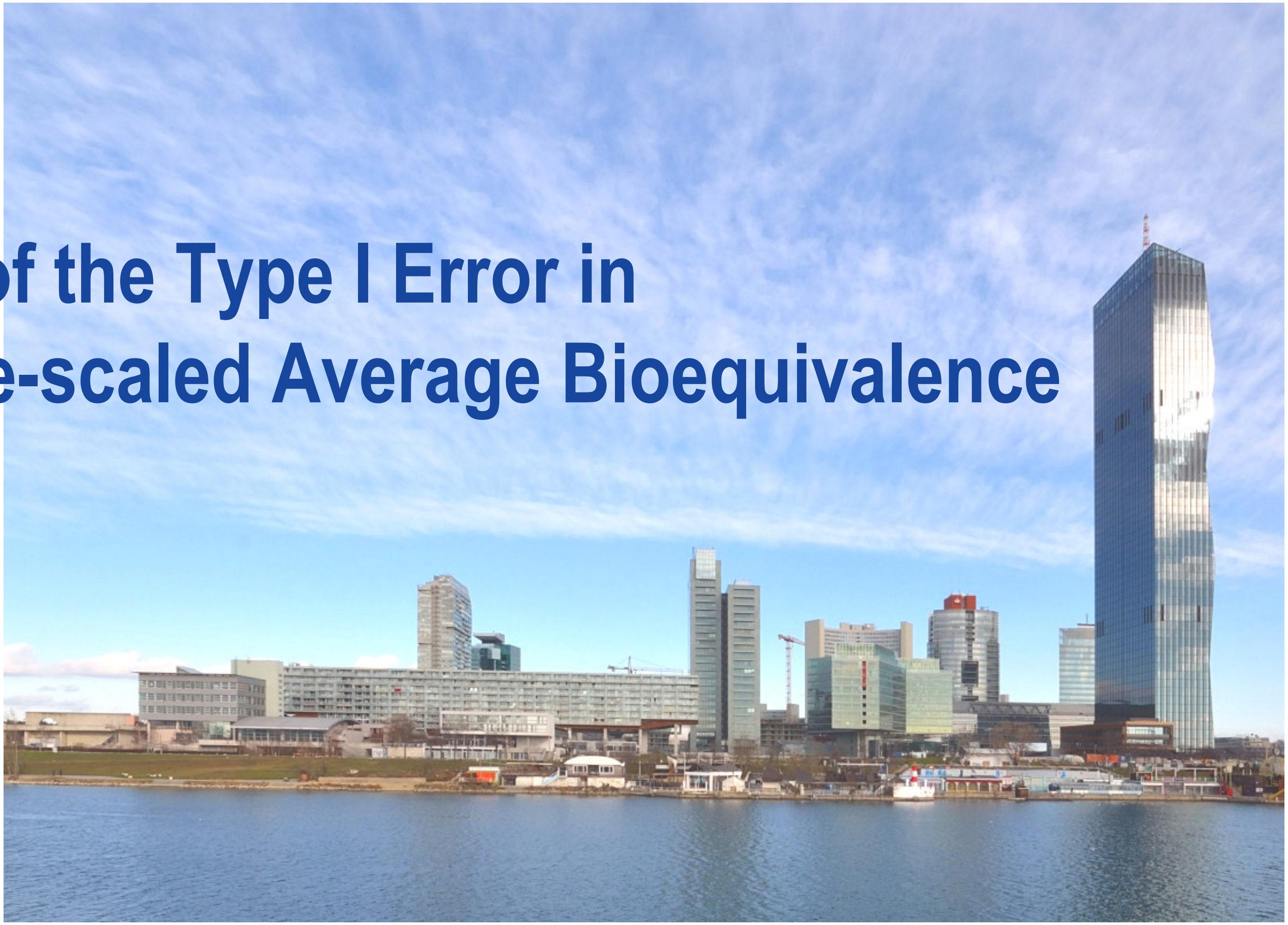
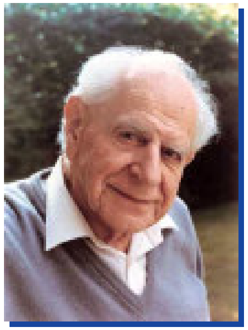


# Inflation of the Type I Error in Reference-scaled Average Bioequivalence

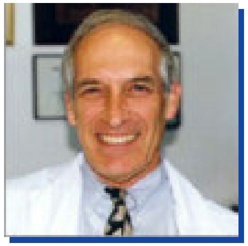
Helmut Schütz

# 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



Leslie Z. Benet

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

# Bioequivalence

## BE = (Desired) result of a comparative bioavailability study.

- Generally only for extravascular routes. Exceptions for IV:
  - Excipients which may interact with the API (complex formation).
  - Case-by-case: Liposomal formulations, emulsions.
- Same active substance.
  - Focus on the ‘core’ API
    (*different* salts, esters, isomers, complexes contain the *same* API).
- Same molar dose.
- Clinically not relevant difference: Δ 20% (NTIDs 10%, HVD(P)s >20%).
- 100(1 – 2*α*) confidence interval of PK-metrics within [1 – Δ, (1 – Δ)$^{-1}$].
  - $AUC_{0-t}$ (extent of BA)
  - $C_{max}$ (rate of BA)
  - $t_{max}$, $AUC_{0-\tau}$, $C_{max,ss}$, $C_{min,ss}$, $C_{\tau,ss}$, *%PTF*, partial *AUC*s, …

# Study Designs

≥1 Test Treatment(s) compared to ≥1 Reference Treatment(s).

- Parallel Group(s)
  - APIs with (very) long half-lives.
  - Studies in patients.
- Crossover
  - Preferred design in BE.
  - More powerful than parallel (based on within subject variance).
- Replicate crossover
  - At least one treatment is administered more than once.
  - Allows estimation of within subject variance of treatment(s).
  - Required for reference-scaling.

