Type I Error and power

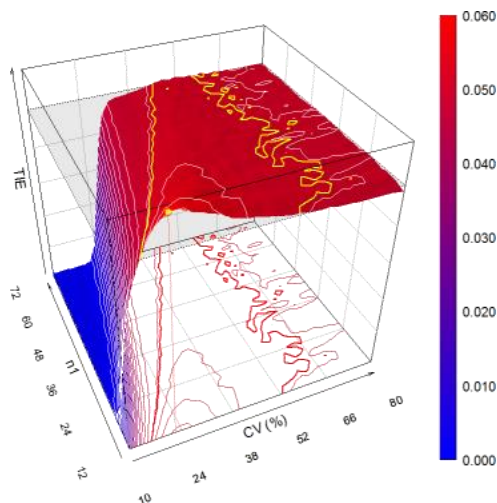
- Fixed sample  $2 \times 2 \times 2$  design ( $\alpha$  0.05). *GMR* 0.95, *CV* 10 – 80%, *n* 12 – 72



# Group-Sequential Designs

## Type I Error

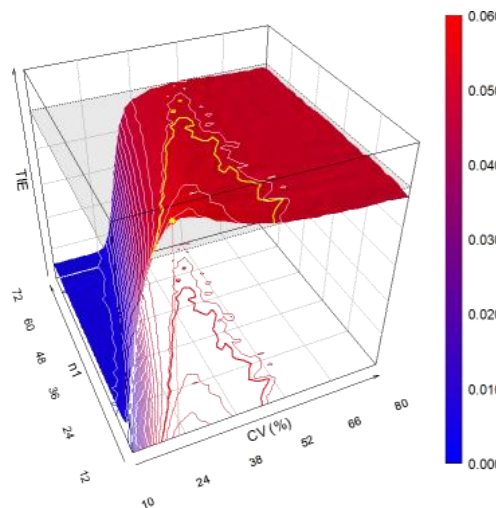
Haybittle/Peto  
 $\alpha_1$  0.001,  $\alpha_2$  0.049



Maximum 0.05849

$\alpha_2$  0.0413 needed  
to control the TIE

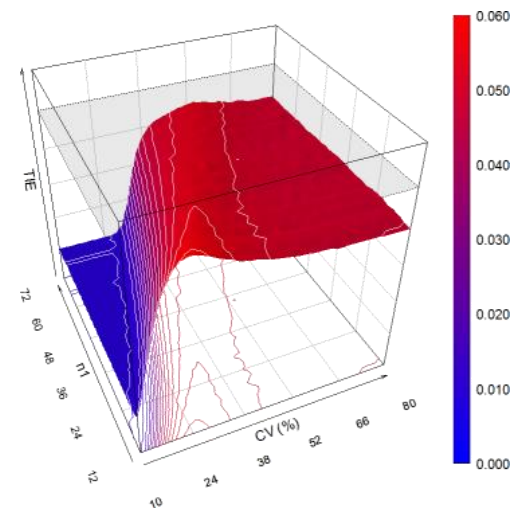
O'Brien/Fleming  
 $\alpha_1$  0.005,  $\alpha_2$  0.048



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 TIE  $\leq 0.0494$ .
  - 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 TIE is possible!  
 $\alpha$ -spending functions can control the TIE (but are *not* mentioned in the guidance).



# (Adaptive) Sequential Two-Stage Designs

Fixed stage 1 sample size ( $n_1$ ), sample size re-estimation in the interim.

- 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.
- Contrary to common beliefs no analytical proof of controlling the TIE exist.

