Reference-scaling and Control of the Type I Error

Helmut Schütz

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!

Assumptions

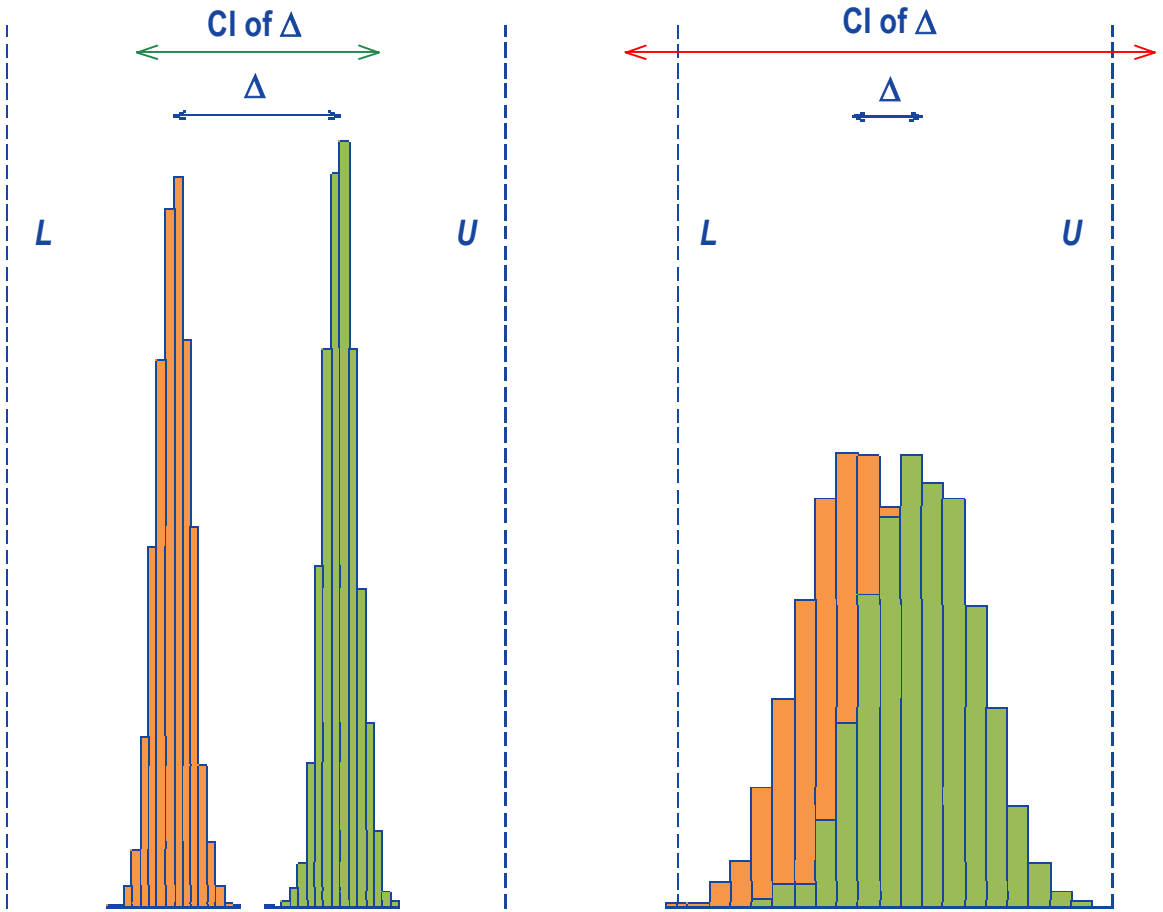
High variability can be

- an intrinsic property of the drug itself (low absorption and/or inter-occasion clearance) and/or
- attributed to the product's performance.
  - Physiology (enteric coated formulations and gastric emptying).
  - Absorption: rate of drug release and absorption window.
  - Influence of excipients
    - on gastric motility and/or
    - on transporters.

