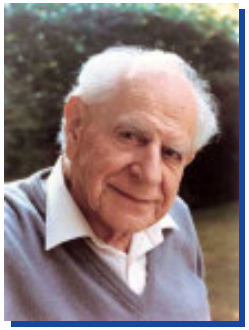
Statistical Planning and Evaluation of Bioequivalence Studies

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.

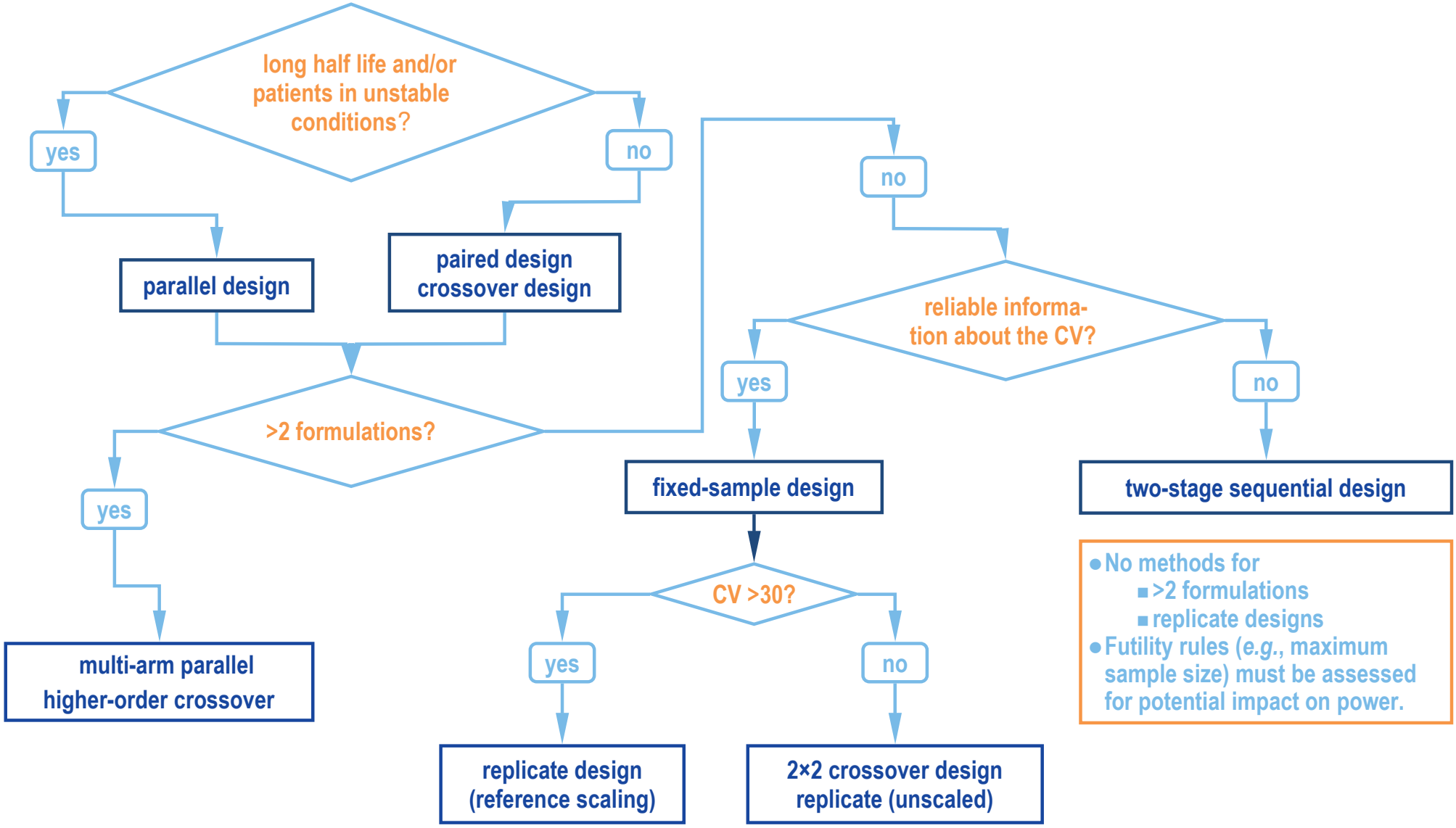


Karl R. Popper

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

Leslie Z. Benet

Study Designs



Study Designs

The more 'sophisticated' a design is, the more information can be extracted.

