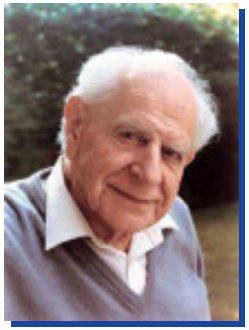
Two-Stage Sequential Designs
Industry Perspective

Helmut Schütz

To bear in Remembrance...

Whenever a theory appears to you as the only possible one, take this as a sign that you have neither understood the theory nor the problem which it was intended to solve.



Karl R. Popper

Even though it's *applied* science we're dealin' with, it still is – *science*!

Leslie Z. Benet

Assumptions

All models rely on assumptions.

- Bioequivalence as a surrogate for therapeutic equivalance.
  - Studies in healthy volunteers in order to minimize variability (*i.e.*, lower sample sizes than in patients).
  - Current emphasis on *in vivo* release ('human dissolution apparatus').
- Concentrations in the sample matrix reflect concentrations at the target receptor site.
  - In the strict sense only valid in steady state.
  - *In vivo* similarity in healthy volunteers can be extrapolated to the patient population(s).
- $f = \mu_T / \mu_R$ assumes that
  - $D_T = D_R$ *and*
  - inter-occasion clearances are constant.

Assumptions

All models rely on assumptions.

- Log-transformation allows for additive effects required in ANOVA.
- No carry-over effect in the model of crossover studies.
  - Cannot be statistically adjusted.
  - Has to be avoided *by design* (suitable washout).
  - Shown to be a statistical artifact in meta-studies.
  - Exception: Endogenous compounds (biosimilars!)
- Between- and within-subject errors are independently and normally distributed about unity with variances σ^2_s and σ^2_e.
  - If the reference formulation shows higher variability than the test, the 'good' test will be penalized for the 'bad' reference.
- All observations made on different subjects are independent.
  - No monocygotic twins or triplets in the study!

Sample Size

Only power is accessible.

- The required sample size depends on
  - the acceptance range (AR) for bioequivalence;
  - the error variance (s^2) associated with the PK metrics as estimated from
    - published data,
    - a pilot study, or
    - previous studies;
  - the fixed significance level (α);
  - the expected deviation (Δ) from the reference product and;
  - the desired power ($1 - \beta$).
- Three values are *known and fixed* (AR, α, $1 - \beta$),
  one is an *assumption* (Δ), and
  one an *estimate* (s^2).
  Hence, the correct term is 'sample size *estimation*'.

Sample Size

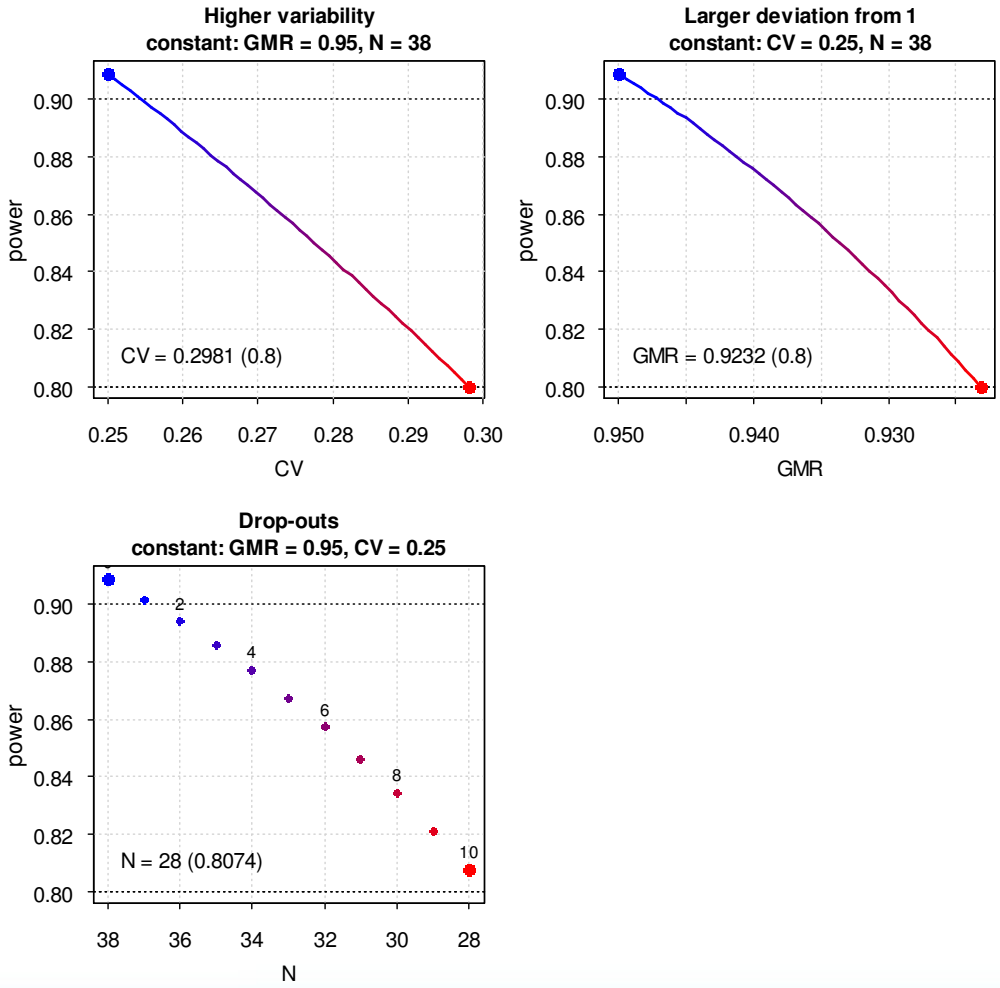
Only power is accessible.

