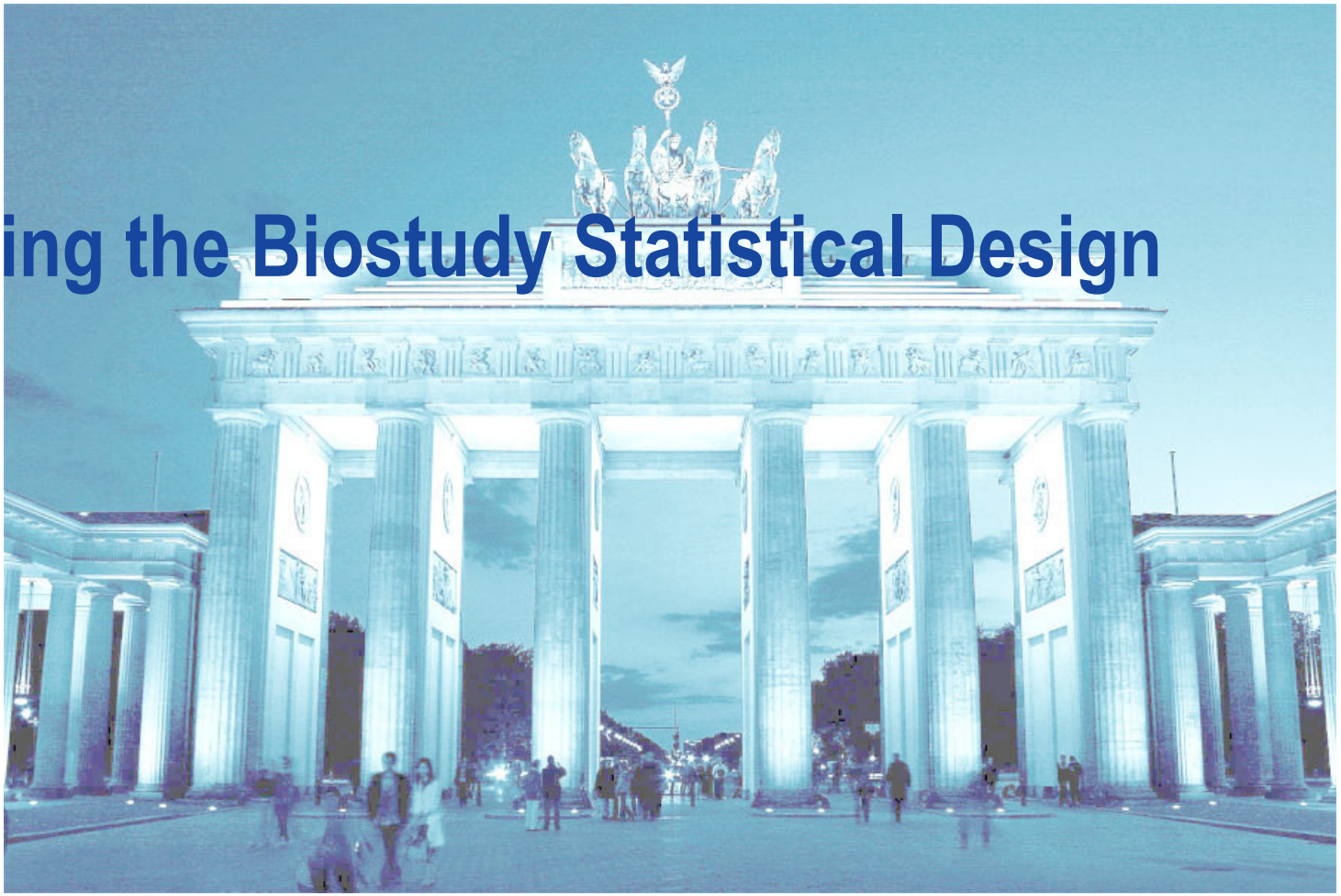
Establishing the Biostudy Statistical Design

Helmut Schütz



Excursion: Assumptions in statistics

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!

Excursion: Error(s)

All *formal* decisions are subjected to two 'Types' of Error.