HVD

HVDP

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

Highly Variable Drugs / Drug Products

It may be almost impossible to demonstrate BE of HVD(P)s with a reasonable sample size

- **Example: *CV* 70%, *GMR* 0.90, target power 80%, 2×2×2 design**

```
library(PowerTOST)
sampleN.TOST(CV=0.7, theta0=0.9, targetpower=0.9, design="2x2x2")
++++++++++++ Equivalence test - TOST ++++++++++++
            Sample size estimation
-----------------------------------------------
Study design:  2x2 crossover
log-transformed data (multiplicative model)
alpha = 0.05, target power = 0.8
BE margins = 0.8 ... 1.25
True ratio = 0.9,  CV = 0.7
Sample size (total)
 n     power
358   0.801175
```

- **Since HVD(P)s are safe and efficacious some jurisdictions accept a larger 'not clinically relevant' difference**
  - **The BE limits can be *scaled* based on the variability of the reference.**

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 the 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: $C_{ss,min}$, $C_{ss,τ}$, partial *AUC*s
    - Health Canada: *AUC*

HVD(P)s – Reference-scaling

Models (in log-scale)

- **ABE Model:**
  - A difference Δ of $\leq$20% is considered to be clinically not relevant.
  - The limits [*L*, *U*] of the acceptance range are fixed to $\log(1-\Delta) = \log((1-\Delta)^{-1})$ 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 – 2α) 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).

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

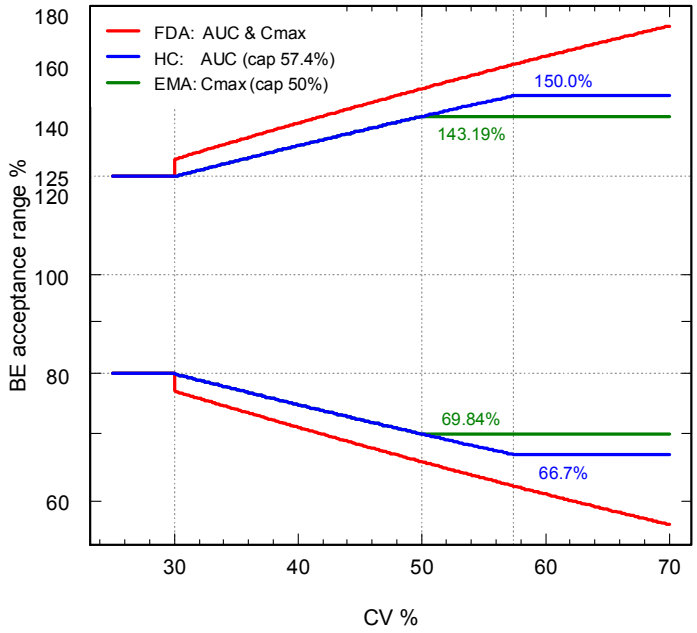
HVD(P)s – Reference-scaling

Regulatory Approaches

- Bioequivalence limits derived from σ_0 and σ_{wR}

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

- FDA
  - Scaling σ_{wR} 0.25 (θ_S 0.893) but applicable at CV_{wR} ≥30%.
  - Discontinuity at CV_{wR} 30%.
- EMA
  - Scaling σ_0 0.2936 (θ_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

Regulatory Approaches

- Scaled limits based on variability of the reference
  - EMA: IR C_{max} only; MR (additionally $C_{max,ss}$, $C_{min,ss}$, $C_{τ,ss}$, partial *AUC*s)
  - FDA: C_{max} and *AUC*
  - HC: *AUC* only

EMA	
CV_{wR}	BE limits (%)
≤30	80.00 – 125.00
35	77.23 – 129.48
40	74.62 – 134.02
45	72.15 – 138.59
≥50	69.84 – 143.19

FDA	
CV_{wR}	BE limits (%)
≤30	80.00 – 125.00
35	73.83 – 135.45
40	70.90 – 141.04
45	68.16 – 146.71
50	65.60 – 152.45
60	60.96 – 164.04
80	53.38 – 187.35
100	47.56 – 210.25

HC	
CV_{wR}	BE limits (%)
≤30	80.00 – 125.00
35	77.23 – 129.48
40	74.62 – 143.02
45	72.15 – 138.59
50	69.84 – 143.19
≥57.4	66.67 – 150.00