- The sample size is searched in an iterative procedure until at least the desired power is obtained.
  - Exact methods for ABE in parallel, crossover, and replicate designs available.
  - Simulations suggested for Group-Sequential and Two-Stage Designs.
- According to ICH E9 a sensitivity analysis is mandatory to explore the impact on power if values deviate from assumptions.

Sample Size

Example

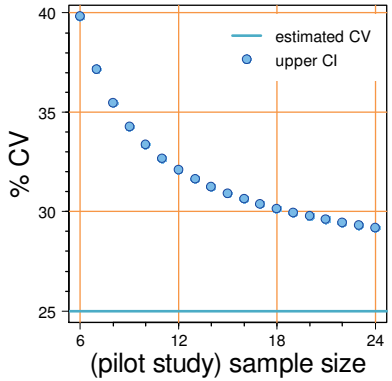
- 2×2×2, assumed *GMR* 0.95, CV_w 0.25, desired power 0.9, min. acceptable power 0.8.
  - Sample size 38 (power 0.909)
  - CV_w can increase to 0.298 (rel. +19%)
  - *GMR* can decrease to 0.923 (rel. –2.8%)
  - 10 drop-outs acceptable (rel. –26%)
  - Most critical is the *GMR*!



Dealing with Uncertainty

Nothing is 'carved in stone'.

- Never assume perfectly matching products.
  - Generally a Δ of not better than 5% should be assumed (0.950 – 1.053).
  - For HVD(P)s do not assume a Δ of <10% (0.900 – 1.111).
- Do not use the *CV* but one of its confidence limits.
  - Suggested α 0.2 (here: the producer's risk).
  - For ABE the upper CL.
  - For reference-scaling 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.

Dealing with Uncertainty

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 α is possible, *i.e.*,
    a small α in the interim and a large one in the final analysis.
    Examples: Haybittle/Peto (α_1 0.001, α_2 0.049), O'Brien/Fleming (α_1 0.005, α_2 0.048), Zheng et al. (α_1 0.01, α_2 0.04).
    May require α-spending functions (Lan/DeMets, Jennison/Turnbull) in order to control the Type I Error.

Dealing with Uncertainty

(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'
    1. The same adjusted α is applied in both stages (regardless whether a study stops in the first stage or proceeds to the second stage).
    2. An unadjusted α may be used in the first stage, dependent on interim power.
  - All published methods are valid only for a range of combinations of stage 1 sample sizes, *CV*s, *GMR*s, and desired power.
  - Contrary to common believes no analytical proof of keeping the TIE exist. It is the responsibility of the sponsor to demonstrate (*e.g.*, in simulations) that the consumer risk is preserved.

Excursion

Type I Error.