- α: Probability of Type I Error (aka Risk Type I)
- β: Probability of Type II Error (aka Risk Type II)

Example from the justice system – which presumes that the defendant is *not guilty*:

Verdict	Defendant *innocent*	Defendant *guilty*
Presumption of innocence *rejected* (*guilty*)	wrong	correct
Presumption of innocence *accepted* (*not guilty*)	correct	wrong

Excursion: Hypotheses

In statistical terminology

- Null hypothesis (H_0): innocent
- Alternative hypothesis (H_a aka H_1): guilty

Decision	Null hypothesis *true*	Null hypothesis *false*
H_0 rejected	Type I Error	Correct (accept H_a)
Failed to reject H_0	Correct (accept H_0)	Type II Error

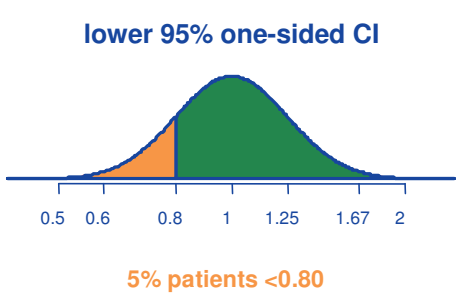
In BE the Null hypothesis is bioinequivalence ($\mu_T \neq \mu_R$)!

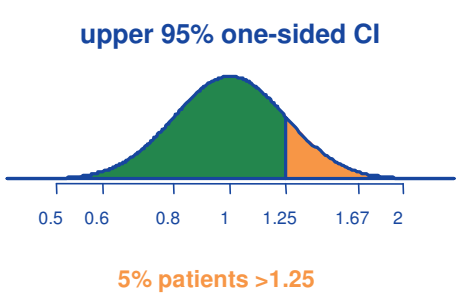
Decision	Null hypothesis *true*	Null hypothesis *false*
H_0 rejected	Patient's risk (α)	Correct (BE)
Failed to reject H_0	Correct (not BE)	Producer's risk (β)

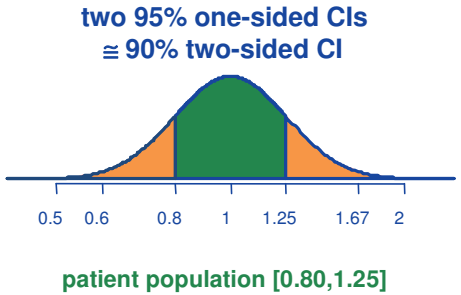
Excursion: Type I Error

α: Patient's risk to be treated with an inequivalent formulation (H_0 *falsely* rejected)

- BA of the test compared to reference in a *particular* patient is considered to be risky *either* below 0.80 *or* above 1.25.
  - If we keep the risk of *particular* patients at α 0.05 (5%), the risk of the entire *population* of patients (where BA <0.80 *and* >1.25) is 2α (10%) – expressed as a confidence interval: $100(1 - 2\alpha) = 90\%$.
  - However, since in a patient BA cannot be <0.80 *and* >1.25 *at the same time*, the patient's risk from a 90% CI is still 5%!







Excursion: Type II Error

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

- **Fixed in study planning to 0.1 – ≤0.2 (10 – ≤20%), where power = 1 – β = ≥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!

α 0.05	BE
not BE	β 0.20

0.20 = 1/5

- ***A posteriori* (*post hoc*) power is irrelevant!**

  Either a study has demonstrated bioequivalence or not.

Review of Guidelines

Minimum sample size.

- 12 WHO, EU, CAN, USA, AUS, NZ, AR, MZ, ASEAN States, RSA, Russia ('Red Book'), EAEU, Ukraine.
  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), Brazil, USA (replicate designs intended for RSABE), EU (RTR|TRT replicate designs intended for ABEL).
- 'Sufficient number' Japan, 'adequate' India.

Maximum sample size.

- Generally *not* specified (decided by IEC/IRB and/or local Authorities).

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 (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*'.**

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: α 0.05, target power 80% (β 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
  2×2×2 study (power 83.5%).

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

  - Exact methods for average bioequiivalence (ABE) in parallel, crossover, and replicate designs are available.
  - Simulations suggested for Group-Sequential and Two-Stage Designs (GSD, TSD).
  - Simulations mandatory: Reference-scaled average bioequivalence (FDA: RSABE), average bioequivalence with expanding limits (EMA: ABEL).

