

# Sample Size Estimation

Helmut Schütz

# 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 monozygotic twins or triplets in the study!

# Excursion: Type II Error

$\beta$ : Producer's risk to get no approval of an **equivalent** formulation ( $H_0$  *falsely* not rejected)

- Fixed in study planning to  $0.1 - \leq 0.2$  (10 -  $\leq 20\%$ ), where power =  $1 - \beta = \geq 80 - 90\%$

If all assumptions in sample size estimations turn out to be correct and power was set to 80%,

**one out of five studies will fail just by chance!**

$\alpha$ 0.05	BE
not BE	$\beta$ 0.20

← 0.20 = 1/5

- A *posteriori* (post hoc) power is irrelevant  
**Either** a study has demonstrated bioequivalence **or** not!  
 There is no need to 'justify' the sample size once the study was done!

# Review of Guidelines

## Minimum Sample Size.

- 12 WHO, EU, CAN, NZ, AUS, Brazil, AR, MZ, ASEAN States, RSA, Russia ('Red Book'), EAEU, Ukraine
- 12 USA *'A pilot study that documents BE can be appropriate, provided its design and execution are suitable and a sufficient number of subjects (e.g., 12) have completed the study.'*
- 18 Russia (2008)
- 20 RSA (MR formulations)
- 24 Saudia Arabia (12 to 24 if statistically justifiable)
- 24 USA (replicate designs intended for RSABE)
- 24 EU (RTR|TRT replicate designs intended for ABEL)
- 'Sufficient number' Japan
- 'Adequate' India

# Review of Guidelines

## Maximum sample size (pivotal study)

- Generally *not* specified (decided by IEC/IRB and/or local Authorities)
- ICH E9, Section 3.5 states:  
 The number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed.

## Sample size (pilot study)

- Is ICH E9 also applicable?
- If yes (likely), what is a 'reliable' answer?

# Power vs. Sample Size

It is not possible to *directly* obtain the required sample size

- 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 ( $\alpha$ );
  - the expected deviation ( $\Delta$ ) from the reference product and;
  - the desired power ( $1 - \beta$ ).
- Three values are *known and fixed* (AR,  $\alpha$ ,  $1 - \beta$ ), one is an *assumption* ( $\Delta$ ), and one an *estimate* ( $s^2$ ).
  - Hence, the correct term is ‘sample size *estimation*’
  - and not ‘sample size *calculation*’

# Power vs. Sample Size

## Only power is accessible

- The sample size is searched in an iterative procedure until at least the desired power is obtained

Example:  $\alpha$  0.05, target power 80% ( $\beta$  0.2),  
 expected  $GMR$  0.95,  $CV_{intra}$  20%  $\rightarrow$   
 minimum sample size 19 (power 81.3%),  
 rounded *up* to the next even number in  
 a  $2 \times 2 \times 2$  study (power 83.5%)

$n$	power (%)
16	73.5
17	76.4
18	79.1
19	81.3
20	83.5

- Exact methods for ABE in parallel, crossover, and replicate designs are available
- Simulations suggested for Group-Sequential and Two-Stage Designs
- Simulations mandatory for reference-scaling methods