- **Hierarchy of designs:**
  - Full replicate (RTRT | TRTR or RTR | TRT)
    - Partial replicate (RRT | RTR | TRR)
      - 2×2×2 crossover (RT | TR)
        - Parallel (R | T)
- **Variances which can be estimated:**

Parallel:	total variance (between + within subjects)
2×2×2 crossover:	+ between, within subjects
Partial replicate:	+ within subjects (of R)
Full replicate:	+ within subjects (of R and T)

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

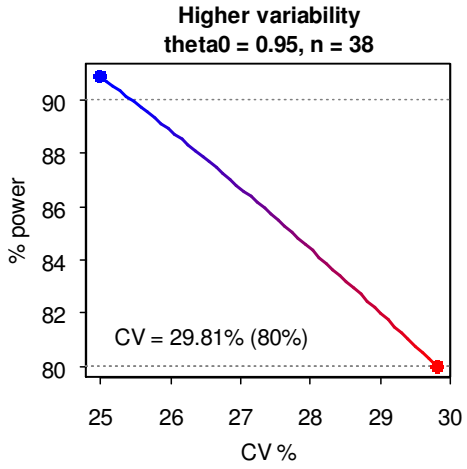
Only power is accessible.

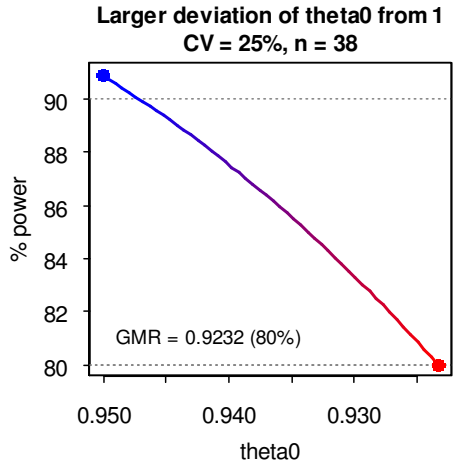
- The sample size is searched in an iterative procedure until at least the desired power is obtained.
  - Exact methods for ABE in parallel, crossover, and replicate designs available.
  - Simulations required for all reference-scaled ABE methods.
- BE has to be shown for all relevant PK metrics.
  - Since for the EMA SABE is only acceptable for C_{max}, the sample size might be mandated by – also highly variable – *AUC*.
  - Might lead to the paradox situation of approving products with large deviations in C_{max}.
- According to ICH E9 a sensitivity analysis is mandatory to explore the impact on power if values deviate from assumptions.

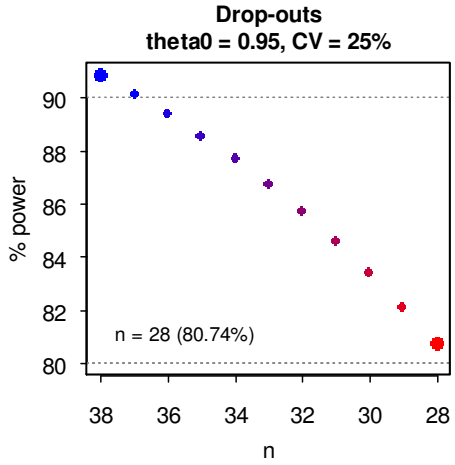
Sample Size

Example

- 2×2×2, assumed *GMR* 0.95, CV_w 25%, desired power 90%, min. acceptable power 80%.
  - Sample size 38 (power 90.9%)
  - CV_w can increase to 29.8% (rel. +19%)
  - *GMR* can decrease to 0.923 (rel. –2.8%)
  - 10 dropouts acceptable (rel. –26%)
  - Most critical is the *GMR*!