Notation of cross-over designs: treatments × sequences × periods

Power *vs.* Sample Size

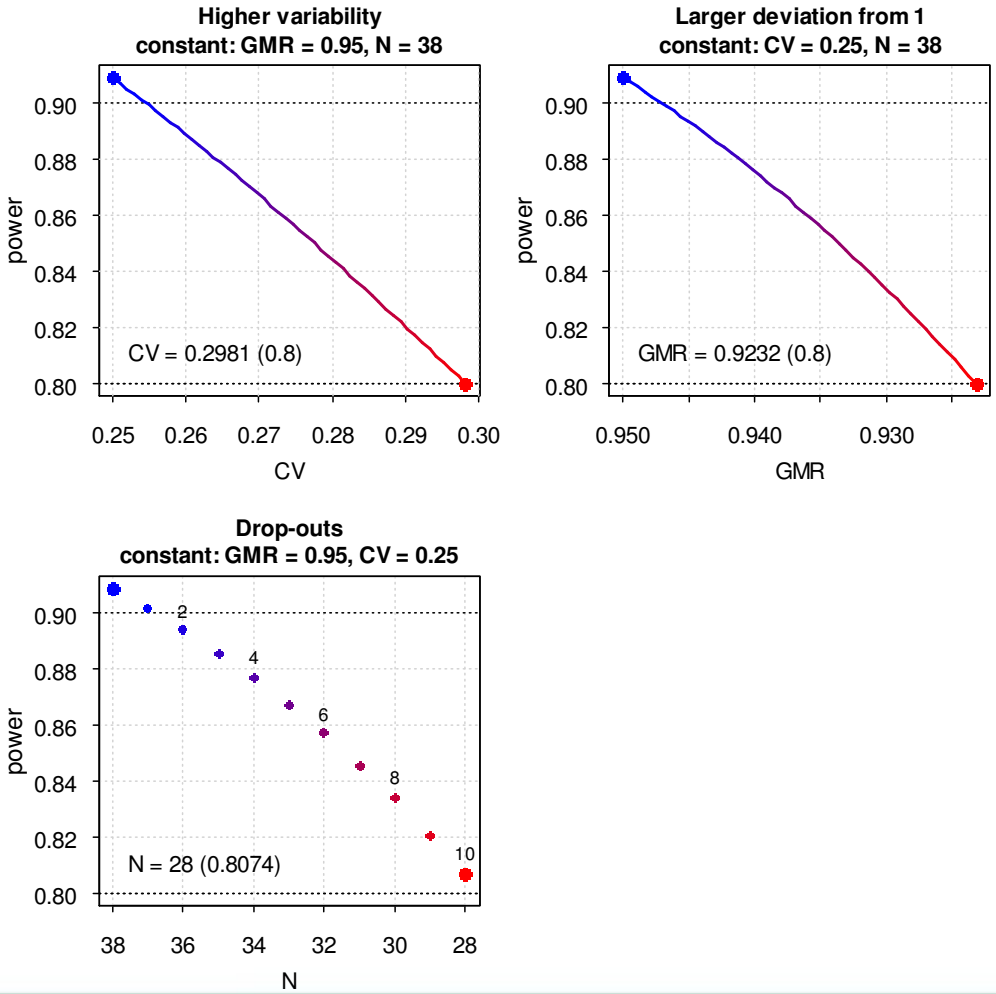
How many subjects are ‘enough’?

- Most guidelines recommend 80 – 90% power.
  - If a study is planned for ≤70% power, problems with the ethics committee are possible (ICH E9).
  - If a study is planned for >90% power (especially with low variability drugs), additional problems with regulators are possible (‘forced bioequivalence’).
  - Some subjects (‘alternates’) may be added to the estimated sample size according to the expected drop-out 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

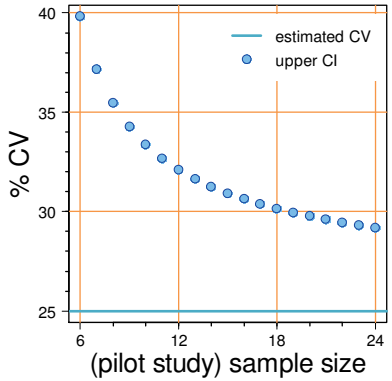
- Assumed *GMR* 0.95, CV_w 0.25, desired power 0.9, min. acceptable power 0.8.
  - Sample size 38 (power 0.909)
  - CV_w can increase to 0.298 (rel. +19%)
  - *GMR* can decrease to 0.923 (rel. –2.8%)
  - 10 drop-outs acceptable (rel. –26%)
  - Most critical is the *GMR*!



Dealing with Uncertainty

Nothing is ‘carved in stone’.

- Never assume perfectly matching products.
  - Generally a Δ of not better than 5% should be assumed (0.9500 – 1.0526).
  - For HVD(P)s do not assume a Δ of <10% (0.9000 – 1.1111).
- 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 or upper CL.
- Alternatives exist.
  - 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.