# Power vs. Sample Size

## Which sample size is ‘large enough’?

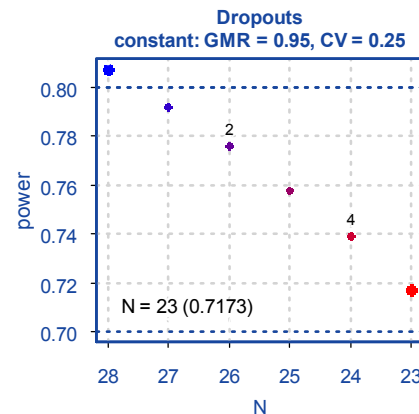
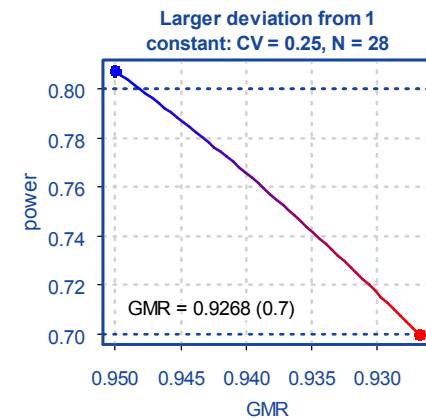
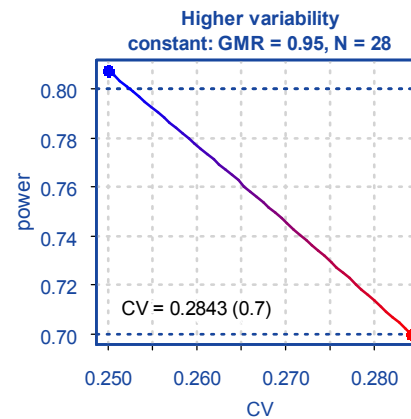
- Most guidelines recommend 80 – 90% power for pivotal studies
  - EMA Appropriate sample size calculation [*sic*].  
Sample size depends on  $\alpha$  (fixed), BE-limits (fixed),  $\Delta$  (assumed), and desired power
  - If a study is planned for  $\leq 70\%$  power, problems with the ethics committee are possible (ICH E9)
  - If a study is planned for  $> 90\%$  power (especially for drugs with low variability), additional problems with regulators are possible (‘forced bioequivalence’)
  - Some subjects (‘alternates’) may be added to the estimated sample size according to the expected dropout-rate – especially for studies with more than two periods or multiple-dose studies
- According to ICH E9 a sensitivity analysis is mandatory to explore the impact on power if values deviate from assumptions



# Power Analysis

## Example 2×2×2, ABE

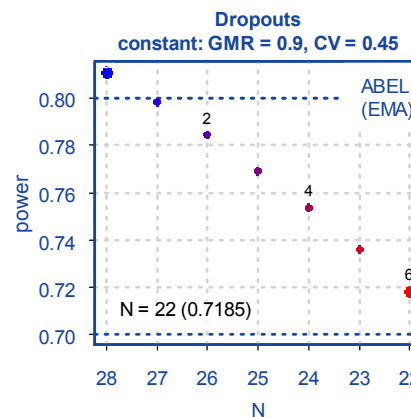
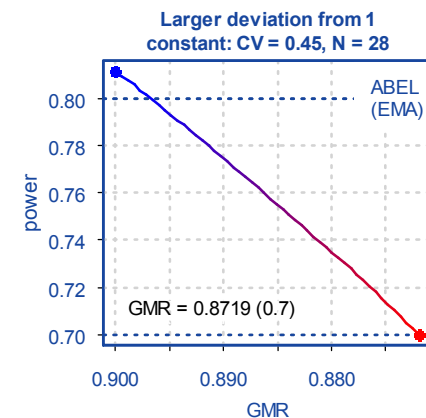
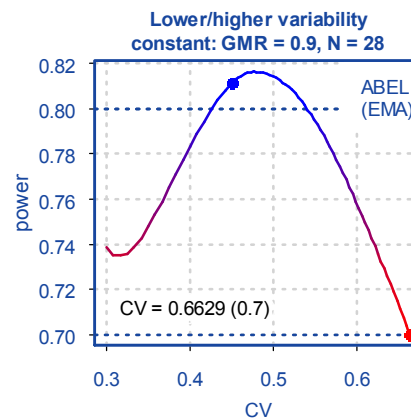
- Assumed *GMR* 0.95,  $CV_w$  0.25, desired power 0.8, min. acceptable power 0.7.
  - Sample size 28 (power 0.807)
  - $CV_w$  can increase to 0.284 (rel. +14%)
  - GMR* can decrease to 0.927 (rel. -2.4%)
  - 5 drop-outs acceptable (rel. -18%)
  - Most critical is the *GMR*!



# Power Analysis

## Example 2×2×4, ABEL

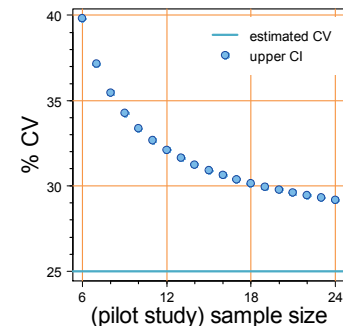
- Assumed *GMR* 0.90,  
 $CV_{WR}$  0.45, desired power 0.8,  
min. acceptable power 0.7.
  - Sample size 28 (power 0.811)
  - $CV_w$  can increase to 0.663  
(rel. +47%)
  - GMR* can decrease to 0.872  
(rel. -3.1%)
  - 6 drop-outs acceptable  
(rel. -21%)
  - Most critical is the *GMR*!