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 80.00 – 125.00%.
  - 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

- Pitfalls and suggestions
  - The applicant should justify that the calculated intra-subject variability is a reliable estimate and that it is not the result of outliers.
    - EMA Q&A-document (Rev. 7, March 2011), Data set I:
      RTRT | TRTR full replicate, 77 subjects, unbalanced, incomplete.
    - CV_{wR} 46.96% → apply ABEL (>30%)
    - Scaled acceptance range: 71.23 – 140.40%.
    - Method A: 90% CI 107.11 – 124.89% ⊂ AR; PE 115.66% ⊂ 80.00 – 125.00%.
    - Method B: 90% CI 107.17 – 124.97% ⊂ AR; PE 115.73% ⊂ 80.00 – 125.00%.
    - But there *are* two severe outliers!
      By excluding subjects 45 and 52, the CV_{wR} drops to 32.16%.
    - New scaled acceptance range: 78.79 – 126.93%.
      Almost no more gain compared to the conventional ABE limits.
    - Outliers have to be only excluded for the calculation of CV_{wR} but *kept* for the calculation of the CI.

HVD(P)s – Reference-scaling

The EMA's Approach

- Pitfalls and suggestions
  - Incomplete data (missing periods).
    - Even if one has no data of T (*e.g.*, a subject dropped out after the second period in sequence RRT) do not exclude the subject from the calculation of CV_{wR}. The estimate will be more accurate.
    - Must be unambigously stated in the protocol. Example for the partial replicate design (RRT|RTR|TRR):
      - » Data set for the estimation of CV_{wR}:
        All subjects with two administrations of R regardless of any other missing periods.
      - » Data set for the calculation of the CI:
        All subjects with at least one administration of T and at least one adminstration of R.

HVD(P)s – Reference-scaling

The EMA's Approach

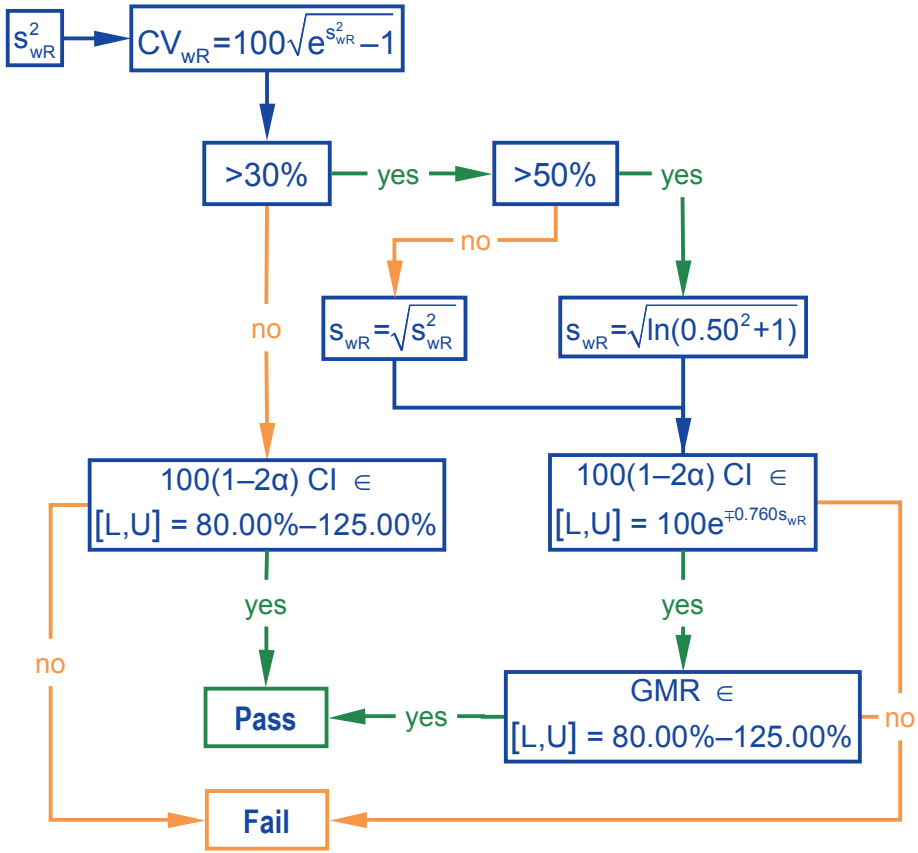
- Pitfalls and suggestions
  - ≥12 subjects in sequence RTR of the 3-period full replicate design (Q&A-document, Rev. 12 June 2015)
    - With sample sizes for the commonly applied T/R-ratio of 0.90 for HVD(P)s and ≥80% power this issue is practically not relevant.
    - Would affect only studies with extreme dropout-rates (>42%)!