- In BE the Null Hypothesis (H_0) is *inequivalence*.
  - TIE = Probability of falsely rejecting H_0 (*i.e.*, accepting H_1 and claiming BE).
  - Can be calculated for the nominal significance level (α) assuming a point estimate at one of the limits of the acceptance range.
    - Example: 2×2×2 crossover, *CV* 20%, *n* 20, α 0.05, θ_0 0.80 or 1.25.

```
library(PowerTOST)
AL <- c(1-0.20, 1/(1-0.20)) # common acceptance range: 0.80–1.25
power.TOST(CV=0.20, n=20, alpha=0.05, theta0=AL[1])
[1] 0.0499999
power.TOST(CV=0.20, n=20, alpha=0.05, theta0=AL[2])
[1] 0.0499999
```

    - TOST is not a uniformly most powerful test.

```
power.TOST(CV=0.20, n=12, alpha=0.05, theta0=AL[2])
[1] 0.04976374
```

    - However, the TIE never exceeds the nominal level.

```
power.TOST(CV=0.20, n=72, alpha=0.05, theta0=AL[2])
[1] 0.05
```

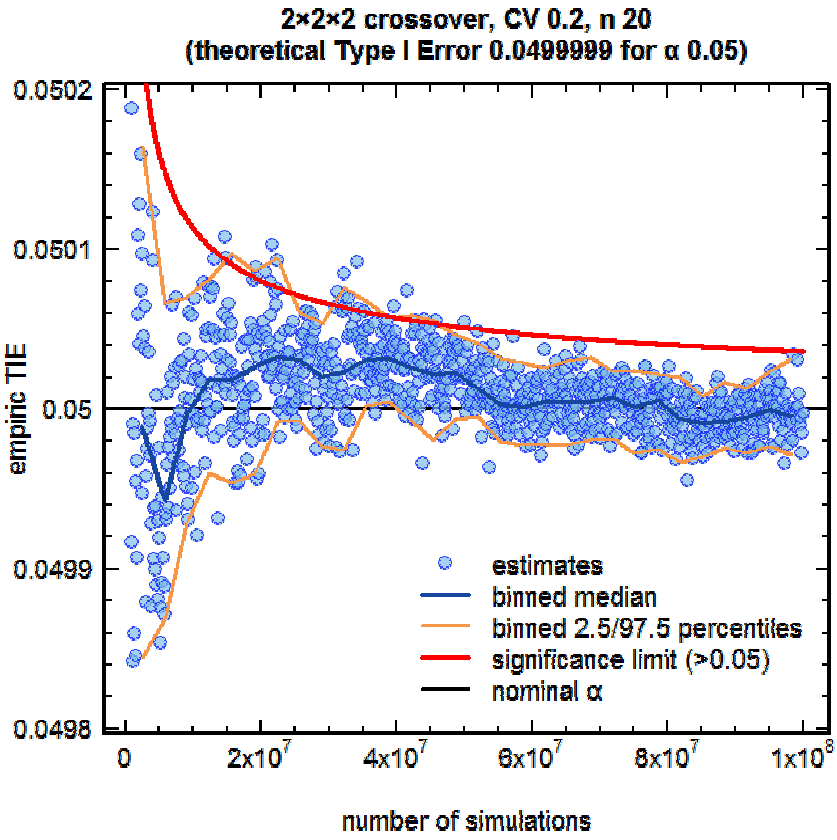
Excursion

Type I Error.

- Alternatively perform simulations to obtain an empiric TIE.

```
power.TOST.sim(CV=0.20, n=20, alpha=0.05, theta0=AL[2],
                    nsims=1e8)
[1] 0.04999703
```