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 (α) assuming a *GMR* (θ_0) at one of the limits of the acceptance range [θ_1 , θ_2].
    - Example: 2×2×2 cross-over, *CV* 20%, *n* 20, α 0.05, θ_0 = [θ_1 0.80 or θ_2 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 1.4-2. 2016. https://cran.r-project.org/package=PowerTOST

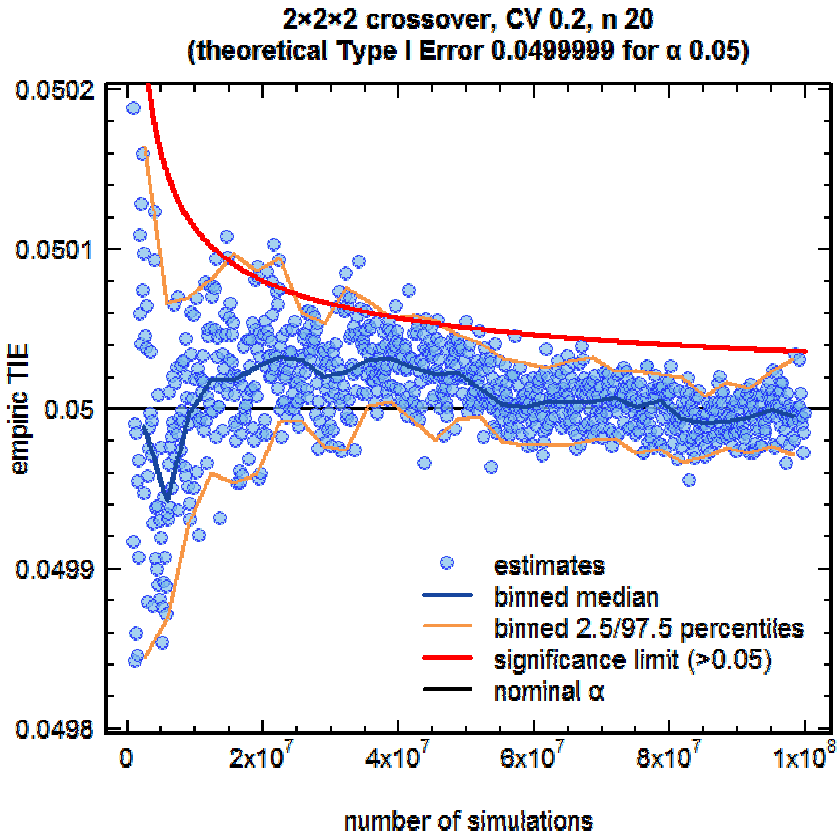
Excursion

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
```

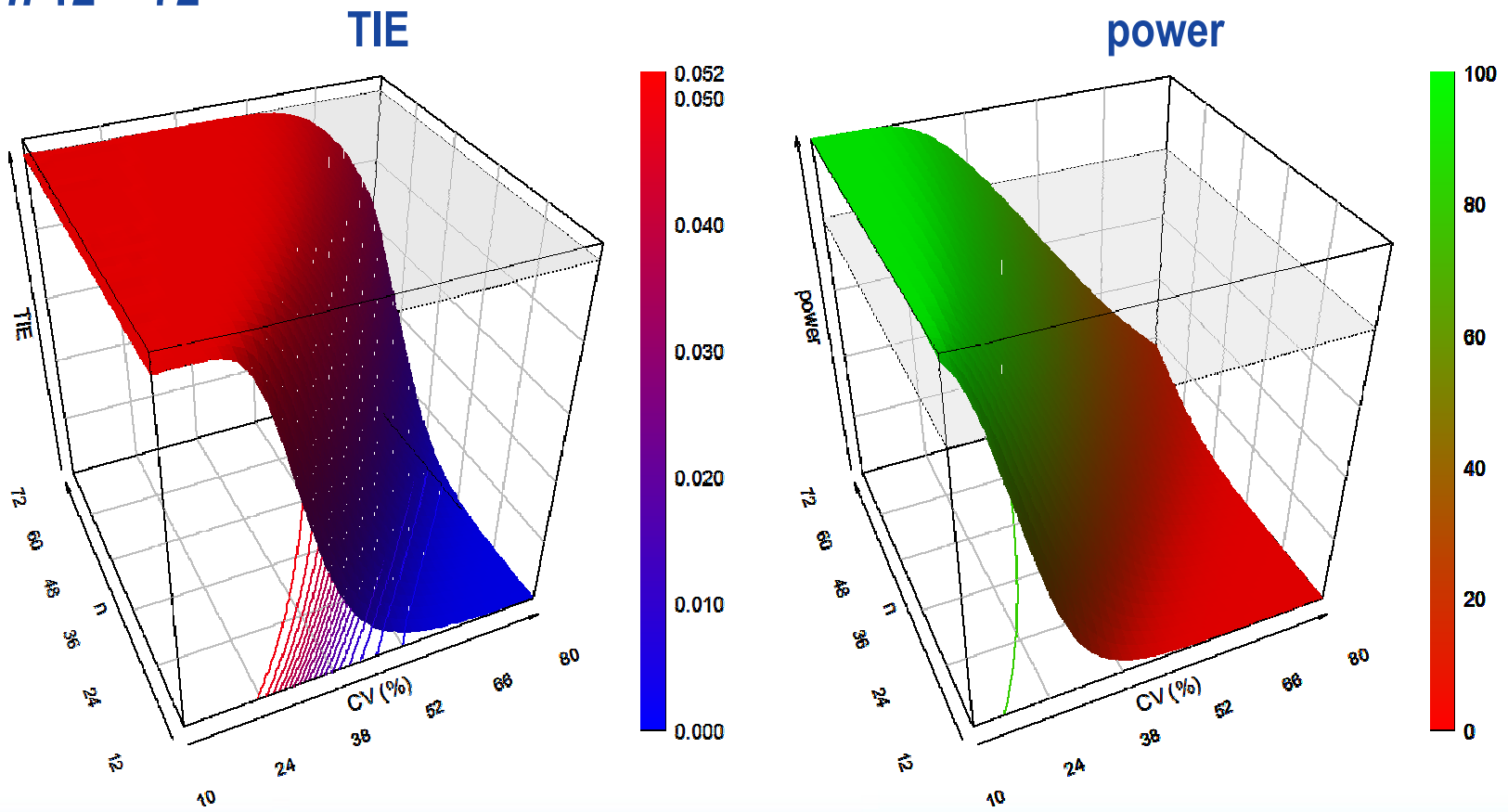
- In other settings (*i.e.*, frameworks like Two-Stage Designs or reference-scaled ABE) analytical solutions for power – and therefore, the TIE – are not possible:
  Simulations are required.



Excursion

Type I Error and power.

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



Dealing with Uncertainty

Variability in the study different from assumption.

- If higher, we gain power. Demonstrate BE even for a worse *GMR*.
- If lower, we loose power. Chances to demonstrate BE decreases and we might loose *a lot* of money (repeat a failed study).
- (Adaptive) Two-Stage Designs
  - First publication in 2008.
  - Many follow-ups (different *GMR*s, power, parallel designs, futility rules).
  - Acceptable according to GLs (EMA 2010, AUS 2011, HC 2012, FDA 2013, Russia 2013, NZ 2015).

(Adaptive) Sequential Two-Stage Designs

Methods by Potvin *et al.* (2008) were the first validated frameworks 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' (Schütz 2015)**
    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.**

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

(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

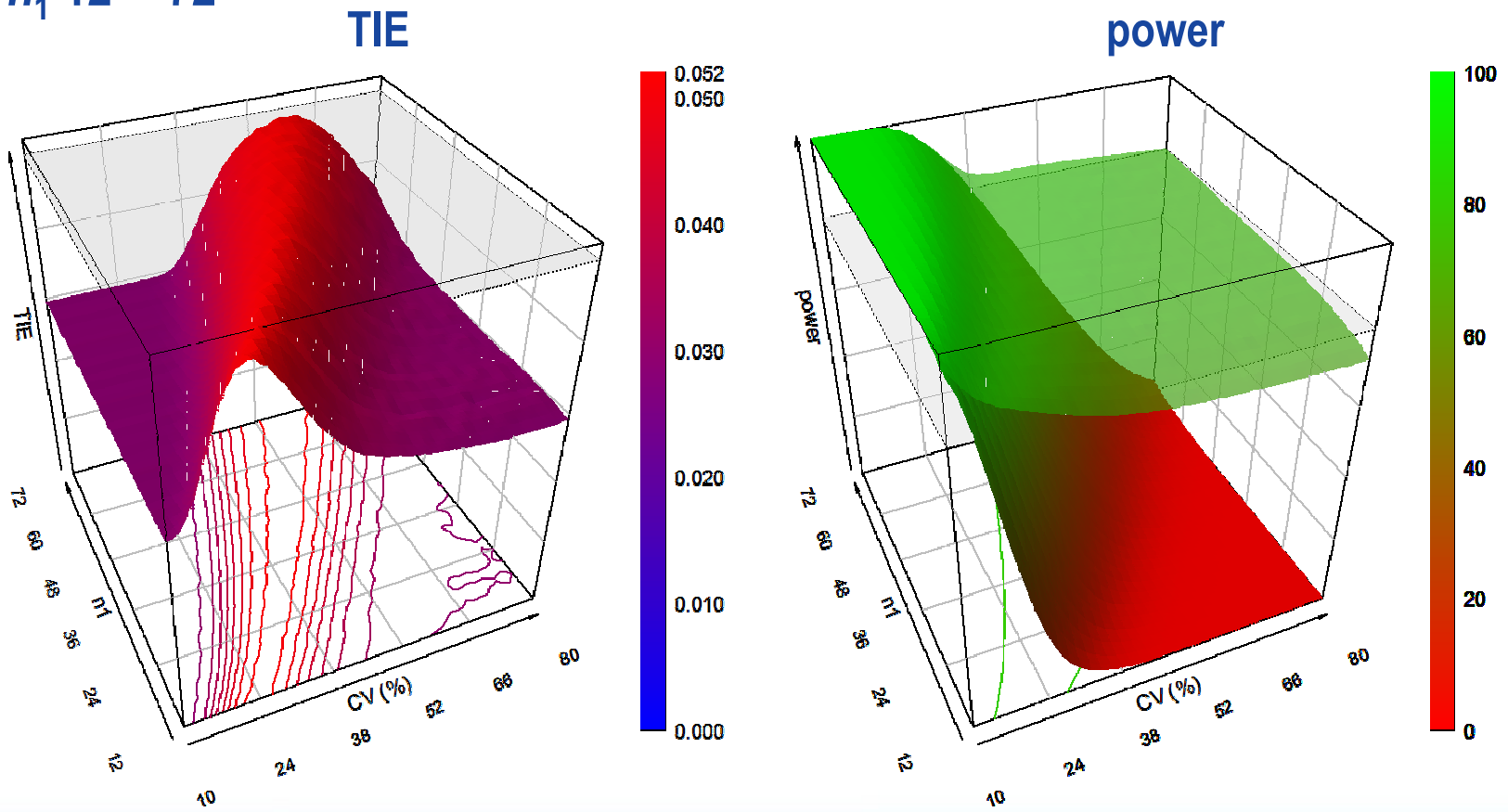
- Xu *et al.* (2015). *GMR* 0.95, target power 80%, futility for the $(1–2\alpha_1)$ CI.

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

Excursion

Type I Error and power.

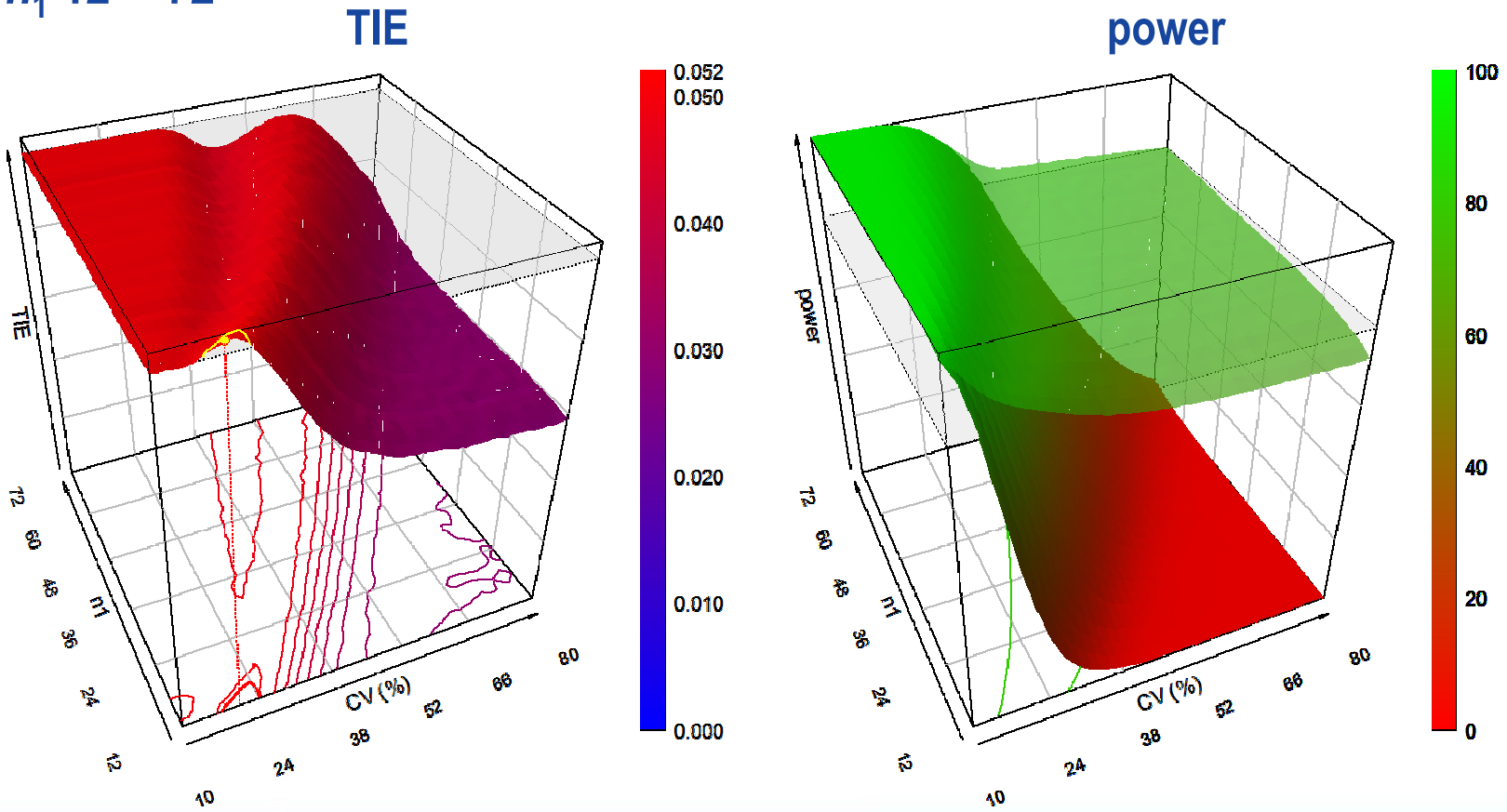