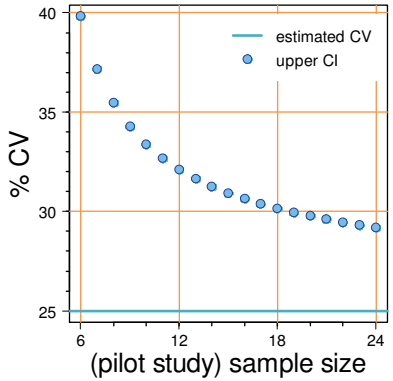




Dealing with Uncertainty

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 to 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, on 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, normal distributed data with known variance, and one interim at N/2.
  - That's not what we have in BE: equivalence, lognormal data with unknown variance. Furthermore – due to dropouts – the interim might not be 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 (0.001 | 0.049), O'Brien/Fleming (0.005 | 0.048).
    May need 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, 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 size, *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 in simulations that the consumer risk is preserved.

Parallel Designs

Two or more groups

- Advantages
  - Studies of endogenous compounds in healthy volunteers or patients where a feedback-loop prevents a crossover.
  - Studies in patients, where the condition of the disease irreversibly changes.
  - Straigthforward statistical analysis.
- Disadvantages
  - Higher sample sizes than in crossovers to achieve desired power.

Crossover Designs

Two-sequence, two-period, two-treatment (aka 2×2×2)

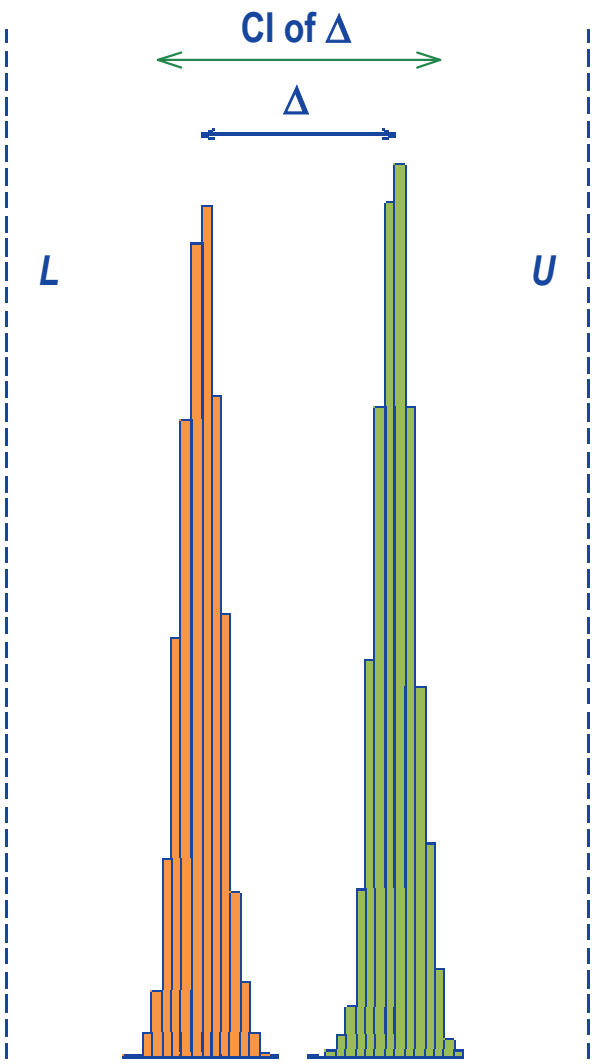
- Advantages
  - Accounts for potential period effects.
  - Healthy volunteers or patients with stable conditions (*e.g.*, asthma).
  - Globally applied standard protocol for bioequivalence, drug-drug of food-drug interaction studies.
  - Straigthforward statistical analysis.
- Disadvantages
  - Not optimal for drugs with long half life
    → parallel design.
  - Not optimal for highly variable drugs / drug products
    → replicate design with reference-scaling.