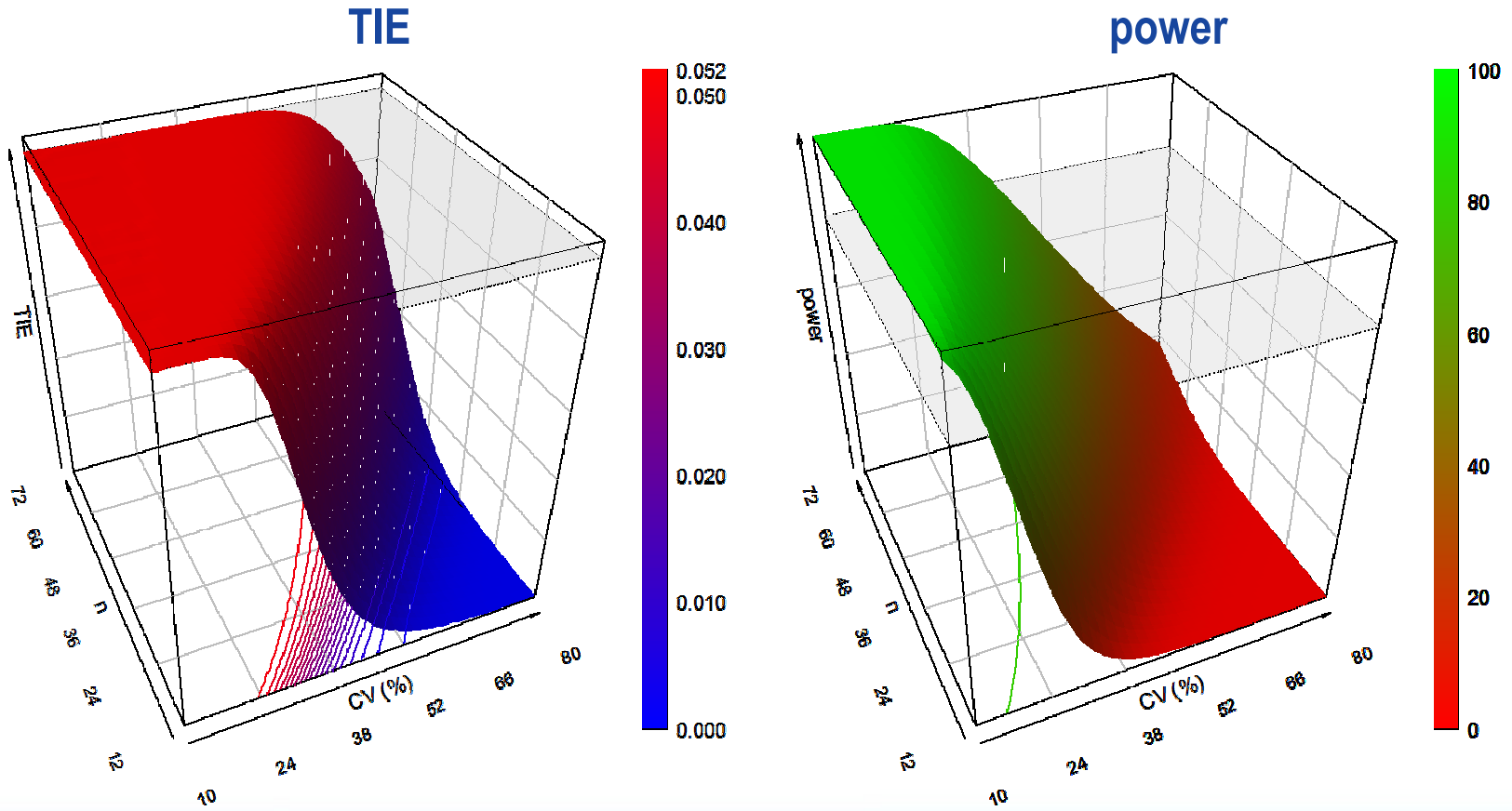
- In other settings (*i.e.*, frameworks like Two-Stage Designs or reference-scaled ABE) analytical solutions for power – and therefore, the TIE – are not possible.



Excursion

Type I Error and power.