- 'Type 1' TSD (Potvin Method B, α_{adj} 0.0294). *GMR* 0.95, *CV* 10 – 80%, n_1 12 – 72



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



(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%, *GMR* 0.95, target power 80%.
    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

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

High variability

Assumptions (again).

- 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.

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

High variability

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 (Session 2).**
  - **The BE limits can be *scaled* based on the variability of the reference.**
  - **Details in part II.**

Fasting/fed studies to assess food effects

Only necessary for MR products (EMA 2014)

- BE must be demonstrated both in fasting and fed state.
  - Three approaches recommended in the GL:
    - (1) A fully randomized 2×4×4 cross-over study (T and R; both fasting and fed).
    - Or two cross-over studies (different designs):
      - (2) A 2×2×2 cross-over in fasting state and a 2×6×3 cross-over, where T and R are administered in fed state and T additionally in fasting state.
      - (3) A 2×2×2 cross-over in fasting state and a 2×2×2 cross-over in fed state.

Fasting/fed studies to assess food effects

Only necessary for MR products (EMA 2014)

- BE must be demonstrated both in fasting and fed state.

$T_{fast.}$	R_{fed}	$R_{fast.}$	T_{fed}
$R_{fast.}$	$T_{fast.}$	T_{fed}	R_{fed}
T_{fed}	$R_{fast.}$	R_{fed}	$T_{fast.}$
R_{fed}	T_{fed}	$T_{fast.}$	$R_{fast.}$

  - Fully randomized 2×4×4 study (approach 1).
    - Pro: The comparison is done based on a common (pooled) variance of fasting and fed state. It is possible to assess not only BE in fasting and fed state but also the food effect of T and R in a cross-over (higher power than in the two 2×2×2 studies evaluated as parallel groups).