# Study Designs

## The more ‘sophisticated’ a design is, the more information can be extracted.

Information

- 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 $\sigma^2_s$ and $\sigma^2_e$.
  - 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!

# Excursion 1

## Type I Error.

- In BE the Null Hypothesis is *inequivalence*.
  - TIE = Probability of falsely rejecting the Null (*i.e.*, claiming BE).
  - Can be calculated for the nominal significance level ($\alpha$) assuming a PE at one of the limits of the acceptance range.
    - Example: 2×2×2 crossover, *CV* 20%, *n* 20, $\alpha$ 0.05, $\theta_0$ 1.25.

```
library(PowerTOST)
AL <- c(0.80, 1.25) # common range for ABE
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 its nominal level.

```
power.TOST(CV=0.20, n=120, alpha=0.05, theta0=AL[2])
[1] 0.05
```

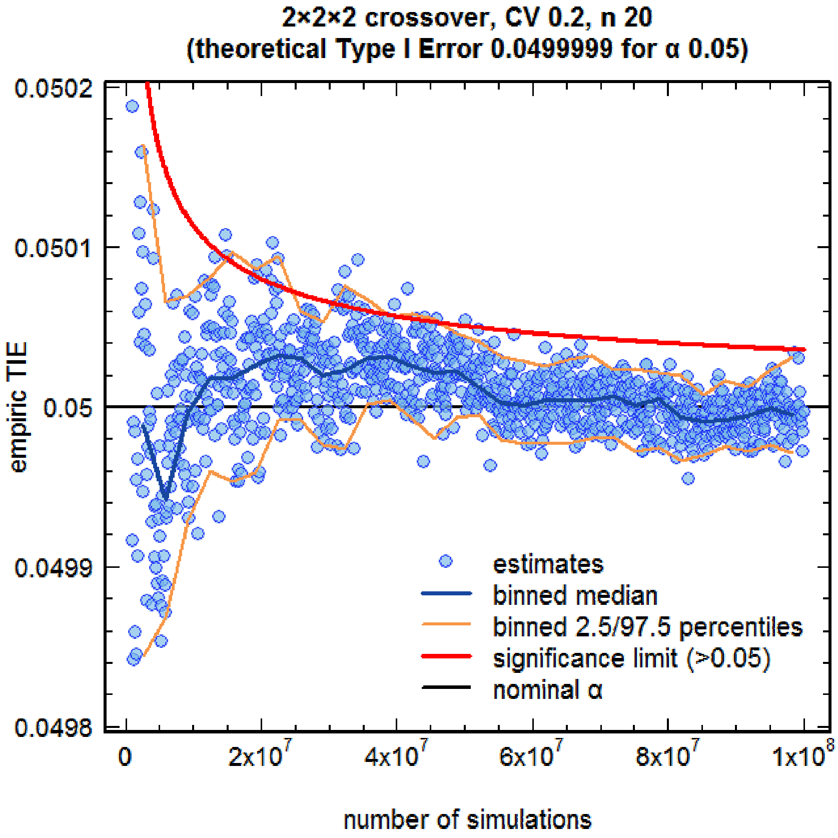
# Excursion 1

## Type I Error.

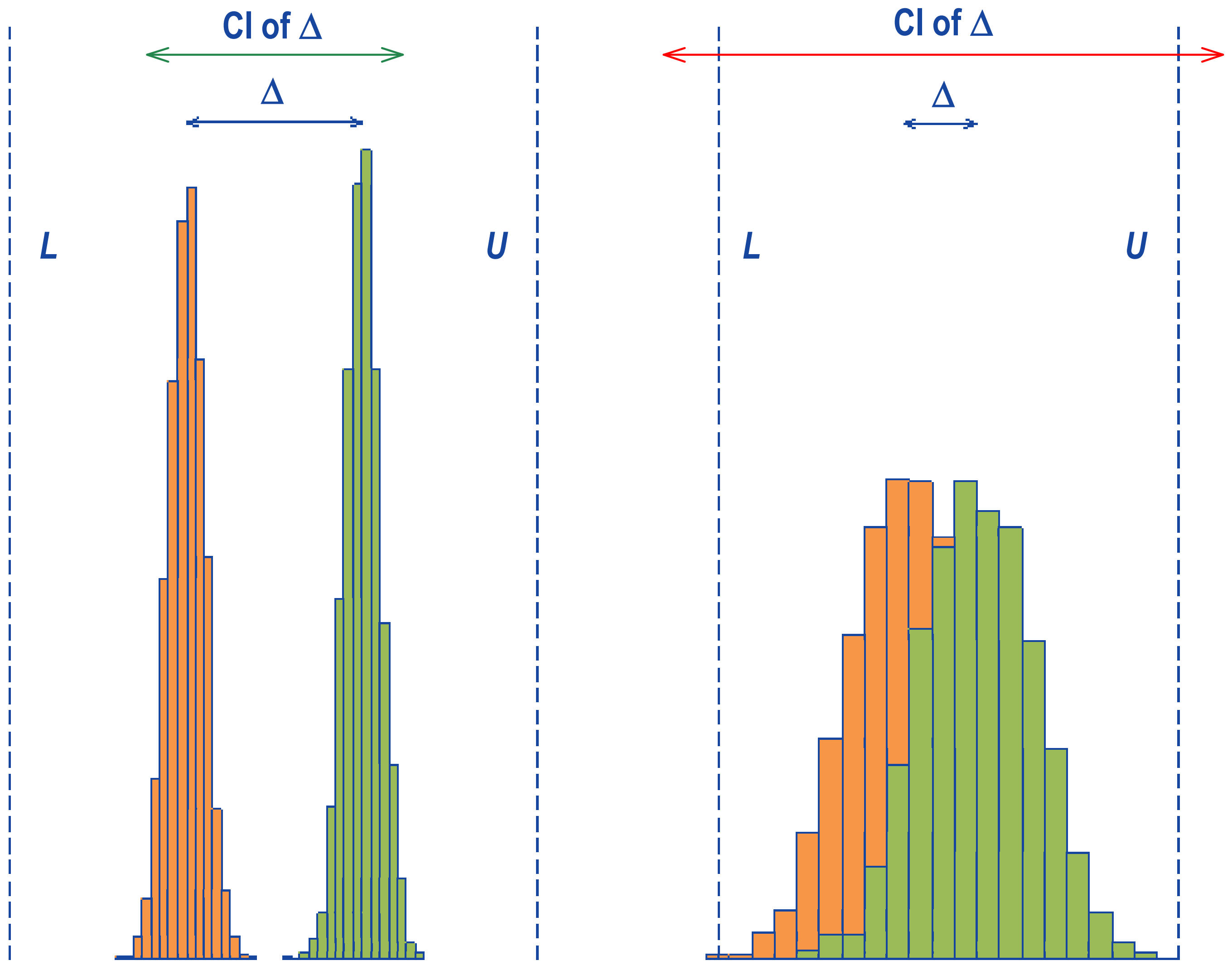
- Alternatively perform simulations to obtain an empiric TIE.