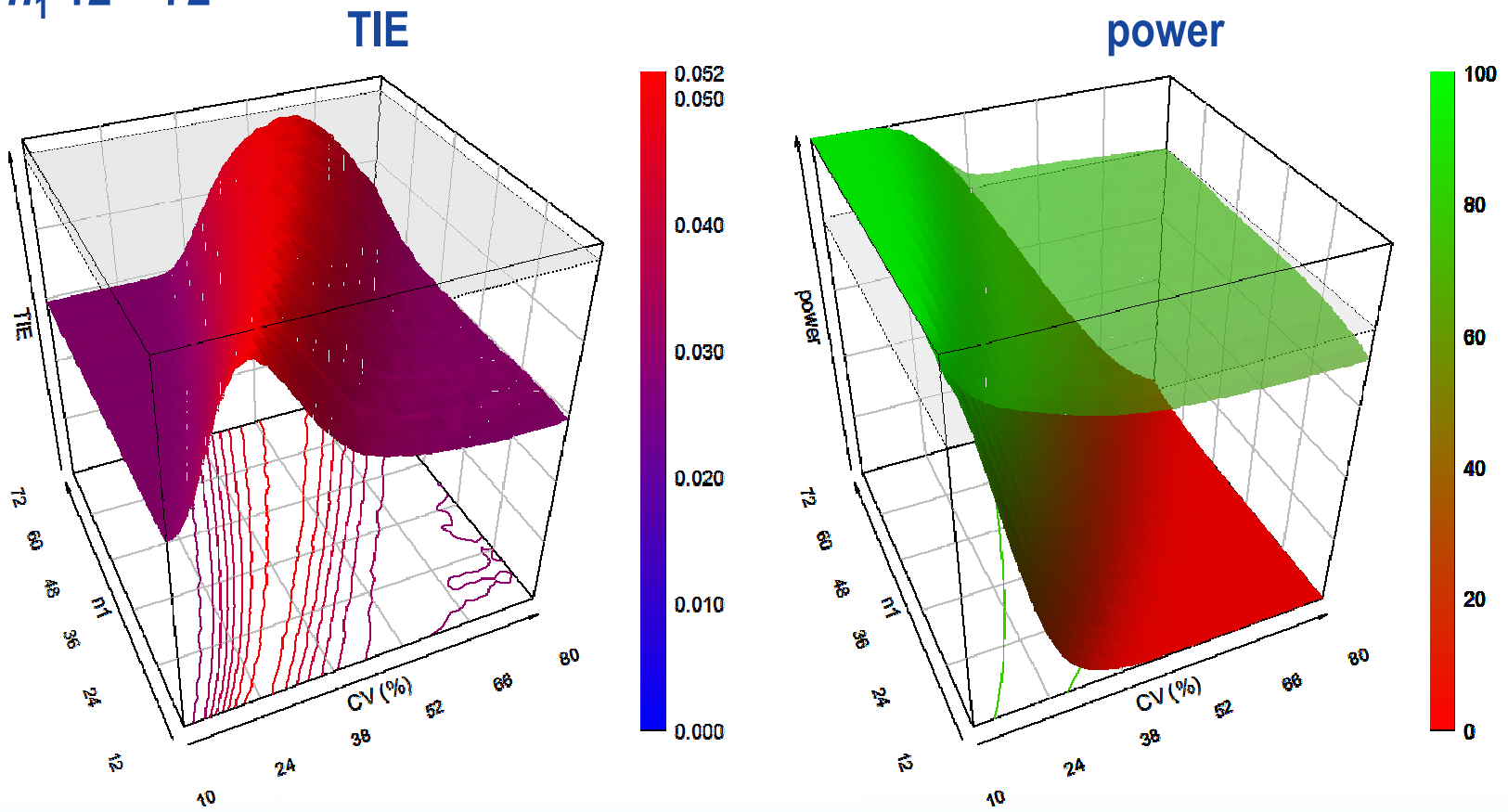
- Fixed sample 2×2×2 design (α 0.05). *GMR* 0.95, *CV* 10 – 80%, *n* 12 –72



Excursion

Type I Error and power.

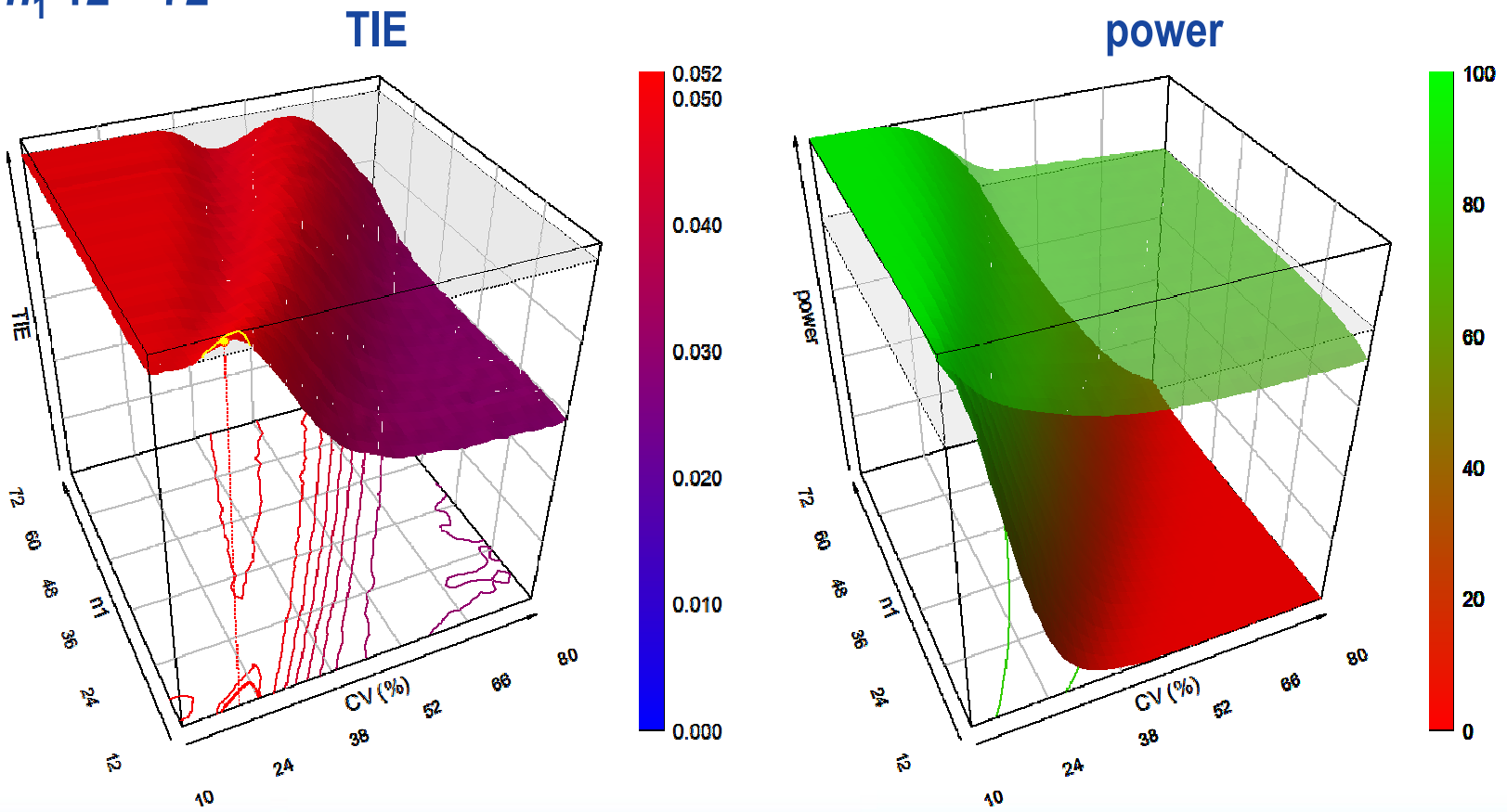
- 'Type 1' TSD (Potvin Method B, α_{adj} 0.0294). *GMR* 0.95, *CV* 10 – 80%, n_1 12 – 72