    - Con: The sample size is lead by the – likely – higher variability of fed state. Unclear whether the evaluation is unbiased in a simultaneous evaluation – *maybe* the 'leave one out' approach can be used.
  - Two cross-over studies.
    - Alternative 1 (approach 2):
      - 2×2×2: BE ($T_{fasting}$ *vs.* $R_{fasting}$).
      - 2×6×3: BE (T_{fed} *vs.* R_{fed}) and food effect of Test (T_{fed} *vs.* $T_{fasting}$).
    - Alternative 2 (approach 3):
      - 2×2×2: BE ($T_{fasting}$ *vs.* $R_{fasting}$) and 2×2×2: BE (T_{fed} *vs.* R_{fed}).

Fasting/fed studies to assess food effects

Only necessary for MR products (EMA 2014)

- BE must be demonstrated both in fasting and fed state.
  - Alternative 1 (2×2×2 and 2×6×3).
    - Pro: Since for most products the variability in fed state is larger than in fasting state, sample sizes can be different.
      The assessment of the food effect of the Test is performed in the second cross-over study and thus powerful.
    - Con: The food effect of the Reference is not directly accessible. A comparison of its food effects between studies is statistically demanding.
  - Alternative 2 (two 2×2×2 studies).
    - Pro: Sample sizes can be different.
      Both studies can be performed in Two-Stage Sequential Designs allowing to increase the sample size if necessary (Sessions 4.II and 10).
      The food effects of both T and R can be assessed as parallel groups.
    - Con: The comparison of food effects is much less powerful than in a cross-over design. The outcome might be inconclusive (due to lacking power).

$T_{fast.}$	$R_{fast.}$
$R_{fast.}$	$T_{fast.}$

T_{fed}	R_{fed}	$T_{fast.}$
R_{fed}	$T_{fast.}$	T_{fed}
$T_{fast.}$	T_{fed}	R_{fed}
T_{fed}	$T_{fast.}$	R_{fed}
R_{fed}	T_{fed}	$T_{fast.}$
$T_{fast.}$	R_{fed}	T_{fed}

$T_{fast.}$	$R_{fast.}$
$R_{fast.}$	$T_{fast.}$

T_{fed}	R_{fed}
R_{fed}	T_{fed}