# Dealing with Uncertainty

## Nothing is 'carved in stone'

- Never assume perfectly matching products
  - Generally a  $\Delta$  of *not* better than 5% should be assumed (GMR 0.9500 – 1.0526)
  - For HVD(P)s do not assume a  $\Delta$  of <10% (GMR 0.9000 – 1.1111)
- 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 the lower or upper CL
- Precision of estimates
  - Improves with  $n^2$
  - In order to double the precision one has to quadruple the sample size



# Problems

## The EMA's 'appropriate sample size calculation'

- The purpose of a pilot study (amongst others) is to obtain estimates of the *GMR* and *CV* which can be used to design the pivotal study
- In a strict sense it is not possible to demonstrate bioequivalence in a pilot study which is – by definition – exploratory
- However, in the past some agencies (Scandinavian countries, Germany) accepted pilot studies as evidence of BE if stated as such in the protocol
  - Repeating a passing pilot (even in a larger sample size) may fail by pure chance (producer's risk =  $1 - \text{power}$ )
  - Hence, this approach was considered unethical
- Nowadays, European regulatory agencies are seemingly more strict (follow the 'cook book')

Still acceptable for the FDA...

# Excursion

## Type I Error

- In BE the Null Hypothesis ( $H_0$ ) is *inequivalence*
  - TIE = Probability of falsely rejecting  $H_0$  (i.e., accepting  $H_a$  and claiming BE)
  - Can be calculated for the nominal significance level ( $\alpha$ ) assuming a *GMR* ( $\theta_0$ ) at one of the limits of the acceptance range  $[\theta_1, \theta_2]$

- Example: 2x2x2 crossover, CV 20%,  $n$  20,  $\alpha$  0.05,  $\theta_0 = [\theta_1 \text{ 0.80 or } \theta_2 \text{ 1.25}]$

```
library(PowerTOST)
```

```
AR <- 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=AR[1])
```

```
[1] 0.0499999
```

```
power.TOST(CV=0.20, n=20, alpha=0.05, theta0=AR[2])
```

```
[1] 0.0499999
```

- TOST is not a uniformly most powerful (UMP) test

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

```
[1] 0.04976374
```

- However, the TIE never exceeds the nominal level

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

```
[1] 0.05
```

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 1.4-7. 2018. <https://cran.r-project.org/package=PowerTOST>

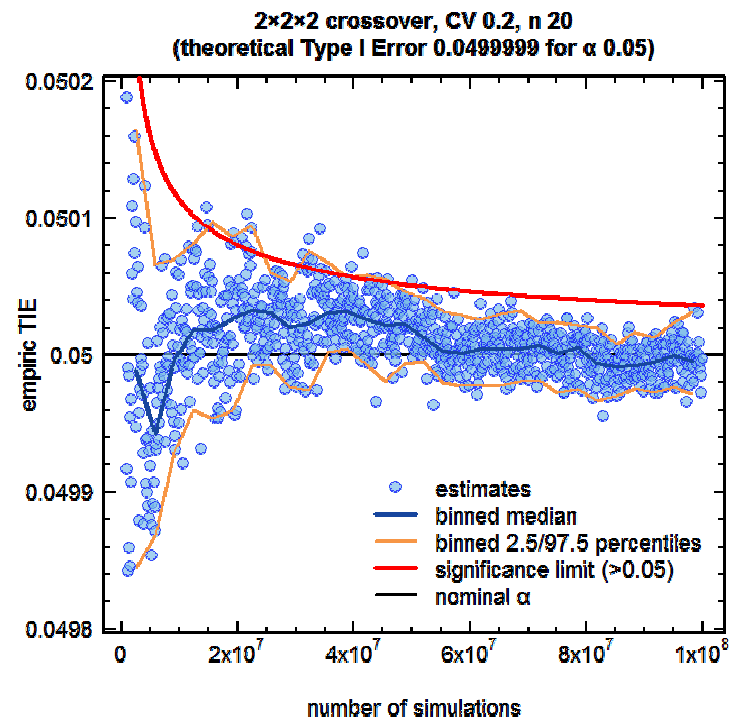
# Excursion

## Type I Error

- Alternatively perform simulations to obtain an *empiric* Type I Error  
`power.TOST.sim(CV=0.20, n=20, alpha=0.05, theta0=AR[2], nsims=1e8)`