Excursion

Type I Error and power.

- 'Type 2' TSD (Potvin Method C, α_{adj} 0.05|0.0294). *GMR* 0.95, *CV* 10 – 80%, n_1 12 – 72



Group-Sequential Designs

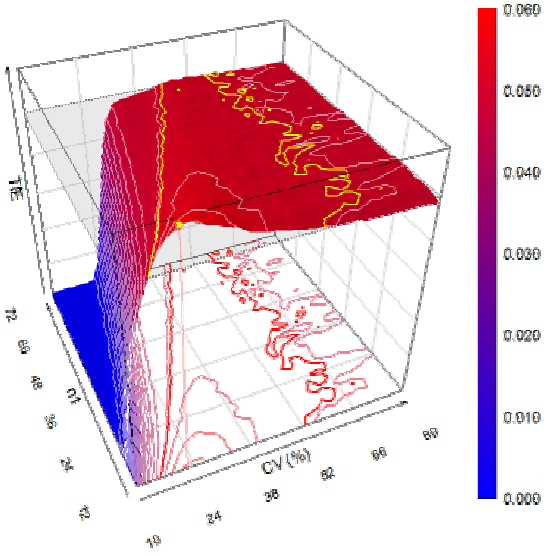
Long and accepted tradition in clinical research (phase III).

- 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 α is possible, *i.e.*,
    - a small α 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 (α_1 0.001, α_2 0.049), O'Brien/Fleming (α_1 0.005, α_2 0.048).
    - *Not* developed for crossover designs and sample size re-estimation (fixed n_1 and variable *N*): Lower α_2 or α-spending functions (Lan/DeMets, Jennison/Turnbull) are needed in order to control the Type I Error.
    - Zheng et al. (2015) for BE in crossovers (α_1 0.01, α_2 0.04) keeps the TIE.

Group-Sequential Designs

Type I Error.