Fasting/fed studies to assess food effects

Only necessary for MR products (EMA 2014)

- BE must be demonstrated both in fasting and fed state.

T_{fed}	R_{fed}	$T_{fast.}$	$R_{fast.}$
R_{fed}	T_{fed}	$R_{fast.}$	$T_{fast.}$

  - My preferred alternative: A *partly* randomized 2×2×4 study.
    - In the first part (per. 1 & 2: fed) subjects are randomized like in a 2×2×2 study.
    - In the second part (per. 3 & 4: fasting) the *same* subjects are randomized like in a 2×2×2 study.
    - Pros:
      - » BE can be demonstrated in fed state (part 1) and fasting state (part 2) in a conventional cross-over design.
      - » If one has reliable information about the variabilities in fed and fasting state ($CV_{fasting} < CV_{fed}$) , it is possible to perform part 2 in fewer subjects.
      - » It is possible to design the two parts as Two-Stage Sequential Designs allowing to increase the sample size if necessary (Sessions 4.II and 10).
      - » The food effect of T and R can be evaluated as a *paired* design, which is almost as powerful as a cross-over.
    - Con: The paired design relies on *no* period effect. However, that's common in assessing linear PK by innovators: MD AUC_{0-T} *vs.* SD $AUC_{0-\infty}$.

Fasting/fed studies to assess food effects

Only necessary for MR products (EMA 2014)

- As long as BE is demonstrated both in fasting and fed state a *different* food effect (of Test and Reference) will not loose the war.
  - A similar food effect is not required – only 'nice to know'. Failing might be pure chance (lack of power – especially if two 2×2×2 studies were performed).
    - If the Test shows a *significantly lower* food effect than the Reference, the EMA welcomes the 'better' product.
    - However, in such a case the applicant could prefer not to claim 'essential similarity' (generic pathway: 2001/83/EC Art. 10(1)) but opt for a 'hybrid application' (additional clinical studies: 2001/83/EC Art. 10(3)) instead.