CV_{wR} (%)	N	n_{RTR}	max. dropout-rate (%)
25	42	21	42.9
30	50	25	52.0
40	40	20	47.8
50	42	21	42.9
60	48	24	50.0
70	60	30	60.0
80	74	37	67.6

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

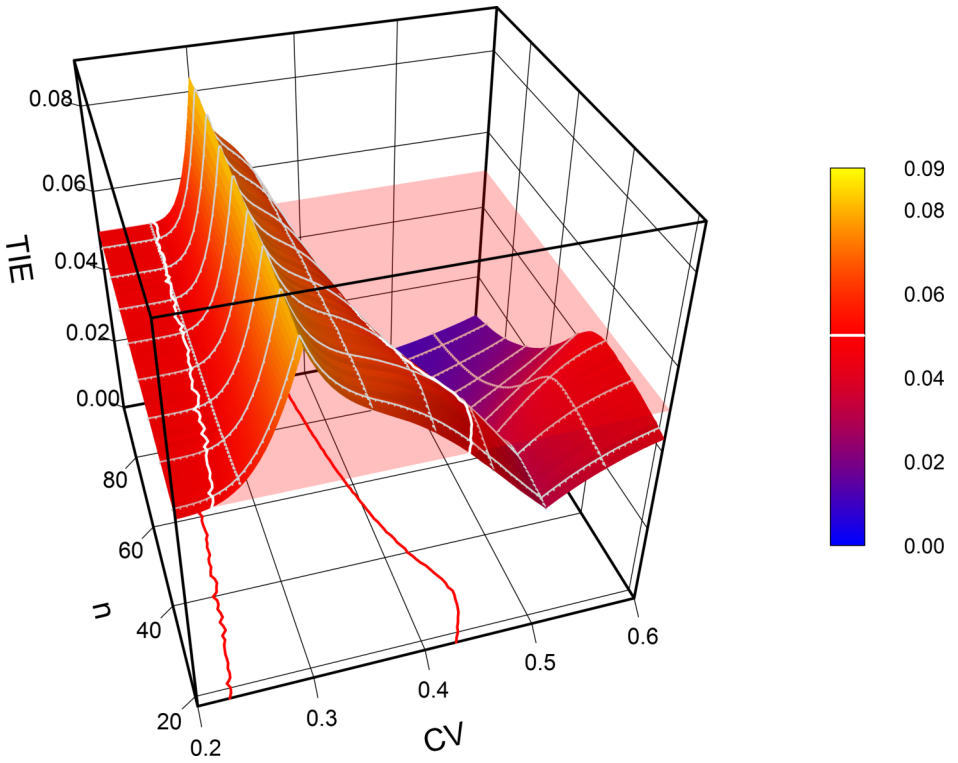
Assessing the Type I Error (TIE)

- 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, Patterson & Jones 2012).
- Confirmed.
  - EMA's ABEL: Tóthfalusi & Endrényi 2009, 2017, BEBA-Forum 2013, Wonnemann *et al*. 2015, Muñoz *et al*. 2016, Labes & Schütz 2016, Molins *et al*. 2017.
  - 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* …