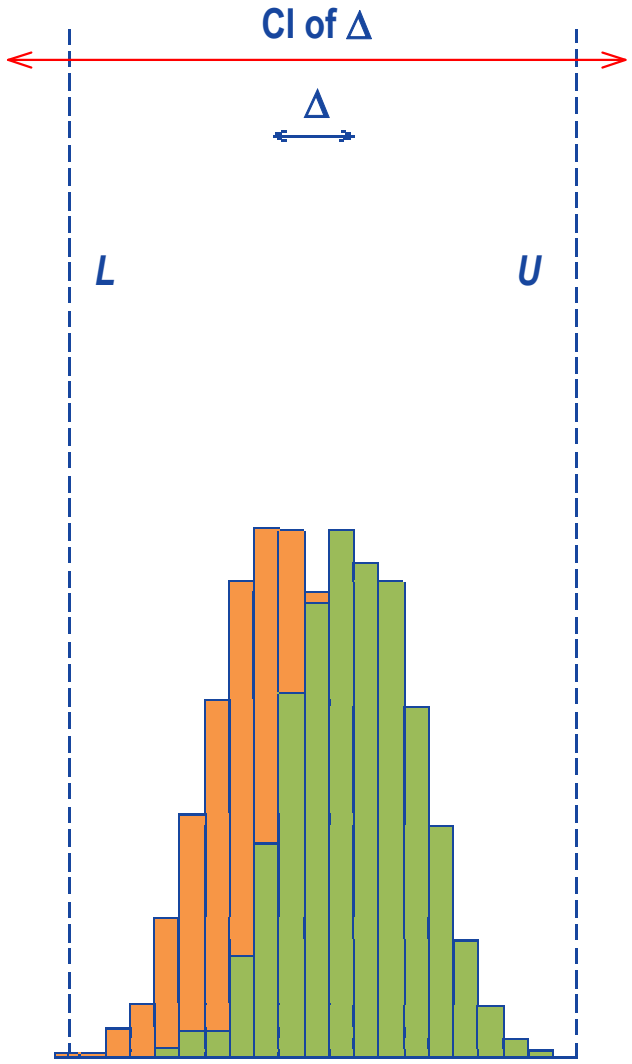
Higher Order Crossover Designs

Latin Squares (3×3, 4×4, …), Williams' Designs (6×3, 4×4, …)

- Advantages
  - Standard designs for establishment of dose proportionality.
  - Allows to choose between candidate test formulations in a pilot study or comparison of a test formulation with two references.
  - Food-effect of T and R in one study.
  - Statistically more demanding than 2×2×2.
- Disadvantages
  - No consensus how pooled variances should be handled.
    - EMA: Ignore 'not relevant' treatment arms.
    - FDA: Full model.

Highly Variable Drugs / Drug Products





Counterintuitive concept of BE:

Two formulations with a large difference in means are declared bioequivalent if variances are low, but not BE – even if the difference is quite small – due to high variability.

Modified from Tothfálusi *et al.* (2009), Fig. 1

HVD(P)s – Reference-scaling

It may be almost impossible to demonstrate BE with a reasonable sample size.

- **Reference-scaling (*i.e.*, widening the acceptance range based of the variability of the reference) in 2010 introduced by the FDA and EMA.**
  - **Requires a replicate design, where at least the reference product is administered twice.**
  - **Smaller sample sizes compared to a standard 2×2×2 design but outweighed by increased number of periods.**
  - **Similar total number of individual treatments.**
  - **Any replicate design can be evaluated for 'classical' (unscaled) Average Bioequivalence (ABE) as well. Switching CV_{wR} 30%:**
    - **FDA: *AUC* and C_{max}**
    - **EMA: C_{max}; MR products additionally: C_{min}, C_t, partial *AUC*s**
    - **HC: *AUC***

HVD(P)s – Reference-scaling

Models (in log-scale)

- **ABE Model**
  - A difference Δ of ≤20% is considered to be clinically not relevant.
  - The limits of the acceptance range are fixed to $\ln(1-\Delta) = \ln((1-\Delta)^{-1})$ or $L \sim -0.2231$ and $U \sim +0.2231$.
  - The consumer risk is fixed with 0.05. BE is concluded if the $100(1-2\alpha)$ confidence interval lies entirely within the acceptance range.