```
power.TOST.sim(CV=0.20, n=20, alpha=0.05, theta0=AL[2],
               nsims=1e8)
[1] 0.0499970
```

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



# 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 and in 2016 by Health Canada.
  - 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: $C_{ss,min}$, $C_{ss,\tau}$, partial *AUC*s
    - Health Canada: *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 [*L*, *U*] of the acceptance range are fixed to log(1 – Δ) = log((1 – Δ)$^{-1}$) or *L* ~ –0.2231 and *U* ~ +0.2231.
  - The consumer risk is fixed with 0.05. BE is concluded if the 100(1 – 2*α*) confidence interval lies entirely within the acceptance range.

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

- SABEL Model:
  - Switching condition $\theta_S$ is derived from the regulatory standardized variation $\sigma_0$ (proportionality between acceptance limits in log-scale and $\sigma_{wR}$ in the highly variable region).

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

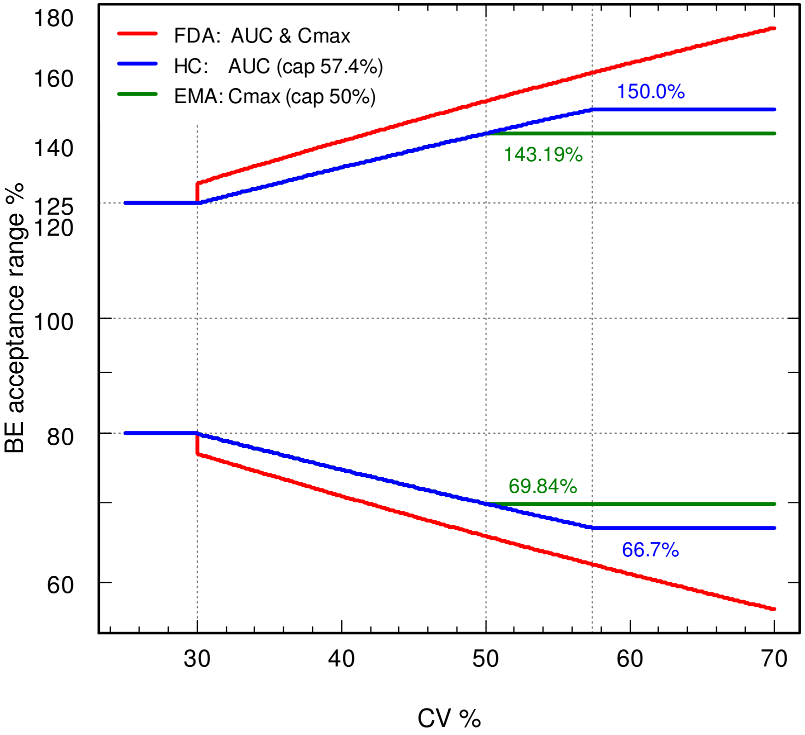
# HVD(P)s – Reference-scaling

## Regulatory Approaches.

- Bioequivalence limits derived from $\sigma_0$ and $\sigma_{wR}$

$$\theta_s = \frac{\log(1.25)}{\sigma_0},\ [L,U] = e^{\pm\theta_S \cdot \sigma_{wR}}$$

- FDA
  - Scaling $\sigma_{wR}$ 0.25 ($\theta_S$ 0.893) but applicable at $CV_{wR} \geq 30\%$.