[1] 0.04999703

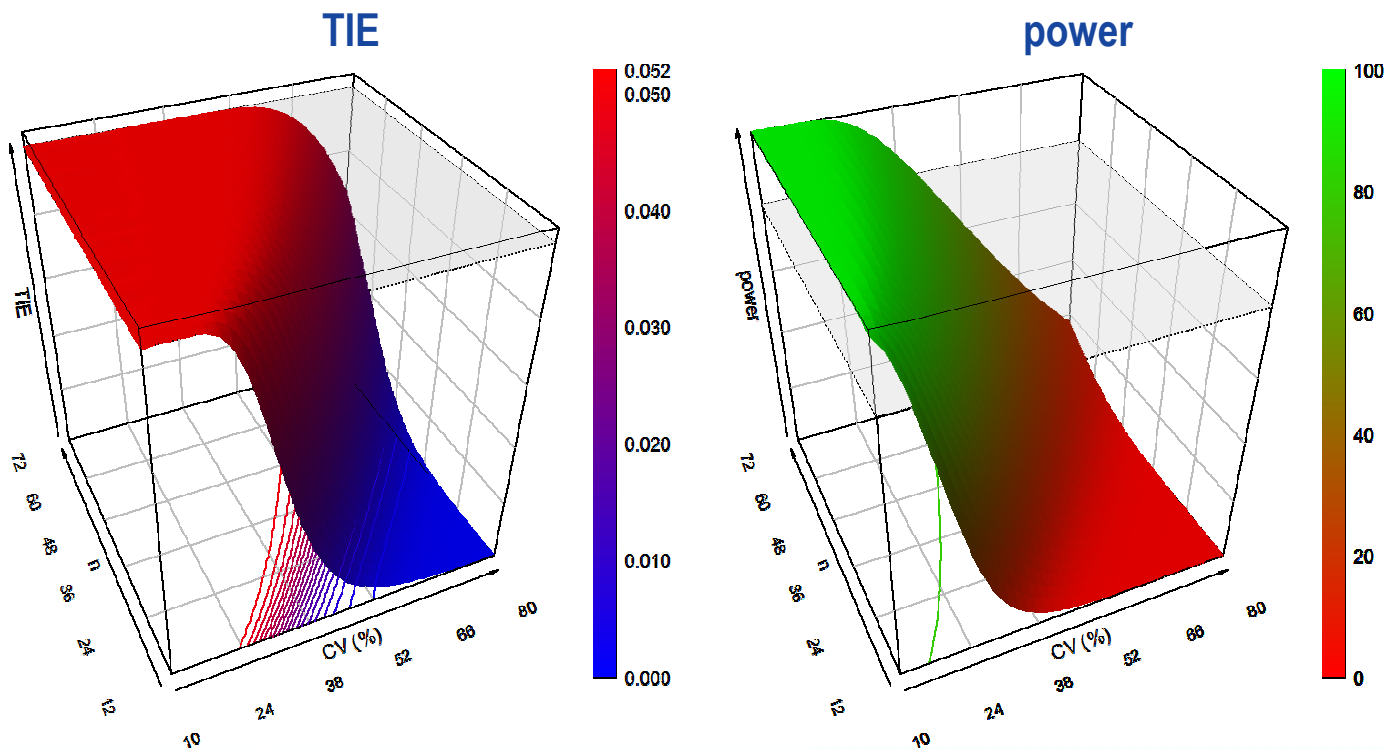
- In other settings (*i.e.*, frameworks like Two-Stage Designs or reference-scaled ABE) analytical solutions for power – and therefore, the TIE – are not possible:  
 Simulations are required.



# 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



# R Package PowerTOST

## Examples

- Install the package from CRAN if necessary and attach it
 

```
if (!("PowerTOST" %in% installed.packages()[, "Package"])) {
  install.packages("PowerTOST")
}
library(PowerTOST)
```
- **ABE**
  - **2×2×2 crossover,  $CV_{intra}$  25%,  $\theta_0$  0.95, targetpower 90%.**

```
sampleN.TOST(CV=0.25, theta0=0.95, targetpower=0.9,
             print=FALSE)[["Sample size"]]
[1] 38
```
  - **2×2×2 crossover,  $CV_{intra}$  10%, NTID (AR 90.00–111.11%),  $\theta_0$  0.95.**

```
sampleN.TOST(CV=0.10, theta0=0.95, theta1=0.9,
             print=FALSE)[["Sample size"]]
[1] 44
```
  - **Parallel design,  $CV_{total}$  40%,  $\theta_0$  0.95.**

```
sampleN.TOST(CV=0.20, theta0=0.95, design="parallel",
             print=FALSE)[["Sample size"]]
[1] 130
```