$$-\theta_A \leq \mu_T - \mu_R \leq +\theta_A$$

- **SABEL Model**
  - Switching condition θ_S is derived from the regulatory standardized variation σ_0 (proportionality between acceptance limits in log-scale and σ_{wR} in the highly variable region).

$$-\theta_S \leq \frac{\mu_T - \mu_R}{\sigma_{wR}} \leq +\theta_S$$

HVD(P)s – Reference-scaling

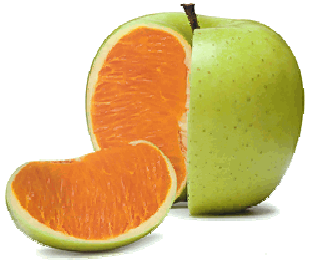
The EMA's Approach

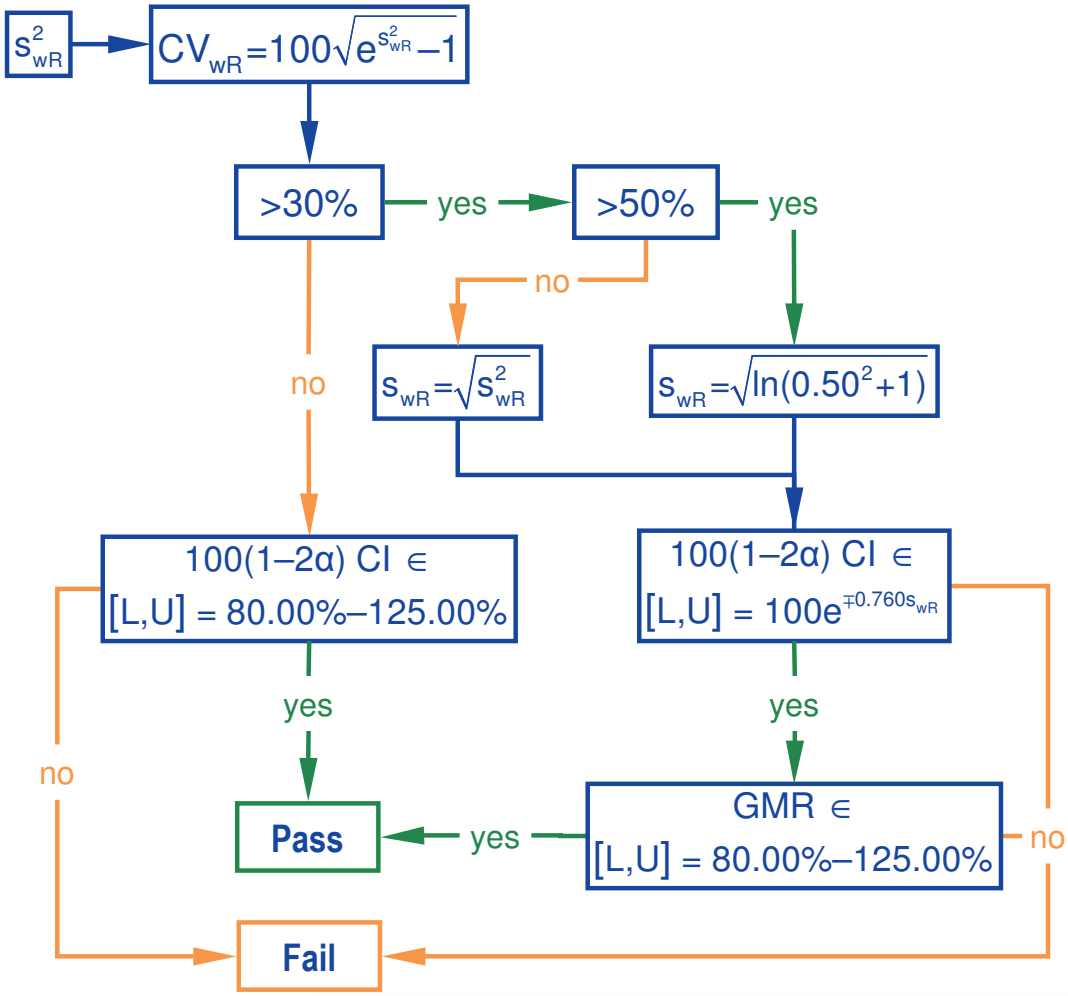
- Average Bioequivalence with Expanding Limits (crippled from Endrényi and Tóthfalusi 2009)
  - Justification that the widened acceptance range is clinically not relevant (important – different to the FDA).
  - Assumes identical variances of T and R [*sic*] like in a 2×2×2.
  - All fixed effects model according to the Q&A-document preferred.
  - Mixed-effects model (allowing for unequival variances) is 'not compatible with CHMP guideline'…
  - Scaling limited at a maximum of CV_{wR} 50% (*i.e.*, to 69.84 – 143.19%).
  - GMR within 0.8000 – 1.2500.
  - Demonstration that CV_{wR} >30% is not caused by outliers (box plots of studentized intra-subject residuals?)…
  - ≥12 subjects in sequence RTR of the 3-period full replicate design.