  - Discontinuity at $CV_{wR}$ 30%.
- EMA
  - Scaling $\sigma_0$ 0.2936 ($\theta_S$ 0.760).
  - Upper cap at $CV_{wR}$ 50%.
- Health Canada
  - Like EMA but upper cap at $CV_{wR}$ 57.4%.



# HVD(P)s – Reference-scaling

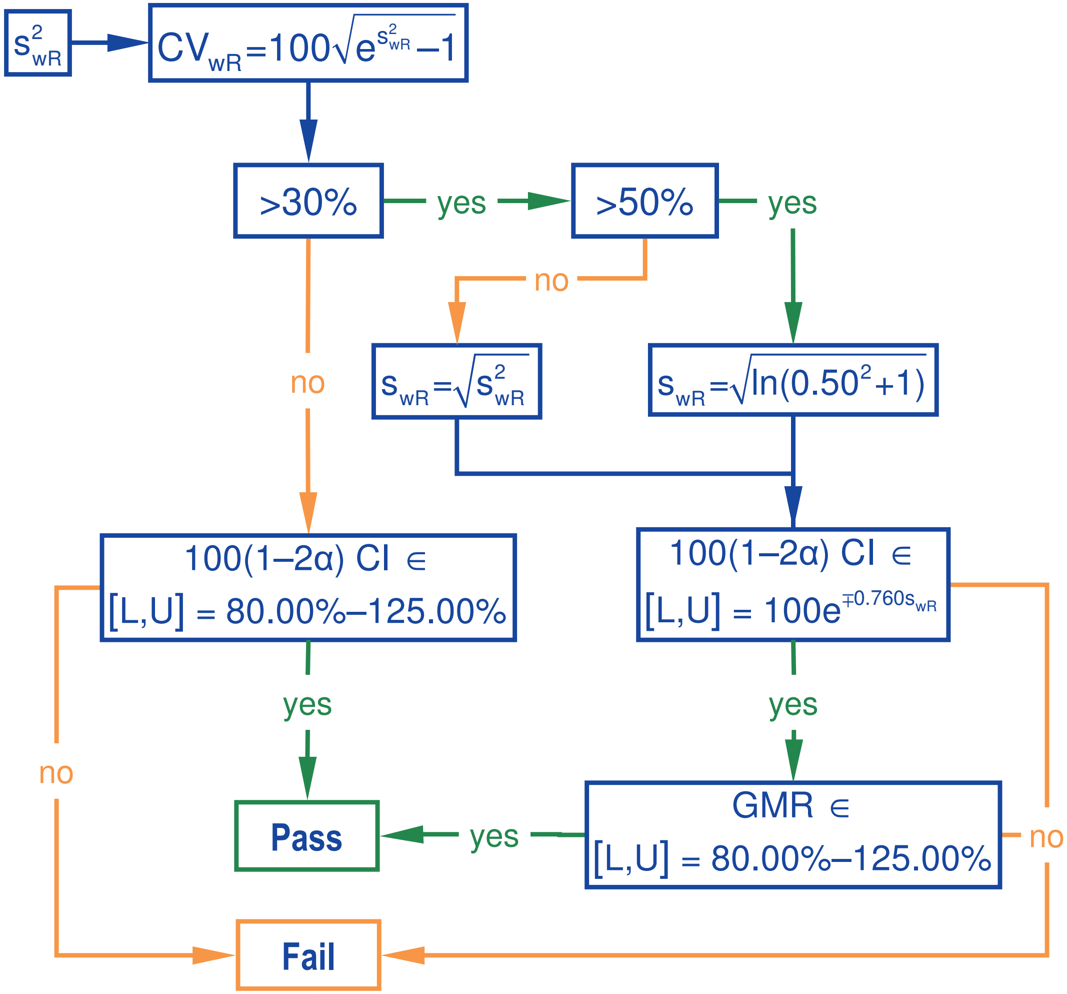
## The EMA's Approach.

- Average Bioequivalence with Expanding Limits – ABEL (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.



$s^2_{wR}$

$CV_{wR}=100\sqrt{e^{s^2_{wR}}-1}$

>30%

yes

>50%

yes

no

no

$s_{wR}=\sqrt{s^2_{wR}}$

$s_{wR}=\sqrt{\ln(0.50^2+1)}$

100(1–2α) CI ∈ [L,U] = 80.00%–125.00%

100(1–2α) CI ∈ [L,U] = $100e^{\mp 0.760 s_{wR}}$

yes

yes

no

Pass

yes

GMR ∈ [L,U] = 80.00%–125.00%

no

Fail

# HVD(P)s – Reference-scaling

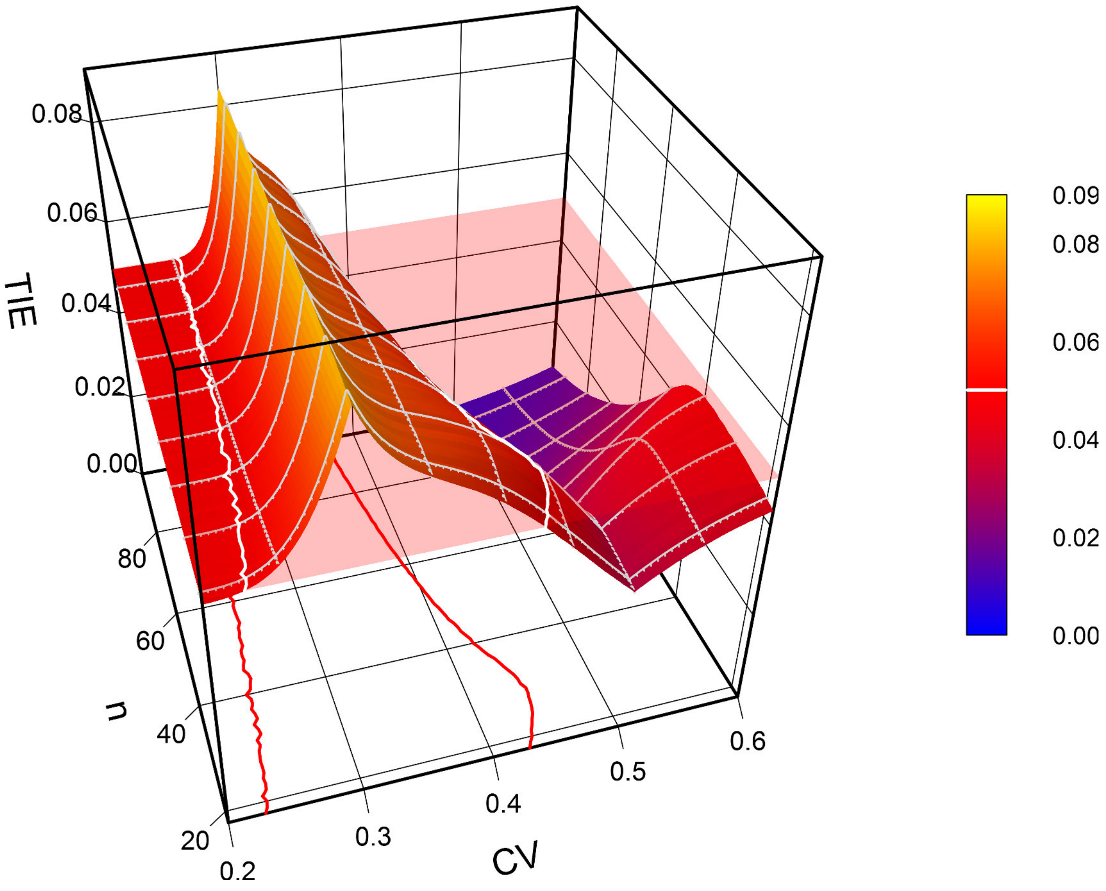
## Assessing the Type I Error (TIE).

- 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, Patterson & Jones 2012).
- Confirmed.
  - EMA's ABEL
    (Tóthfalusi & Endrényi 2009, BEBA-Forum 2013, Wonnemann *et al*. 2015, Muñoz *et al*. 2016, Labes & Schütz 2016).
  - FDA's RSABE
    (Tóthfalusi & Endrényi 2009, BEBA-Forum 2013, Muñoz *et al*. 2016).

# HVD(P)s – Reference-scaling

## Example for ABEL

- 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

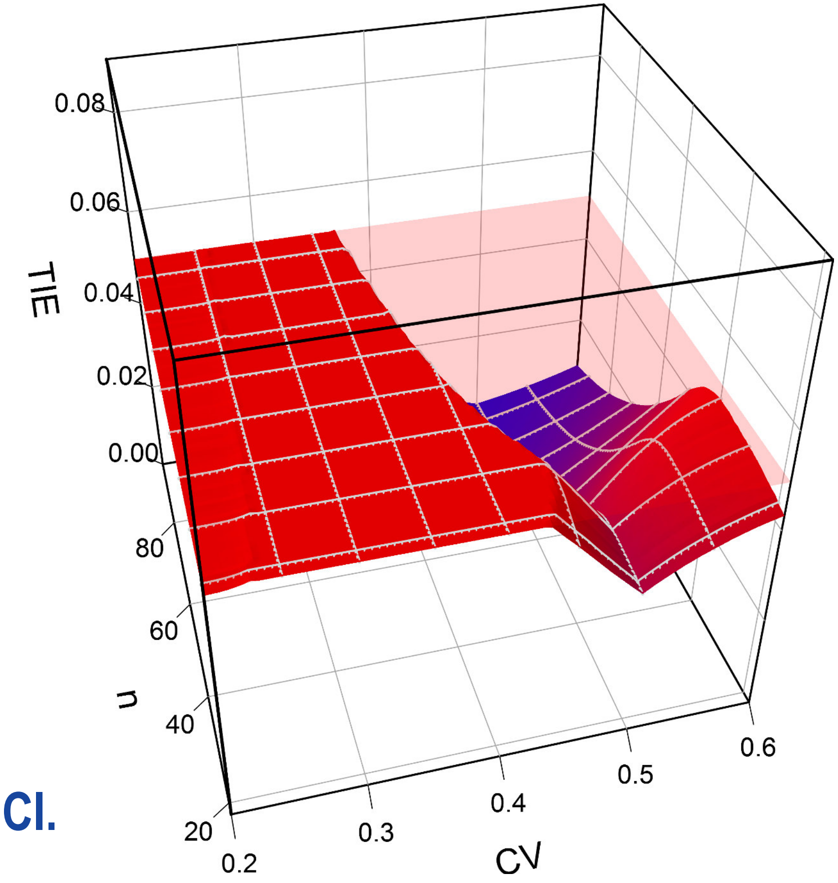
## Outlook.

- Utopia