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

… 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

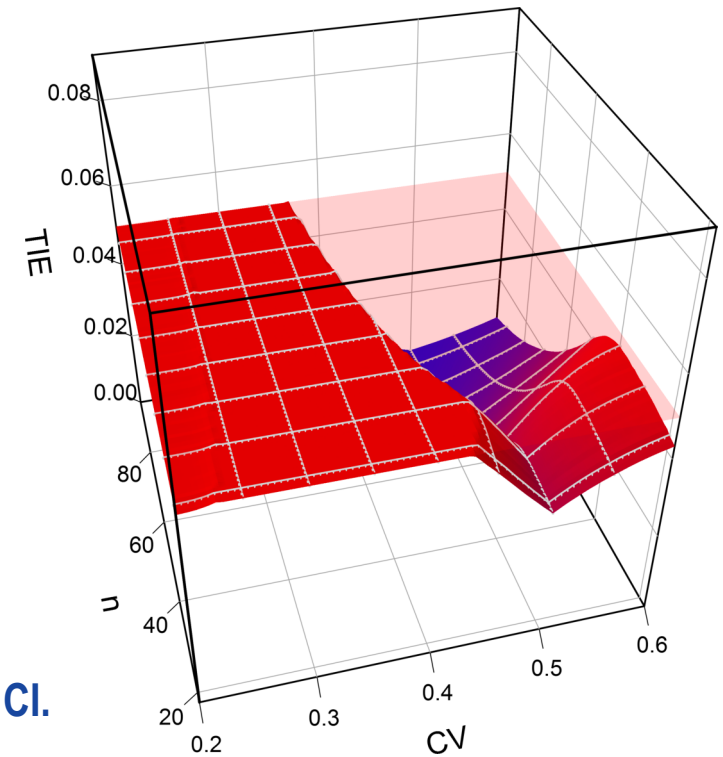
Outlook

- 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 (α 0.025).
  - Drawback: Loss of power, substantial 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 (Labes & Schütz 2016, Molins *et al.* 2017).

ABEL (iteratively adjusted α)

Previous example

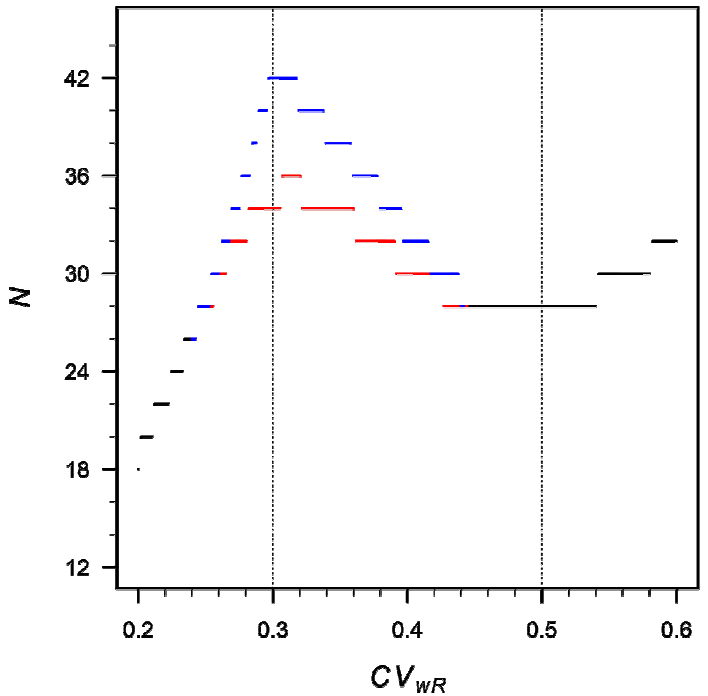
- Algorithm
  - Assess the TIE for the nominal α 0.05.
  - If the TIE $\leq$ 0.05, stop.
  - Otherwise adjust α (downwards) until the TIE = 0.05.
  - At CV_{wR} 30% (dependent on the sample size) α_{adj} is 0.0273 – 0.0300;
    $\rightarrow$ use a 94.00 – 94.54% CI.