HVD(P)s – Reference-scaling

The EMA's Approach

- **Decision Scheme**
  - The Null Hypothesis is *specified* in the face of the data.
  - Acceptance limits themselves become random variables.
  - Type I Error (consumer risk) might be inflated.





HVD(P)s – Reference-scaling

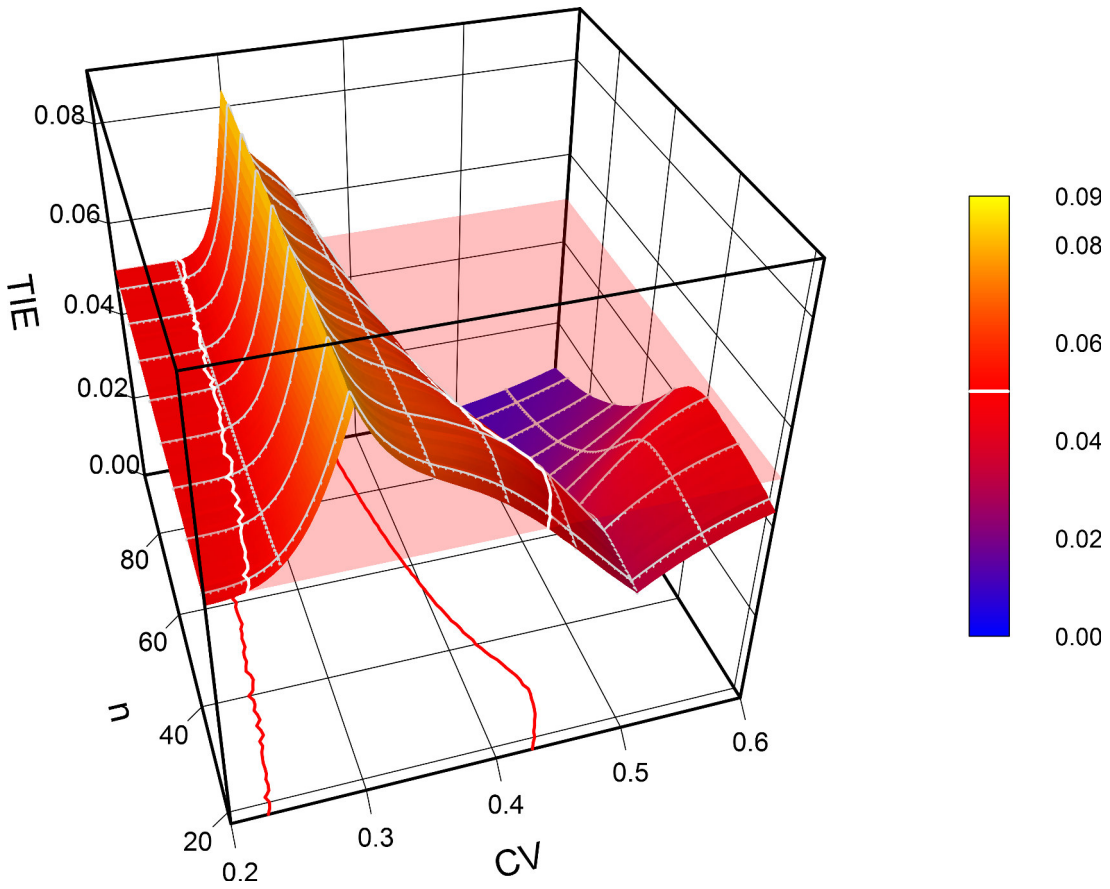
Assessing the Type I Error (TIE)