  - Agencies collect $CV_{wR}$ from submitted studies. Pool them, adjust for designs / degrees of freedom. The EMA publishes 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 [*sic*] that *e.g.*, Bonferroni preserves the consumer risk. Still apply ABEL, but with a 95% CI ($\alpha$ 0.025).
  - Drawback: Loss of power, substantial increase in sample sizes.
- Proposal
  - Iteratively adjust $\alpha$ based on the study's $CV_{wR}$ and sample size – in such a way that the consumer risk is preserved.

# ABEL (iteratively adjusted α)

## Previous example

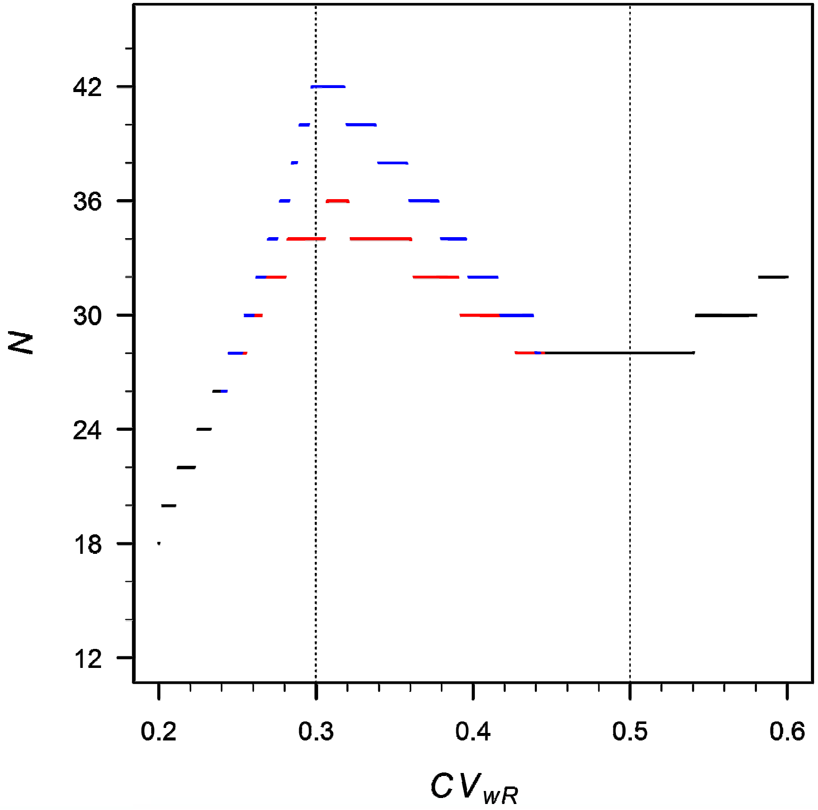
- Algorithm
  - Assess the TIE for the nominal $\alpha$ 0.05.
  - If the TIE $\leq$ 0.05, stop.
  - Otherwise adjust $\alpha$ (downwards) until the TIE $=$ 0.05.
  - At $CV_{wR}$ 30% (dependent on the sample size) $\alpha_{adj}$ is 0.0273 – 0.0300;
    → use a 94.00 – 94.54% CI.



# ABEL (iteratively adjusted *α*)

## Potential impact on the sample size.

- Example: RTRT | TRTR, $\theta_0$ 0.90, target power 0.80.
  - Moderate in the critical region (— —).
    - $CV_{wR}$ 30%: 36 → 42 (+17%);
    - $CV_{wR}$ 35%: 34 → 38 (+12%);
    - $CV_{wR}$ 40%: 30 → 32 ( +7%).
  - None outside (—).



# ABEL (iteratively adjusted *α*)

## Example (RTRT | TRTR, expected $CV_{wR}$ 35%, $\theta_0$ 0.90, target power 0.80); R package PowerTOST (≥1.3-3).

- **Estimate the sample size.**