ABEL (iteratively adjusted α)

Potential impact on the sample size

- Example: RTRT | TRTR, θ_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%, θ_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 α.**

```
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:**

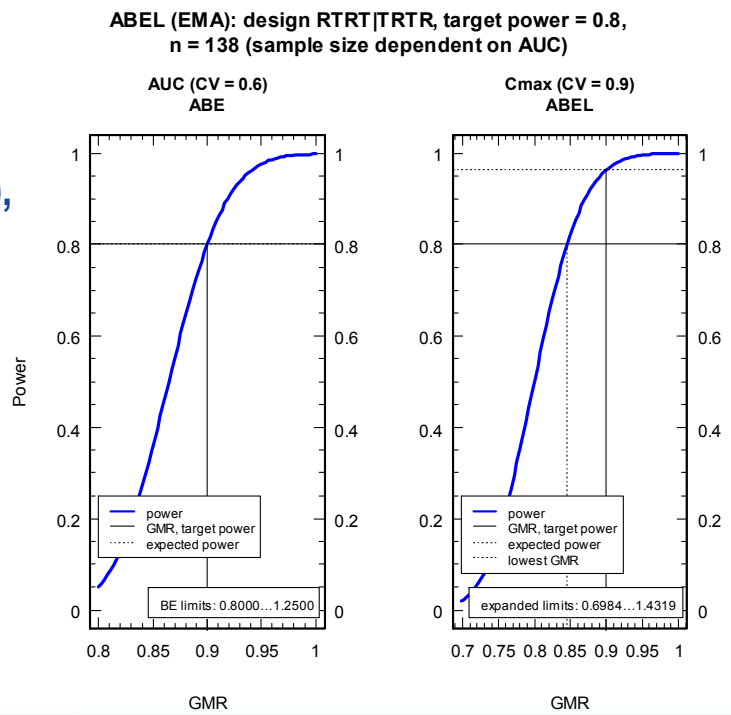
```
sampleN.scABEL.ad(CV=0.35, theta0=0.90, targetpower=0.80, design="2x2x4")
++++++++++++ scaled (widened) ABEL ++++++++++++
            Sample size estimation
        for iteratively adjusted alpha
--------------------------------------------
Study design: 2x2x4 (RTRT|TRTR)
Expected CVwR 0.35
Nominal 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, α_{adj} 0.0363 → 0.0361.**

Excursion 2

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

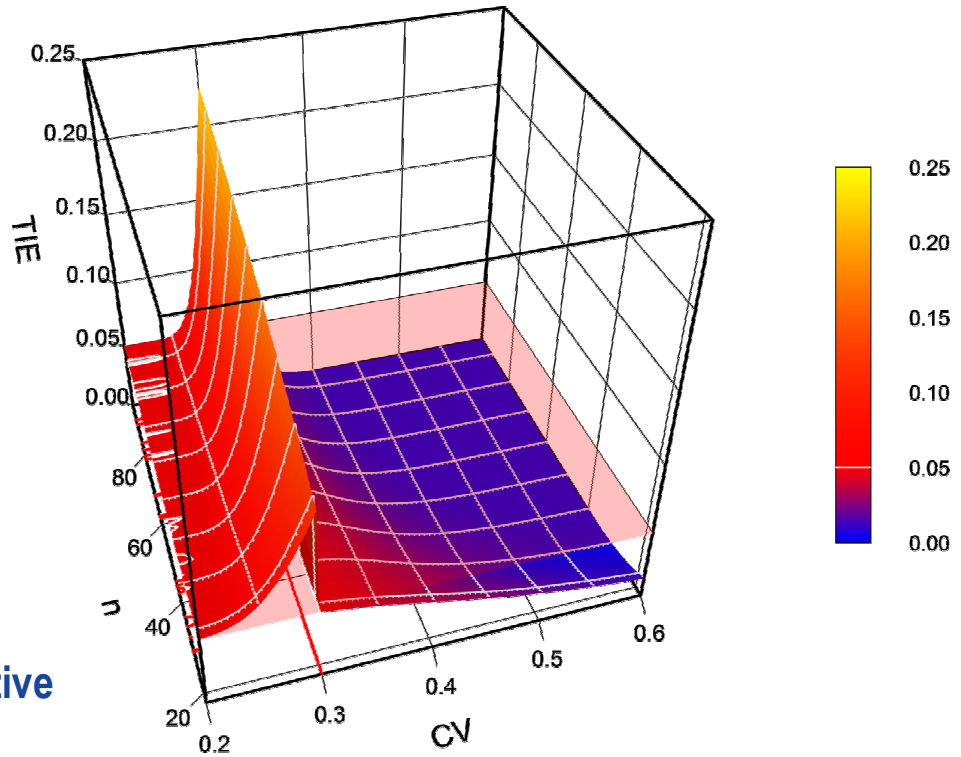
- Some drugs show high variability in *AUC* as well.
  - Since in such a case the sample size will be mandated by *AUC*, products with high deviations in C_{max} will be approved.
  - Example: CV_{wR} 90% (C_{max}), 60% (*AUC*), θ_0 0.90, target power 80% → the study is ‘overpowered’ for C_{max}; C_{max}-*GMR*s of [0.846–1.183] will pass BE. Really desirable?
  - With the FDA’s RSABE the study could be performed in only 34 subjects…