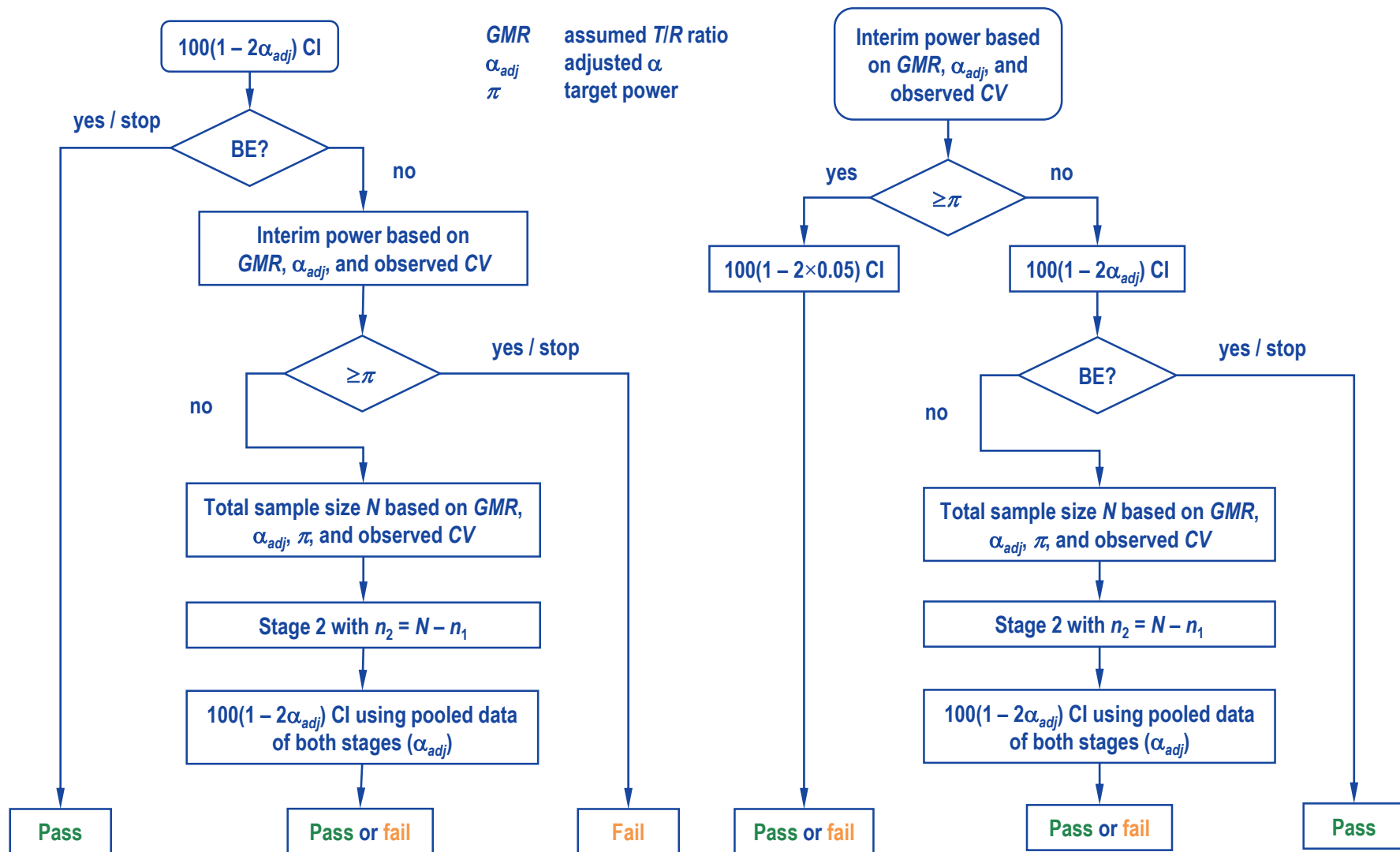
It is the responsibility of the sponsor to demonstrate (e.g., by simulations) that the consumer risk is preserved.

- Fully adaptive methods (*i.e.*, taking also the PE of stage 1 into account) are problematic. May substantially deteriorate power and require a futility criterion. Simulations mandatory.