```
sampleN.scABEL(CV=0.35, theta0=0.90, targetpower=0.80, design="2x2x4",
               details=FALSE, print=FALSE)[["Sample size"]]
[1] 34
```

- **Estimate the empiric TIE for this study.**

```
UL <- scABEL(CV=0.35)[["upper"]] # scaled limit (1.2948 for CVwR 0.35)
power.scABEL(CV=0.35, theta0=UL, n=34, design="2x2x4", nsims=1e6)
[1] 0.065566
```

- **Iteratively adjust $\alpha$.**

```
scABEL.ad(CV=0.35, n=34, design="2x2x4")
+++++++++++ scaled (widened) ABEL +++++++++++
         iteratively adjusted alpha
---------------------------------------------
CVwR 0.35, n(i) 17|17 (N 34)
Nominal alpha                 : 0.05
Null (true) ratio             : 0.9000
Regulatory settings           : EMA (ABEL)
Empiric TIE for alpha 0.0500  : 0.06557
Power for theta0 0.900        : 0.812
Iteratively adjusted alpha    : 0.03630
Empiric TIE for adjusted alpha: 0.05000
Power for theta0 0.900        : 0.773
```

# ABEL (iteratively adjusted *α*)

- **Optionally compensate for the loss in power (0.812 → 0.773) by increasing the sample size:**

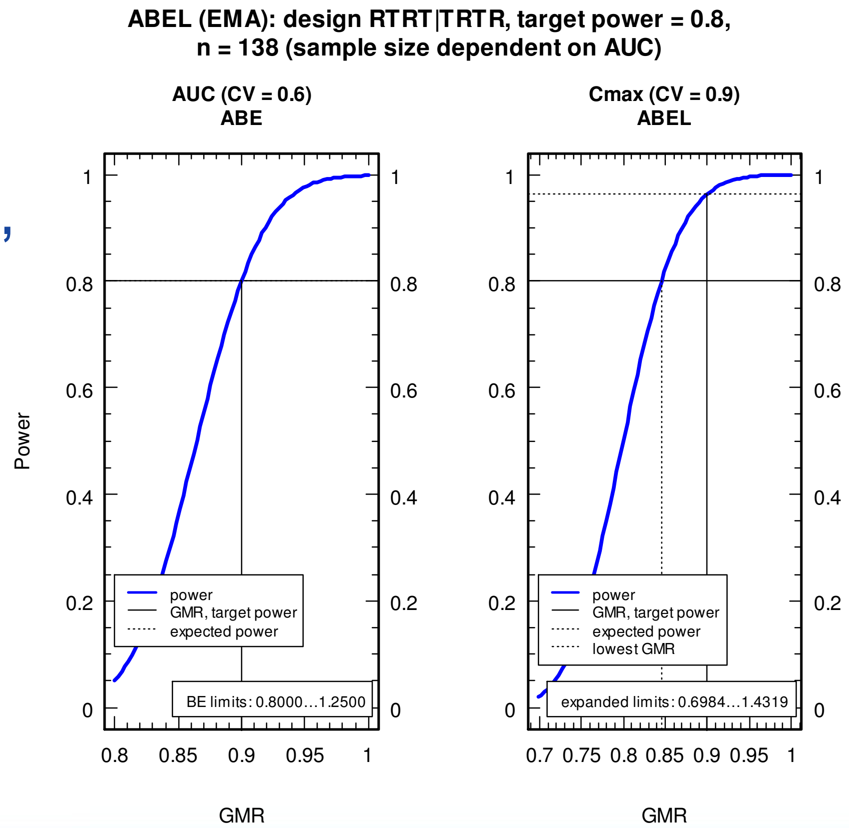
```
sampleN.scABEL.ad(CV=0.35, theta0=0.90, targetpower=0.80, design="2x2x4")
+++++++++++ scaled (widened) ABEL +++++++++++
            Sample size estimation
        for iteratively adjusted alpha
---------------------------------------------
Study design: 2x2x4 (RTRT|TRTR)
Expected CVwR 0.35
Nominal alpha      : 0.05
Null (true) ratio  : 0.9000
Target power       : 0.8
Regulatory settings: EMA (ABEL)
Switching CVwR     : 30%
Regulatory constant: 0.760
Expanded limits    : 0.7723...1.2948
Upper scaling cap  : CVwR 0.5
PE constraints     : 0.8000...1.2500
n  38,   adj. alpha: 0.03610 (power 0.8100), TIE: 0.05000
```

  - **n 34 → 38 (+12%), power 0.773 → 0.810, $\alpha_{adj}$ 0.0363 → 0.0361.**

# Excursion 2

## ‘Side effect’ of allowing ABEL only for $C_{max}$.

- Some drugs show high variability in *AUC* as well.