- TIE = falsely concluding BE at the limits of the acceptance range.
  In ABE the TIE is ≤0.05 at 0.80 and ≤0.05 at 1.25.
- Due to the decision scheme no direct calculation of the TIE at the scaled limits is possible;
  → extensive simulations required (10^6 BE studies mandatory).
- Inflation of the TIE suspected.
  (Chow *et al*. 2002, Willavazie & Morgenthien 2006, Chow & Liu 2009).
- Confirmed.
  - ABEL
    (Tóthfalusi & Endrényi 2009, BEBA-Forum 2013, Wonnemann *et al*. 2015, Muñoz *et al*. 2015, Labes & Schütz 2016).
  - RSABE
    (Tóthfalusi & Endrényi 2009, BEBA-Forum 2013, Muñoz *et al*. 2015).

HVD(P)s – Reference-scaling

Example

- RTRT | TRTR
  sample size 18 – 96
  CV_{wR} 20% – 60%
  - TIE_{max} 0.0837.
  - Relative increase of the consumer risk 67%!



HVD(P)s – Reference-scaling

What is going on here?

- SABE is stated in model *parameters* …

$$-\theta_S \le \frac{\mu_T - \mu_R}{\sigma_{wR}} \le +\theta_S$$

… which are *unknown*.

  - Only their *estimates* (*GMR*, s_{wR}) are accessible in the actual study.
  - At CV_{wR} 30% the decision to scale will be wrong in ~50% of cases.
  - If moving away from 30% the chances of a wrong decision decrease and hence, the TIE.
  - At high *CV*s (>43%) both the scaling cap and the *GMR*-restriction help to maintain the TIE <0.05).

HVD(P)s – Reference-scaling

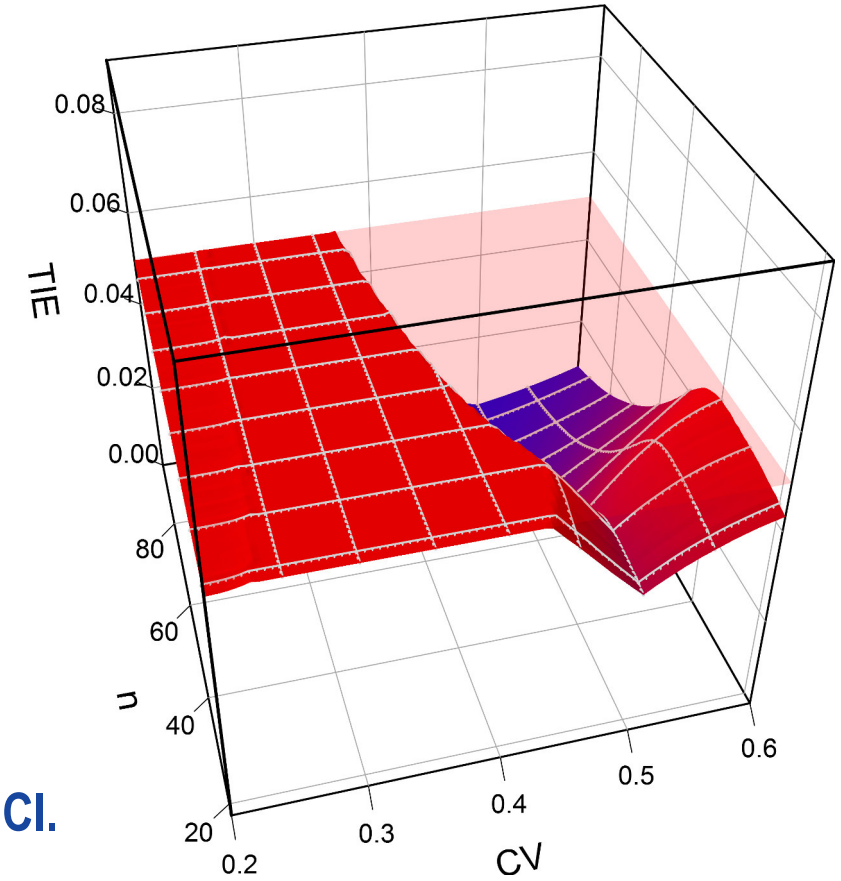
What can we do?