# Type 1 and Type 2

*GMR*  
 $\alpha_{adj}$   
 $\pi$

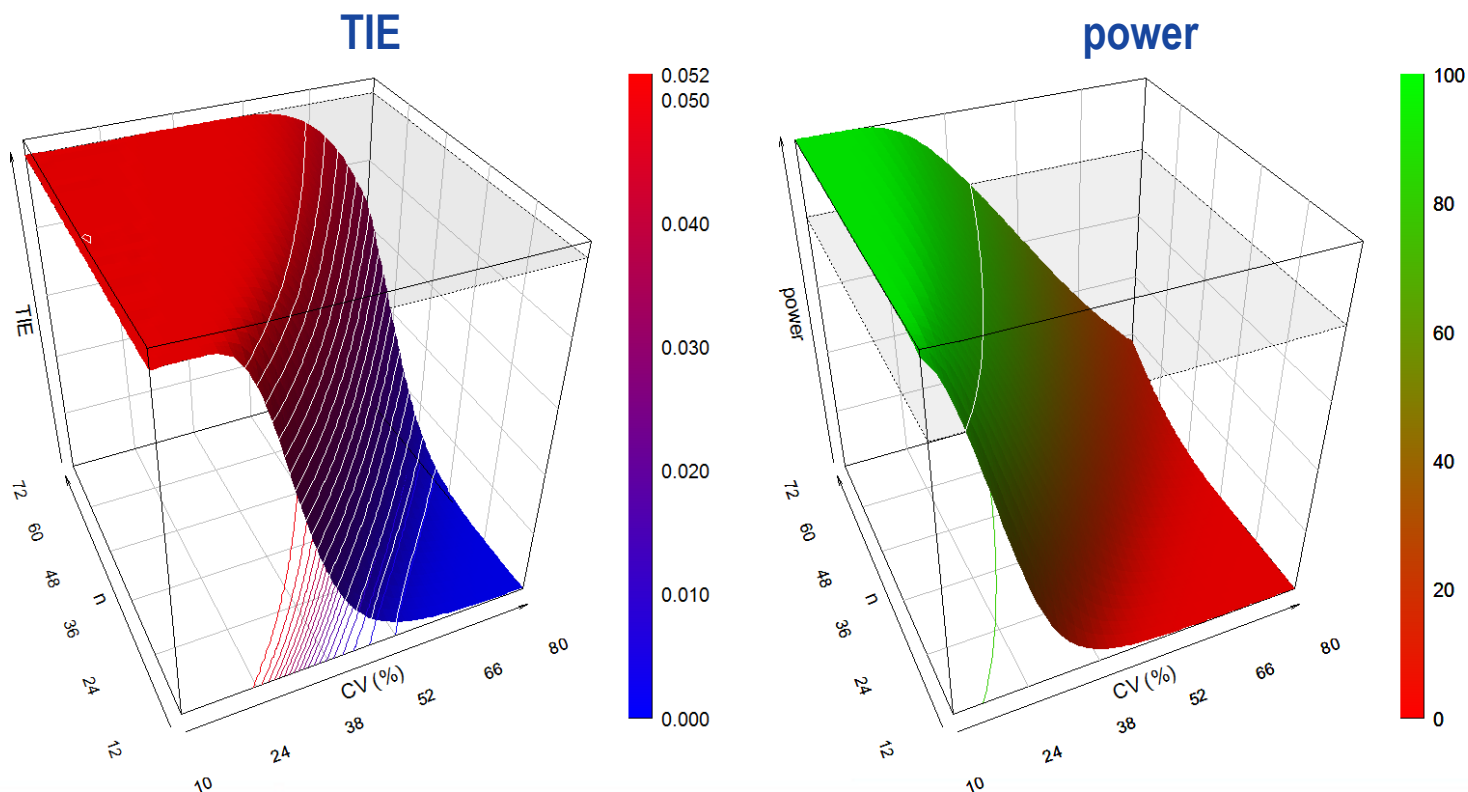
assumed *T/R* ratio  
adjusted  $\alpha$   
target power