# R Package PowerTOST

- ABEL (reference-scaling according to the EMA)

- 4-period full replicate,  $CV_{WR}$  35%,  $\theta_0$  0.90.

```
sampleN.scABEL(CV=0.35, theta0=0.90, design="2x2x4", details=TRUE)
```

```
+++++++ scaled (widened) ABEL ++++++
          Sample size estimation
          (simulation based on ANOVA evaluation)
```

```
-----
Study design: 2x2x4 (full replicate)
```

```
alpha = 0.05, target power = 0.8
```

```
CVw(T) = 0.35; CVw(R) = 0.35
```

```
True ratio = 0.9
```

```
ABE limits / PE constraint = 0.8 ... 1.25
```

```
EMA regulatory settings
```

```
- CVswitch = 0.3
```

```
- cap on scABEL if CVw(R) > 0.5
```

```
- regulatory constant = 0.76
```

```
- pe constraint applied
```

```
Sample size search
```

```
  n   power
30  0.7702
32  0.7929
34  0.8118
```

# R Package PowerTOST

- ABEL (reference-scaling according to the EMA, iteratively adjusted  $\alpha$  to preserve the consumer risk at  $\leq 0.05$ : Labes and Schütz 2016)

- 4-period full replicate,  $CV_{WR}$  35%,  $\theta_0$  0.90.

```
sampleN.scABEL.ad(CV=0.35, theta0=0.90, design="2x2x4", details=TRUE)
```

```
+++++++ scaled (widened) ABEL ++++++
```

```
Sample size estimation
```

```
for iteratively adjusted alpha'
```

```
-----  
Study design: 2x2x4 (RTRT|TRTR)
```

```
Expected CVWR 0.35
```

```
Nominal alpha      : 0.05
```

```
True ratio         : 0.9000
```

```
Target power      : 0.8
```

```
Regulatory settings: EMA (ABEL)
```

```
Switching CVWR    : 0.3
```

```
Regulatory constant: 0.76
```

```
Expanded limits    : 0.7723...1.2948
```

```
Upper scaling cap  : CVWR > 0.5
```

```
PE constraints     : 0.8000 ... 1.2500
```

```
n 34, nomin. alpha: 0.05000 (power 0.8118), TIE: 0.0656
```

```
n 34,  adj. alpha: 0.03630 (power 0.7728)
```

```
n 38,  adj. alpha: 0.03610 (power 0.8100), TIE: 0.05000
```

# Remedies, Outlook

## Pilot study

- For applicants
  - Sample size as large as the budget allows
    - Increases the precision of estimates
    - Adjusting for the uncertainty of the *GMR* (even with a Bayesian method) leads to sample sizes of the pivotal study which likely are not feasible
    - Take all available information about the *GMR* into account (e.g., from *IVIVC*) but always allow for a safety margin (don't be overly optimistic)
  - For ABE consider a Two-Stage Design
    - Adjusts the sample size based on the *CV* observed in the first stage
    - Adjusting for the observed *GMR* is generally not possible (compromises power)
    - Include a futility criterion for early stopping

# Remedies, Outlook

## Pilot study

- For applicants
  - Reference-scaling (ABEL)
    - If the expected  $CV_{wR}$  is within 30 – 50% and the actual  $CV_{wR}$  is larger, power increases (more expansion of limits)
    - Some companies have a policy for pilot studies:  
Full replicate, 36 subjects
    - Even if the pivotal study is planned as a *partial* replicate design (TRR|RTR|RRT), perform the pilot in a *full* replicate in order to additionally estimate  $CV_{wT}$   
If  $CV_{wT} < CV_{wR}$  there will be incentive in the sample size  
Example
      - »  $CV_{wT}$  35%,  $CV_{wR}$  50% observed in the full replicate pilot.  
Sample size for a partial replicate design **33**
      - » If the pilot was performed in a partial replicate (no information about  $CV_{wT}$ ) one has to *assume* that  $CV_{wT} = CV_{wR}$   
Sample size for a partial replicate design **39**