- Utopia
  - Agencies collect CV_{wR} from submitted studies. Pool them, adjust for designs / degrees of freedom. The EMA publishs a fixed acceptance range in the product-specific guidance. No need for replicate studies any more. 2×2×2 crossovers evaluated by ABE would be sufficient.
- Halfbaked
  - Hope that *e.g.*, Bonferroni preserves the consumer risk. Still apply ABEL, but with a 95% CI (α 0.025).
  - Drawback: Loss of power, substanial increase in sample sizes.
- Proposal
  - Iteratively adjust α based on the study's CV_{wR} and sample size – in such a way that the consumer risk is preserved.

HVD(P)s – Reference-scaling

Previous example

- Algorithm
  - Assess the TIE for the nominal α 0.05.
  - If the TIE $\leq$ 0.05, stop.
  - Otherwise adjust α (downwards) until the TIE = 0.05.
  - At CV_{wR} 30% (dependent on the sample size) α_{adj} is 0.0273 – 0.0300;
    $\rightarrow$ use a 94.00 – 94.54% CI.



Statistical Planning and Evaluation of Bioequivalence Studies

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.

Tóthfalusi L, Endrényi. *Sample Sizes for Designing Bioequivalence Studies for Highly Variable Drugs*. J Pharm Pharmaceut Sci. 2011; 15(1): 73–84.

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 1.3-5. 2016.
Available from: 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, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. *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.

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, Schabenberger O, Zucker D, Brittain D, Proschan M. *Internal pilot studies I: type I error rate of the naive t-test*. Stat Med. 1999; 18: 3481–91.

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

Potvin D, DiLiberti CE, HauckWW, Parr AF, Schuirmann DJ, Smith RA. *Sequential design approaches for bioequivalence studies with crossover designs*. Pharm Stat. 2008; 7: 245–262. DOI 10.1002/pst.294

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

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

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

Chow S-C, Shao J, Wang H. *Individual bioequivalence testing under 2×3 designs*. Stat Med. 2002; 21(5): 629–48. DOI 10.1002/sim.1056

Willavize SA, Morgenthien EA. *Comparison of models for average bioequivalence in replicated crossover designs*. Pharm Stat. 2006; 5(3): 201–11. DOI 10.1002/pst.212

Endrényi L, Tóthfalusi L. *Regulatory Conditions for the Determination of Bioequivalence of Highly Variable Drugs*. J Pharm Pharmaceut Sci. 2009; 12(1): 138–49.

Wonnemann M, Frömke C, Koch A. *Inflation of the Type I Error: Investigations on Regulatory Recommendations for Bioequivalence of Highly Variable Drugs*. Pharm Res. 2015; 32(1): 135–43. DOI 10.1007/s11095-014-1450-z

Muñoz J, Daniel Alcaide D, Ocaña J. *Consumer's risk in the EMA and FDA regulatory approaches for bioequivalence in highly variable drugs*. Stat Med. (Epub 17 Nov 2015). DOI 10.1002/sim.6834

Labes D, Schütz H. *Inflation of Type I Error in the Evaluation of Scaled Average Bioequivalence, and a Method for its Control*. Submitted to Pharm Res. 2016.