# Excursion

## Type I Error and power

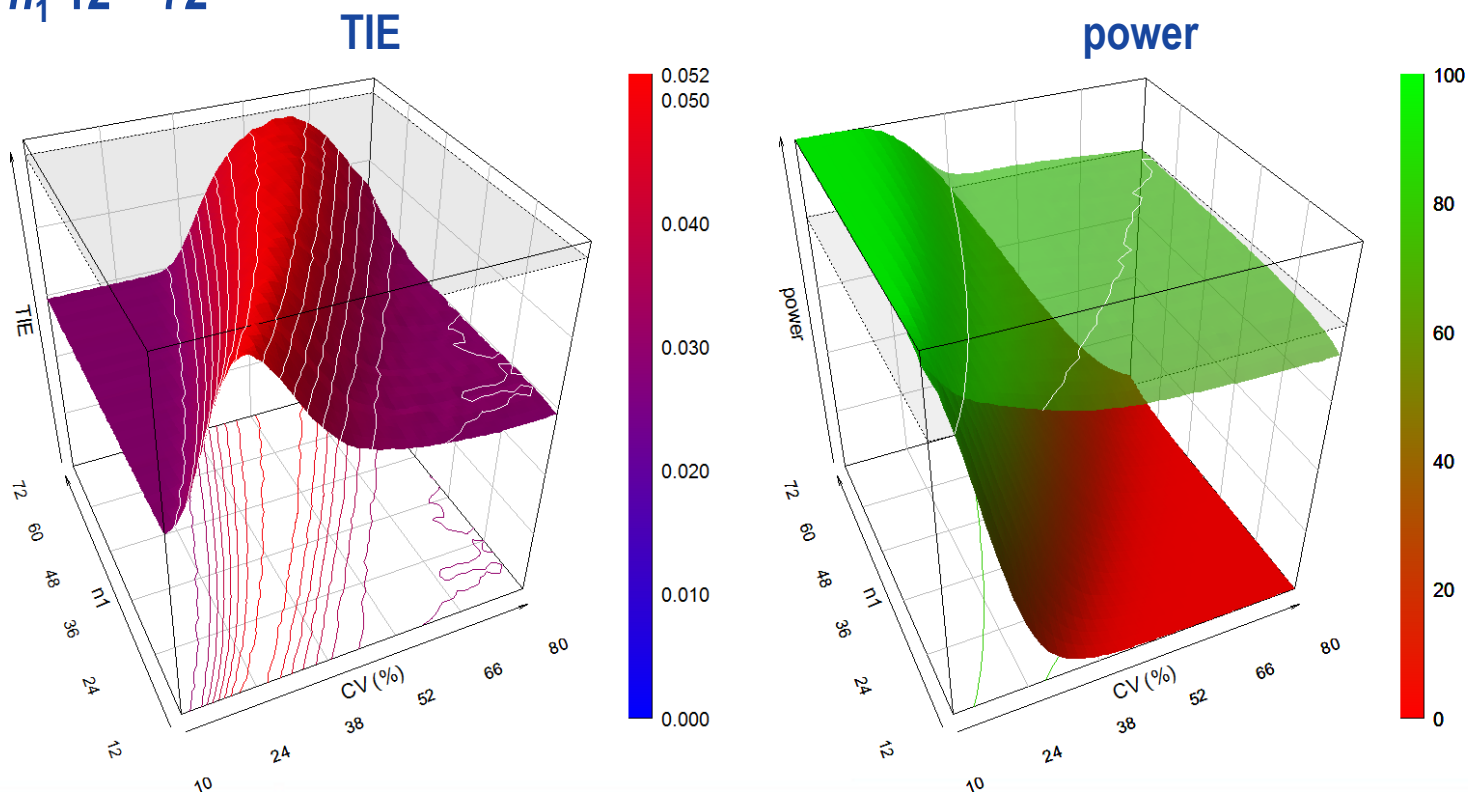
- Fixed sample  $2 \times 2 \times 2$  design ( $\alpha 0.05$ ). *GMR* 0.95, *CV* 10 – 80%, *n* 12 – 72



# Excursion

## 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

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 <sub>w</sub>	$\alpha_{adj}$	TIE <sub>max</sub>
Potvin <i>et al.</i> (2008)	1	B	0.95	80%	10 – 100%	0.0294	0.0485
	2	C					0.0510
Montague <i>et al.</i> (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 <sub>w</sub>	Futility region	$\alpha_1$	$\alpha_2$	TIE <sub>max</sub>
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

# (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.’
  - 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 2017...