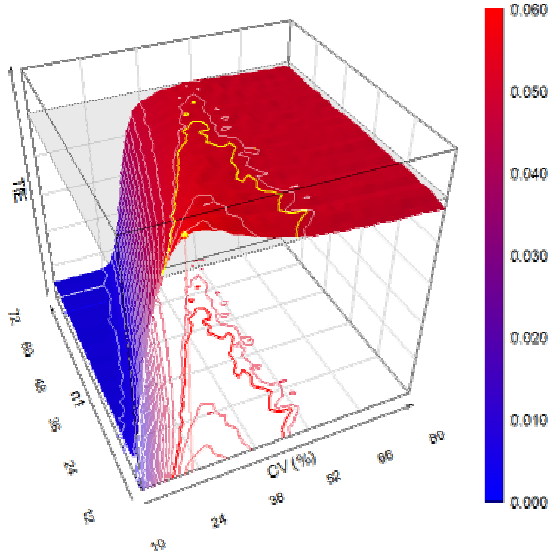
Haybittle/Peto
α_1 0.001, α_2 0.049



Maximum 0.05849

α_2 0.0413 needed
to control the TIE

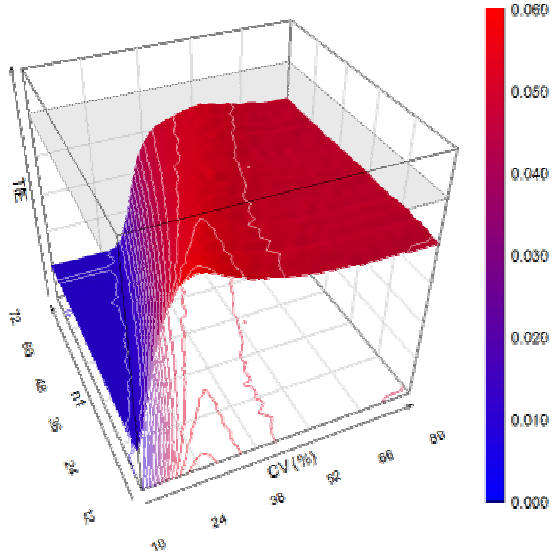
O'Brien/Fleming
α_1 0.005, α_2 0.048



Maximum 0.05700

α_2 0.0415 needed
to control the TIE

Zheng et al.
α_1 0.01, α_2 0.04



Maximum 0.04878

Group-Sequential Designs

Review of Guidelines.

- Australia (2004), Canada (Draft 2009)
  - Application of Bonferroni's correction (α_{adj} 0.025).
  - Theoretical TIE ≤0.0494.
  - For *CV*s and samples sizes common in BE the TIE generally is ≤0.04.
- Canada (2012)
  - Pocock's α_{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! α-spending functions can control the TIE (but are *not* mentioned in the guidance).

(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 α_{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%).
  - Two 'Types'
    1. The same adjusted α is applied in both stages (regardless whether a study stops in the first stage or proceeds to the second stage).
    2. An unadjusted α may be used in the first stage, dependent on interim power.

(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	α_{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 (1–2α_1) CI.

Type	Method	CV_w	Futility region	α_1	α_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

(Adaptive) Sequential Two-Stage Designs

Review of Guidelines.

- EMA (Jan 2010)
  - Acceptable.
  - α_{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 α-spending functions or adaptive methods (where α_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 2016…

(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.
  - 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).
- Russia (2013)
  - 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.
  - State stopping criteria unambiguously 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 α_{adj} and validating for TIE and power takes ~20 minutes with the R-package Power2Stage (open source).

(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 some concerns about the validity of methods based on simulations.
  More about that in the second presentation.

(Adaptive) Sequential Two-Stage Designs

Outlook.

- Selecting a candidate formulation from a higher-order crossover; continue with 2×2×2 in the second stage.
- Continue a 2×2×2 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).

Two-Stage Sequential Designs
Industry Perspective

Thank You!
Open Questions?

Helmut Schütz

BEBAC

Consultancy Services for
Bioequivalence and Bioavailability Studies
1070 Vienna, Austria

helmut.schuetz@bebac.at

References

Diletti E, Hauschke D, Steinijans VW. *Sample size determination for bioequivalence assessment by means of confidence intervals*. Int J Clin Pharm Ther Toxicol. 1991;29(1):1–8.

Labes D, Schütz H, Lang B. *PowerTOST: Power and Sample size based on Two One-Sided t-Tests (TOST) for (Bio)Equivalence Studies*. R package version 1.4-2. 2016. https://cran.r-project.org/package=PowerTOST

Pocock SJ. *Group sequential methods in the design and analysis of clinical trials*. Biometrika. 1977;64:191–9.

Gould LA. *Group sequential extension of a standard bioequivalence testing procedure*. J Pharmacokinet Biopharm. 1995;23:57–86. DOI 10.1007/BF02353786

Haybittle JL. *Repeated assessment of results in clinical trials of cancer treatment*. Br J Radiol. 1971;44:793–7. DOI 10.1259/0007-1285-44-526-793

Peto R et al. *Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples*. Br J Cancer. 1977;35:2–39. DOI 10.1038/bjc.1977.1

O'Brien PC, Fleming TR. *A multiple testing procedure for clinical trials*. Biometrics. 1979;35:549–56.