And on the other side of the pond?

Example for the FDA's RSABE

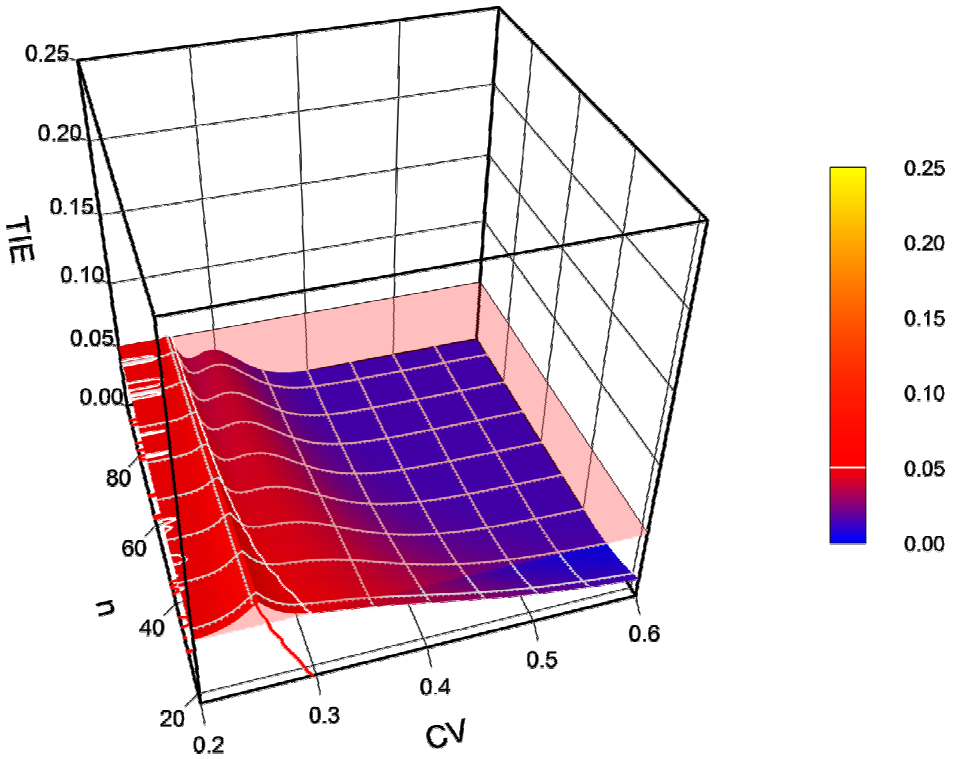
- RTRT | TRTR
  sample size 18 – 96
  CV_{wR} 20% – 60%
  - TIE_{max} 0.2245.
  - Relative increase of the consumer risk 349%!
  - TIE more dependent on the sample size than in ABEL.
  - However, no inflation of the TIE for CV_{wR} >30%; RSABE is very conservative for 'true' HVD(P)s.



And on the other side of the pond?

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.
  - TIE_{max} 0.0668.
  - Lászlo Endrényi: *"Hocus pocus!"*



Reference-scaling and Control of the Type I Error

Thank You!

Open Questions?

Helmut Schütz

BEBAC

Consultancy Services for
Bioequivalence and Bioavailability Studies
1070 Vienna, Austria

helmut.schuetz@bebac.at