    - Whether proving such an advantage for the patient (better compliance) over the reference pays off is another story.

MR and the need for steady state studies

FDA

- MD-study generally not required.

Health Canada

- MD-study only required if accumulation can be expected based on the SD-study:

  $AUC_{0-\tau} / AUC_{0-\infty} > 80\%$ of $AUC_{0-\infty}$,

  where τ is the intended dosing interval.

MR and the need for steady state studies

EMA 2014

- PK metrics in the SD-study:
  - C_{max}, AUC_{0-t}, $AUC_{0-\infty}$.
    - Truncated AUC_{0-72} like for IR-products is not acceptable!
      - » Many MR products show flip-flop PK (absorption slower than elimination).
      - » Hence, the *late* part of the curve represents *absorption*.
  - Cut-off for accumulation: $AUC_{0-\tau}$ / $AUC_{0-\infty}$ >90% (!) of $AUC_{0-\infty}$.
- Prolonged release products
  - *With* accumulation:
    - MD required. PK metrics: $C_{max,ss}$, $C_{\tau,ss}$, $AUC_{0-\tau,ss}$.
  - *Without* accumulation:
    - MD not required.
    - *Additionally* in the SD-study:
      PK metrics representative of the shape of the curve
      (*e.g.*, early and terminal partial *AUC*s).

MR and the need for steady state studies

EMA 2014

- Delayed release products
  - MD not required.
- Multiphasic MR products
  - *Additionally* in the SD-study:
    $C_{max(x)}$, $C_{max(x+1)}$, $pAUC_{(x)}$, $pAUC_{(x+1)}$,
    where x is/are pre-defined cut-off time(s).
  - *With* accumulation:
    - MD required. PK metrics: $C_{max,ss}$, $C_{\tau,ss}$, $AUC_{0-\tau,ss}$.
  - *Without* accumulation:
    - MD not required.

Establishing the Biostudy Statistical Design

Thank You!

Open Questions?

Helmut Schütz

BEBAC

Consultancy Services for
Bioequivalence and Bioavailability Studies
1070 Vienna, Austria

helmut.schuetz@bebac.at