Lan KG, DeMets DL. *Discrete sequential boundaries for clinical trials*. Biometrika. 1983;70:659–63.

Hauck WW, Preston PE, Bois FY. *A Group Sequential Approach to Crossover Trials for Average Bioequivalence*. J Biopharm Stat. 1997;71(1):87–96. DOI 10.1080/10543409708835171

Jennison C, Turnbull BW. *Equivalence tests*. In: Jennison C, Turnbull BW, editors. *Group sequential methods with applications to clinical trials*. Boca Raton: Chapman & Hall/CRC; 1999. p. 142–57.

Wittes J et al. *Internal pilot studies I: type I error rate of the naive t-test*. Stat Med. 1999;18(24):3481–91. DOI 10.1002/(SICI)1097-0258(19991230)18:24<3481::AID-SIM301>3.0.CO;2-C

Potvin D et al. *Sequential design approaches for bioequivalence studies with crossover designs*. Pharmaceut Statist. 2008;7(4):245–62. DOI 10.1002/pst.294

Montague TH et al. *Additional results for 'Sequential design approaches for bioequivalence studies with crossover designs'*. Pharmaceut Statist. 2012;11(1):8–13. DOI 10.1002/pst.483

García-Arieta A, Gordon J. *Bioequivalence Requirements in the European Union: Critical Discussion*. AAPS J. 2012;14(4):738–48. DOI 10.1208/s12248-012-9382-1

Davit B et al. *Guidelines for Bioequivalence of Systemically Available Orally Administered Generic Drug Products: A Survey of Similarities and Differences*. AAPS J. 2013;15(4):974–90. DOI 10.1208/s12248-013-9499-x

Karalis V, Macheras P. *An insight into the properties of a two-stage design in bioequivalence studies*. Pharm Res. 2013;30(7):1824–35. DOI 10.1007/s11095-013-1026-3

Karalis V. *The role of the upper sample size limit in two-stage bioequivalence designs*. Int J Pharm. 2013;456(1):87–94. DOI 10.1016/j.ijpharm.2013.08.013

Fuglsang A. *Futility rules in bioequivalence trials with sequential designs*. AAPS J. 2014;16(1):79–82. DOI 10.1208/s12248-013-9540-0

Fuglsang A. *Sequential Bioequivalence Approaches for Parallel Designs*. AAPS J. 2014;16(3):373–8. DOI 10.1208/s12248-014-9571-1

Karalis V, Macheras P. *On the Statistical Model of the Two-Stage Designs in Bioequivalence Assessment*. J Pharm Pharmacol. 2014;66(1):48–52. DOI 10.1111/jphp.12164

Golkowski D, Friede T, Kieser M. *Blinded sample size reestimation in crossover bioequivalence trials*. Pharmaceut Stat. 2014;13(3):157–62. DOI 10.1002/pst.1617

Jones B, Kenward MG. *Chapters 12–14*. In: Jones B, Kenward MG, editors. *Design and analysis of crossover trials*, Chapman & Hall/CRC; Boca Raton. 2014. p. 365–80.

Schütz H. *Two-stage designs in bioequivalence trials*. Eur J Clin Pharmacol. 2015;71(3):271–81. DOI 10.1007/s00228-015-1806-2

Zheng Ch, Zhao L, Wang J. *Modifications of sequential designs in bioequivalence trials*. Pharmaceut Statist. 2015;14(3):180–8. DOI 10.1002/pst.1672

Kieser M, Rauch G. *Two-stage designs for crossover bioequivalence trials*. Stat Med. 2015;34(16):2403–16. DOI 10.1002/sim.6487

König F, Wolfsegger M, Jaki T, Schütz H, Wasmer G. *Adaptive two-stage bioequivalence trials with early stopping and sample size re-estimation*. Trials. 2015;16(Suppl 2):P218. DOI 10.1186/1745-6215-16-S2-P218

Xu et al. *Optimal adaptive sequential designs for crossover bioequivalence studies*. Pharmaceut Statist. 2016;15(1):15–27. DOI 10.1002/pst.1721

Labes D, Schütz H. *Power2Stage: Power and Sample-Size Distribution of 2-Stage Bioequivalence Studies*. R package version 0.4-3. 2015. https://cran.r-project.org/package=Power2Stage