# Remedies, Outlook

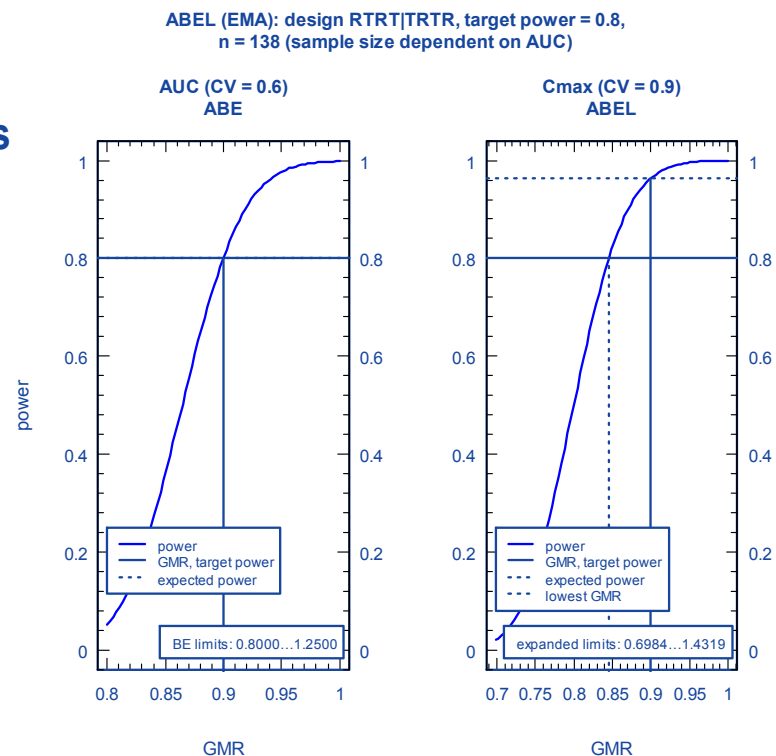
## Pilot study

- For applicants
  - Demonstrating bioequivalence in the pilot
    - State the intention unambiguously in the protocol
    - Give a justification and concentrate on ethics rather than economics
    - Consider a scientific advice in a 'difficult' member state (e.g., Spain, The Netherlands, France)

# Remedies, Outlook

## Pivotal study

- For applicants
  - The EMA's approach of allowing reference-scaling only for  $C_{max}$  has the side effect of accepting products which large deviations if  $AUC$  is highly variable as well
  - The sample size depends on the variability of  $AUC$  which has to be assessed by ABE. Example:
    - » Target power 80%,  $GMR$  0.9 (both PK metrics),  $CV_{wT} = CV_{wR}$  0.6 ( $AUC$ ), 0.9 ( $C_{max}$ )
    - » With 138 subjects required for  $AUC$ , products with a  $GMR$  of 0.846 of  $C_{max}$  will pass ABEL



# Remedies, Outlook

## Pilot study

- Regulatory agencies

- should reconsider accepting BE demonstrated in a pilot study

- Example

- » Pilot:  $n$  24,  $GMR$  0.95,  $CV_w$  0.25, 90% CI 81.98 – 110.09%

- » Pivotal:  $n$  28, power 80.7% (*i.e.*, risk of failure 19.3%)

- Elastic clause in the BE GL (4.1.8 Evaluation – Presentation of data)

If [...] multiple studies have been performed some of which demonstrate BE and some of which do not, the body of evidence must be considered as a whole. Only relevant studies [...] need be considered. The existence of a study which demonstrates BE does not mean that those which do not can be ignored. The applicant should thoroughly discuss the results and justify the claim that BE has been demonstrated. Alternatively, when relevant, a combined analysis of all studies can be provided in addition to the individual study analyses. It is not acceptable to pool together studies which fail to demonstrate BE in the absence of a study that does.

# Remedies, Outlook

## Pivotal study

- Regulatory agencies
  - should reconsider accepting reference-scaling also for *AUC*
    - Was discussed in the Concept Paper 2006 (removed from the EMA's website; available at: <http://bebac.at/downloads/14723106en.pdf>) and the 2<sup>nd</sup> / 3<sup>rd</sup> International Conferences of the Global Bioequivalence Harmonization Initiative (Rockville, September 2016 / Amsterdam, April 2018)
    - RSABE acceptable for the FDA
    - ABEL acceptable for Health Canada (expanded limits up to 66.67 – 150.00%)
    - In June 2017 the WHO opened in pilot phase allowing scaling for *AUC* on a case-by-case basis
      - » 4-period full replicate design mandatory  
'in order to assess the variability associated with each product'
    - Current practice leads to approval of products with large  $\Delta$  in  $C_{max}$   
Although technically valid, is this really desirable?



# Sample Size Estimation

**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)