# (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 x stage) +
period(stage) + formulation
```
    - At least two subjects in the second stage.
    - 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?
      - 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 recommended (Davit et al. 2013; 2<sup>nd</sup> GBHI conference, Rockville 2016).
- Russia (2013), Eurasian Economic Union (2016)
  - Acceptable; Potvin Method B preferred?

# (Adaptive) Sequential Two-Stage Designs

## Futility Rules

- 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 and overrun studies

- Dropouts 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.
- Including more than the re-estimated subjects in the second stage
  - Common practice in fixed sample designs ‘in order to compensate for loss in power based on the expected dropout-rate’.
  - If less dropouts occur in the second stage, the study is ‘overrun’. The chance to show BE increases and therefore, the TIE!
  - Methods exists in the literature (though for parallel designs, superiority testing only) to adjust  $\alpha$  accordingly. Nothing published for equivalence yet.
  - Don’t go there.

# (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).

| $n_1$ | $E[M]$ | Studies stopped<br>in stage 1 (%) | Studies failed<br>in stage 1 (%) | Power in<br>stage 1 (%) | Studies in<br>stage 2 (%) | Final<br>power (%) | Increase of<br>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 some concerns about the validity of methods based on simulations.

# (Adaptive) Sequential Two-Stage Designs

## Outlook

- Selecting a candidate formulation from a higher-order crossover; continue with  $2 \times 2 \times 2$  in the second stage.
- Continue a  $2 \times 2 \times 2$  TSD in a replicate design for reference-scaling.
- Fully adaptive methods (taking the PE of stage 1 into account – without jeopardizing power).
- Exact methods (not relying on simulations).

# Case Study 1

## Potvin 'Method C' (2010 – 2011)

- Study stopped in stage 1
  - $AUC$ : power >80%; passed BE with 90% CI.
  - $C_{max}$ : power <80%; passed BE with 94.12% CI.
- **NL: Adapting the confidence intervals based upon power is not acceptable and also not in accordance with the EMA guideline.\* Confidence intervals should be selected *a priori*, without evaluation of the power. Therefore, the applicant should submit the 94.12% confidence intervals for  $AUC$ .**
  - \* What about: '... choice of how much alpha to spend at the interim analysis is at the company's discretion.'?
    - Failed to show BE of  $AUC$  with 94.12% CI.
    - Study repeated in India in a very (!) large fixed sample design.
    - Failed on  $C_{max}$ . Project cancelled.



# Case Study 2

## Potvin 'Method C' (2011 – 2012)

- Study passed already in stage 1
  - CV in the interim 30.65%,  $n_1$  49.
  - 90% CI since power was 87.3%.
- UK, IE: **Unadjusted  $\alpha$  in stage 1 not acceptable.**
  - Study passed with 94.12% CI as well (*post hoc* switch to 'Method B').
- **AT: The Applicant should demonstrate that the type I error inflation, which can be expected from the chosen approach, did not impact on the decision of bioequivalence.\***
  - \* Unofficial information: Potvin's table contains only a cell for CV 30% and  $n_1$  48...
    - One million studies simulated based on the study's CV and  $n_1$ .
    - Empiric Type I Error 0.0494 (95% CI: 0.0490 – 0.0498).

# Case Study 3

## Potvin 'Method C' (2012 – 2013)

- Protocol synopsis with statistical details submitted to the Spanish Agency (2012).
  - Unofficial feedback (after consultation of AEMPS with the BSWP):
    - Potvin's method is not valid in Europe.
- Question to the Spanish Agency (2013):
 

[...] we'd like to ask about the current status of TSD BE study, [...] if the BE protocol with Potvin's Method C is acceptable now [...].

  - Answer:
    - Potvin's methods are not acceptable in EMA.

# 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.
  - 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.
  - Remarks:  
One could confirm  $\sim 0.0304$  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.**
    - 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.

# Rumors & Chinese Whispers (Part 1)

## TSDs based on simulations

- Scientific Advice in SE (2016).
  - Simulations based on Fuglsang's 'Type 1' TSD for Parallel Groups (2014).
  - Large  $n_1$  (up to 125/group), homo- and heterogenous variances, potentially unequal group sizes due to drop-outs.
  - With  $\alpha_{adj}$  0.0274 the maximum Type I Error was 0.04992.
  - Response:
    - According to the guideline, application of a TSD was accepted provided that the patient's risk is maintained at or below 5%.
    - Confirmed that the statement about Potvin's methods is not public. These types of TSDs are not proven in a strict sense.
    - However, it was acknowledged that the simulations covered a sufficient range of possible outcomes (unequal variances and drop-out rates).
    - [...] the empiric type I error rate should be evaluated with the real data (i.e., the actual group sizes and variances of the study).

# 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.

# Package AdaptiveBE

## Function check.TSD()

- Required:
  - Interim data (*CV* or *MSE*,  $n_1$ , PE or CI), data of the final analysis (*CV* or *MSE*,  $N$ , PE or CI), adjusted alpha(s), the type of the TSD (optionally futility rules).
  - Alternatively (*i.e.*, if not given in the report) the CIs can be used to calculate the CVs and/or the PEs.
- Algorithm:
  - Based on the interim data and the study's framework simulate one million studies in order to obtain the empiric Type I Error.
    - If the TIE  $\leq 0.05$ , stop. Can accept the applicant's results.
    - If not, optimize  $\alpha_{adj}$  with a target TIE of 0.05. Recalculate the study (interim – and optionally – final) and compare conclusions with the reported ones.
      - » If conclusions agree, accept the study (increase of the TIE not *relevant*).
      - » If not (reported passes and adjusted fails), calculate the increase of relative risk. Whether the study is accepted or not lies in the hands of the assessor.

# Package AdaptiveBE

Available at <https://github.com/Helmut01/AdaptiveBE>

- Example 2 of Potvin's 'Method C'
  - The maximum TIE in Table I of in the reference is 0.0510 for CV 20%,  $n_1$  12.
  - I used the reported *MSEs* and sample sizes. The CV in the interim was with 18.21% close to the location of the maximum TIE.
  - The power-calculation was done by the shifted *t*-distribution like in the reference.
  - R-code