  - Since in such a case the sample size will be mandated by *AUC*, products with high deviations in $C_{max}$ will be approved.
  - Example: $CV_{wR}$ 90% ($C_{max}$), 60% (*AUC*), $\theta_0$ 0.90, target power 80% → the study is ‘overpowered’ for $C_{max}$; $C_{max}$-*GMR*s of [0.846–1.183] will pass BE. Really desirable?
  - With the FDA’s RSABE the study could be performed in only 34 subjects…



# Thank You!

## *Open Questions?*

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

# To bear in Remembrance...

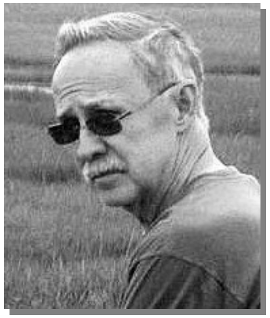
The fundamental cause of trouble in the world today
is that the stupid are cocksure
while the intelligent are full of doubt. Bertrand Russell

100% of all disasters are failures
of design, not analysis.

Ronald G. Marks

My definition of an expert in any field is a person
who knows enough about what's really going on
to be scared.

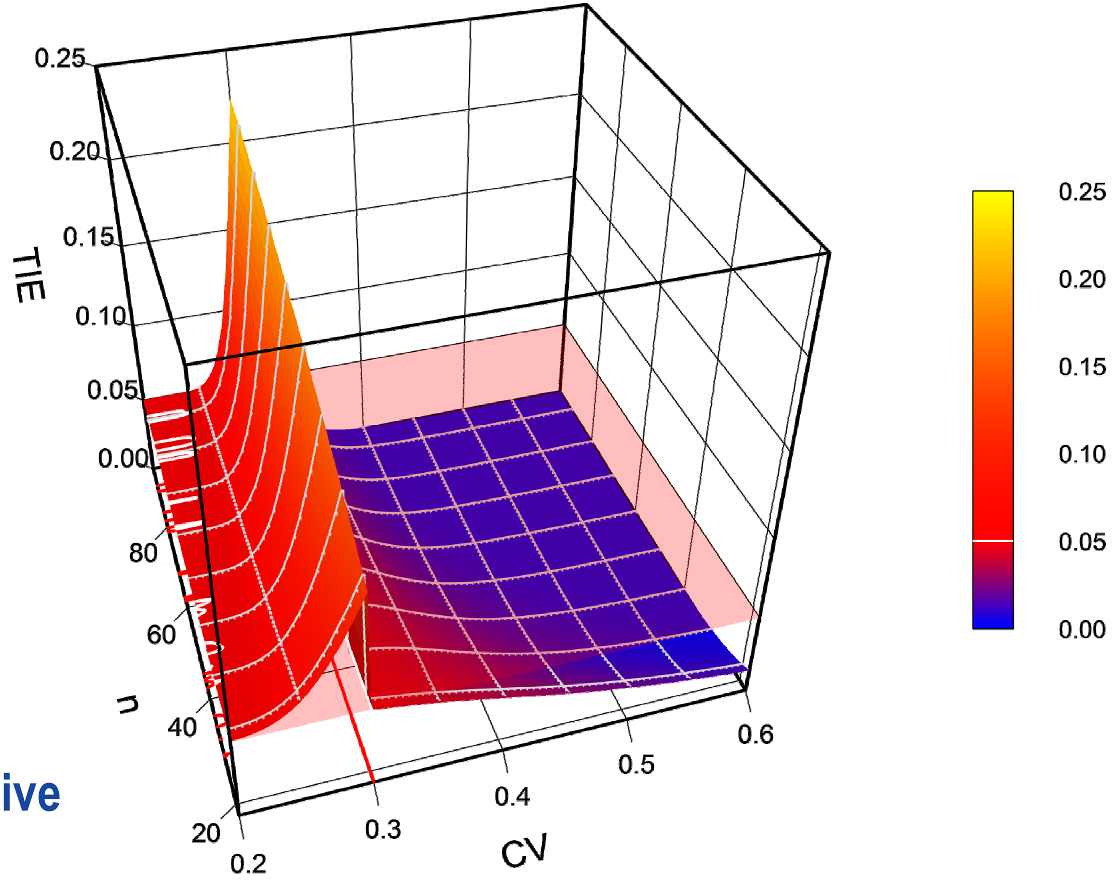
Phillip J. Plauger



# Backup

## Example for the FDA's RSABE

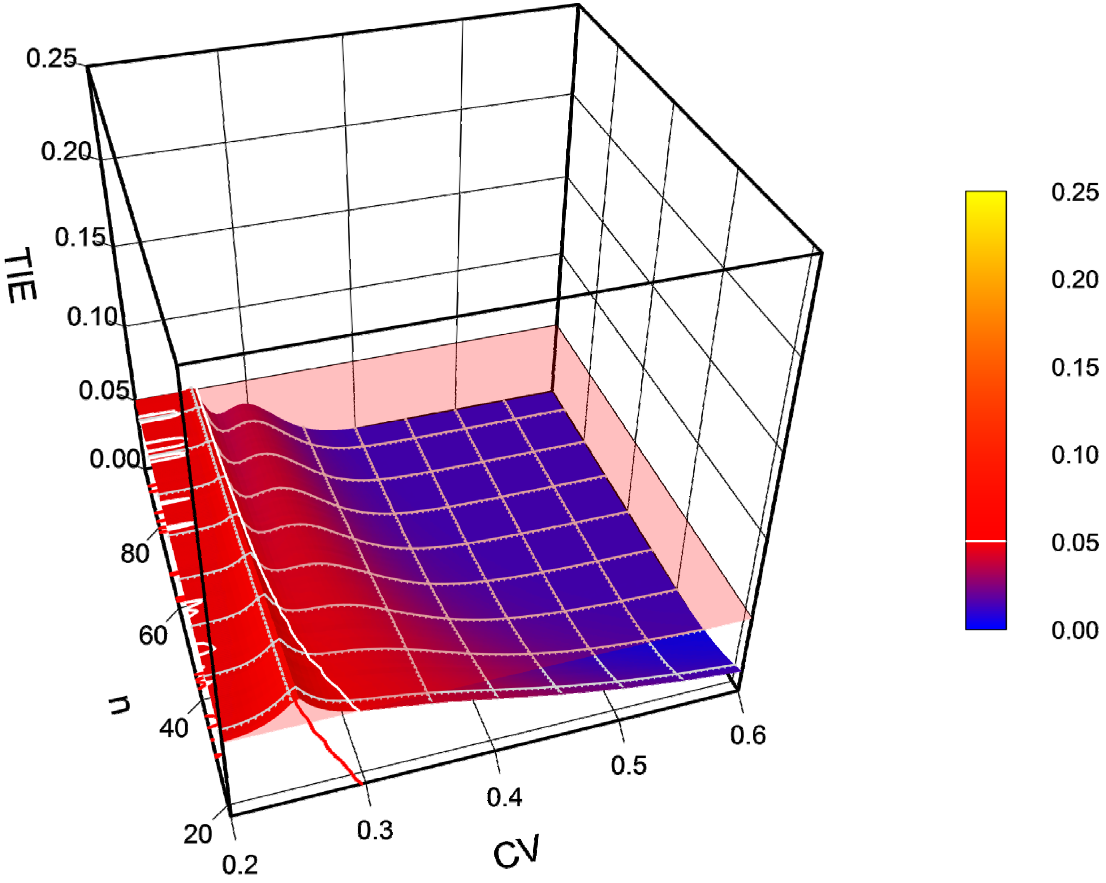
- RTRT | TRTR  
  sample size 18 – 96  
  $CV_{wR}$ 20% – 60%
  - TIE$_{max}$ 0.2245.
  - Relative increase of the consumer risk 349%!
  - TIE more dependent on the sample size than in ABEL.
  - However, no inflation of the TIE for $CV_{wR}$ >30%; RSABE is very conservative for 'true' HVD(P)s.



# Backup

## "FDA's desired consumer risk model" (Davit *et al*. 2012)

- Previous example
  - TIE assessed not at the scaled limits but
    - at 1.25 if $CV_{wR} \leq 25.4\%$ or
    - at $e^{0.893 \cdot \sigma_{wR}}$ otherwise.
  - $\text{TIE}_{\text{max}}$ 0.0668.
  - Lászlo Endrényi: *"Hocus pocus!"*



# References

Schuirmann DJ. *A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability*. J Pharmacokinet Biopharm. 1987; 15(6): 657–80.

Tóthfalusi L *et al*. *Evaluation of the Bioequivalence of Highly-Variable Drugs and Drug Products.* Pharm Res. 2001;18(6): 728–33.

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

Tóthfalusi L, Endrényi L. *Limits for the Scaled Average Bioequivalence of Highly Variable Drugs and Drug Products.* Pharm Res. 2003; 20(3): 382–9.

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

Wolfsegger MJ, Jaki T. *Simultaneous confidence intervals by iteratively adjusted alpha for relative effects in the one-way layout.* Stat Comput. 2006; 16(1): 15–23. DOI 10.1007/s11222-006-5197-1

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

Tóthfalusi L, Endrényi L, García-Arieta A. *Evaluation of Bioequivalence for Highly Variable Drugs with Scaled Average Bioequivalence.* Clin Pharmacokinet. 2009; 48(11): 725–43. DOI 10.2165/11318040-000000000-00000

European Medicines Agency, CHMP. *Guideline on the Investigation of Bioequivalence.* London; 2010 Jan 20. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/01/WC500070039.pdf

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

Davit BM *et al. Implementation of a Reference-Scaled Average Bioequivalence Approach for Highly Variable Generic Drug Products by the US Food and Drug Administration.* AAPS J. 2012; 14(4): 915–24. DOI 10.1208/s12248-012-9406-x

Patterson SD, Jones B. *Viewpoint: observations on scaled average bioequivalence.* Pharm Stat. 2012; 11(1): 1–7. DOI 10.1002/pst.498

European Medicines Agency, CHMP. *Questions & Answers: positions on specific questions addressed to the Pharmacokinetics Working Party (PKWP).* London; 2015 Nov 19. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002963.pdf

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

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 1.3-6. 2016. https://cran.r-project.org/package=PowerTOST

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. 2016; 35(12): 1933–43. 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.