```
library(AdaptiveBE)
check.TSD(Var1=c(0.032634, "MSE"), PE1=c(0.083960, "log"), n1=12,
          Var=c(0.045896, "MSE"), PE=c(0.014439, "log"), N=20,
          alpha0=0.05, alpha1=0.0294, alpha2=0.0294,
          type=2, GMR=0.95, pmethod="shifted")
```



# Package AdaptiveBE

## Function check.TSD()

— Part of the output

TIE for specified  $\alpha$ : 0.05062 ( $>0.05$ )

Applied adjustment is not justified.

Final analysis of pooled data (specified  $\alpha_2$  0.0294)

---

---

94.12% CI: 88.45–116.38% (BE concluded)

Adjusted  $\alpha$  1, 2 : 0.050 | 0.02858, 0.02858

Adjusted CIs : 90.00% | 94.28%, 94.28%

TIE for adjusted  $\alpha$  : 0.04992 (n.s.  $>0.05$ )

Final analysis of pooled data (adjusted  $\alpha_2$  0.02858)

---

---

94.28% CI: 88.36–116.39% (BE concluded)

Since conclusions of both analyses agree,  
can accept the original analysis.

# Package AdaptiveBE

- It was difficult to fabricate an example where the original evaluation would pass and the optimized fail, *i.e.*, a borderline case where the CI was ‘too close’ to the acceptance limits.
  - The maximum TIE reported in any of the publications is 0.0518 (Montague’s ‘Method D’, CV 20%,  $n_1$  12).
  - I used the interim CV and  $n_1$ , a  $PE_1$  of 0.92, and in the final analysis a higher CV (22.3%), a worse PE (0.88), and one drop-out in the second stage ( $N$  45).
  - The power-calculation was done by the shifted  $t$ -distribution like in the reference.
  - R-code
 

```
library(AdaptiveBE)
check.TSD(Var1=c(0.200, "CV"), PE1=c(0.92, "ratio"), n1=12,
          Var=c(0.233, "CV"), PE=c(0.88, "ratio"), N=45,
          alpha0=0.05, alpha1=0.028, alpha2=0.028,
          type=2, GMR=0.90, pmethod="shifted")
```

# Package AdaptiveBE

## Function check.TSD()

### — Part of the output

TIE for specified  $\alpha$ : 0.05153 ( $>0.05$ )

Applied adjustment is not justified.

Final analysis of pooled data (specified  $\alpha_2$  0.028)

---

---

94.40% CI: 80.00–96.80% (BE concluded)

Adjusted  $\alpha$  1, 2 : 0.050 | 0.02709, 0.02709

Adjusted CIs : 90.00% | 94.58%, 94.58%

TIE for adjusted  $\alpha$  : 0.04998 (n.s.  $>0.05$ )

Final analysis of pooled data (adjusted  $\alpha_2$  0.02709)

---

---

94.61% CI: 79.94–96.87% (failed to demonstrate BE)

Accepting the reported analysis could increase the relative consumer risk by ~3.1%.

# 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 implemented in the development release of R-package Power2Stage available at <https://github.com/Detlew/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).  
Correct. But only two posters about BE so far (not published in a peer-reviewed journal).
- In the inverse normal approach one obtains two  $p$ -values (compatible with the GLs requiring a confidence interval?)
- Both in the inverse normal approach and with repeated CIs the final  $\alpha$  is adapted based on the study's data (compatible with the GLs 'pre-specified  $\alpha$ '?)
- Either there is a proof (but *not* for the conditions in BE) or it is not published yet.

# 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](mailto:helmut.schuetz@bebac.at)