

General Hurdles and Pitfalls in BE Studies

Helmut Schütz



Wikimedia Commons • 2013 • Laika ac • CCA-ShareAlike 3.0 Unported

Timing and Project Management

What to do if you have more studies to perform.

- Suggestions

- Start with the most difficult one (*i.e.*, the one which most likely fails) *first!*
 - Variability in fed state commonly higher than in fasting state.
 - Due to potential different food effects of T and R the *GMR* may be worse.
 - » Hence, fed study → fasting study.
- MR: If the GL allows waiving the MD-study, perform the SD-study and assess the *additional* PK metrics (*e.g.*, early and terminal *pAUCs*) for BE.
 - If you fail these PK metrics (but still pass C_{max} , AUC_{0-t} , $AUC_{0-\infty}$) perform the MD-study.
 - » If you have performed the SD- and MD-study and pass required PK metrics in both, the failing *pAUCs* in the SD-study are ‘overruled’.
 - » Since the purpose of *pAUCs* was only to justify waiving the MD-study (which was later performed) there is no reason for an assessor not accepting the application.

Timing and Project Management

What to do if you have more studies to perform.

- **Suggestions**
 - **Variability in steady state is generally lower than after a single dose.**
 - Estimate the *CV* from the *SD*-study.
 - Perform the *MD*-study in a **Two-Stage-Design** where the size of the first stage is **~75%** of a fixed sample design.
 - » Reasonably high chance to pass already in the first stage (due to lower *CV*).
 - » If the *CV* is higher (unlikely!) you still get a second chance.
 - **If ever possible try to perform studies in the same CRO.**
 - If there are problems with the clincial capacity (→ different CROs), employ still the *same* bioanalytical CRO.
 - » If you face capacity problems in bioanalytics (→ different CROs) make sure (!) that the *same* validated method is used.
 - » If ever possible,
 - (a) assure that the same type of instruments are used and
 - (b) run a cross-validation between sites.

Timing and Project Management

Large studies – lacking capacity of the clinical site.

- Suggestions

- Find a larger CRO – even if more expensive!
- If you have to split the estimated sample size into groups:
 - Dose subjects within a limited time frame, e.g., the groups only days apart (sometimes called the ‘staggered approach’).
Group I : period 1, Mo – We → washout → period 2, Mo – We
Group II: period 1, Th – Sa → washout → period 2, Th – Sa
 - Do *not* split groups into equal sizes.
Perform at least one in the maximum capacity of the clinical site.
 - Some jurisdictions (Russian MoH, Eurasian Economic Union, and Saudi FDA always, FDA regularly, EMA rarely) require a statistical test for the ‘group-by-treatment interaction’.
 - » If this test is significant at the 0.1 level, one is *not* allowed to pool the data and is only free to demonstrate BE in the *largest* group.

Timing and Project Management

Large studies – lacking capacity of the clinical site.

- Example

- CV of AUC 30% (no scaling allowed), GMR 0.90, target power 90%, 2×2×4 (reference-scaling of C_{max} intended). Estimated sample size 54.
- Maximum capacity 24 beds.
 - Option 1: Equal group sizes (3 × 18).
 - Option 2a: Two groups with the maximum size (24), the remaining one 6.
 - Option 2b: One group 24, the remaining ones as balanced as possible (16 | 14).
- Let us assume that there are no drop-outs and pooling is not allowed (significant group-by-treatment interaction). Expected power:
 - Option 1: 51% in each of the groups.
 - Option 2a: 62% in the two largest groups (n = 24 each).
 - Option 2b: 62% in the largest group.
- Which one would you prefer – and *why*?

Group Effect

Review of Guidelines

- **FDA 2001**
 - If a crossover study is carried out in two or more groups of subjects (e.g., if for logistical reasons only a limited number of subjects can be studied at one time), the statistical model should be modified to reflect the multigroup nature of the study. In particular, the model should reflect the fact that the periods for the first group are different from the periods for the second group.
 - If the study is carried out in two or more groups and those groups are studied at different clinical sites [...], questions may arise as to whether the results from the several groups should be combined in a single analysis.

Group Effect

Review of Guidelines

- **FDA**
 - No details about the analysis is given in any guidance. However, this text can be found under the FOI:
 - The following statistical model can be applied:
 - » **Group**
 - » **Sequence**
 - » **Treatment**
 - » **Subject (nested within Group × Sequence)**
 - » **Period (nested within Group)**
 - » **Group-by-Sequence Interaction**
 - » **Group-by-Treatment Interaction**
 - **Subject (nested within Group × Sequence) is a random effect and all other effects are fixed effects.**

Group Effect

Review of Guidelines

- FDA
 - FOI (cont'd):
 - If the Group-by-Treatment interaction test is not statistically significant ($p \geq 0.1$), only the Group-by-Treatment term can be dropped from the model.
 - If the Group-by-Treatment interaction is statistically significant ($p < 0.1$), DBE requests that equivalence be demonstrated in one of the groups, provided that the group meets minimum requirements for a complete bioequivalence study.
 - Please note that the statistical analysis for bioequivalence studies dosed in more than one group should commence only after all subjects have been dosed and all pharmacokinetic parameters have been calculated. Statistical analysis to determine bioequivalence within each dosing group should never be initiated prior to dosing the next group; otherwise the study becomes one of sequential design.

Group Effect

Review of Guidelines

- FDA

- FOI (cont'd):

- If ALL of the following criteria are met, it may not be necessary to include Group-by-Treatment in the statistical model:
 - » the clinical study takes place at one site;
 - » all study subjects have been recruited from the same enrollment pool;
 - » all of the subjects have similar demographics;
 - » all enrolled subjects are randomly assigned to treatment groups at study outset.
 - In this latter case, the appropriate statistical model would include only the factors Sequence, Period, Treatment and Subject (nested within Sequence).



Group Effect

Review of Guidelines

- EMA 2010
 - The study should be designed in such a way that the formulation effect can be distinguished from other effects.
 - The precise model to be used for the analysis should be pre-specified in the protocol. The statistical analysis should take into account sources of variation that can be reasonably assumed to have an effect on the response variable.



Group Effect

Models proposed by the FDA

- **Model I**
 - **Fixed effects:**
Group, Sequence, Treatment, Period(Group), Group×Sequence, Group×Treatment
 - **Random effect:**
Subject(Group×Sequence)
 - If the Treatment-by-Group interaction term is not significant at the 0.1 level, data of all groups can be pooled and the term dropped (*i.e.*, proceed with Model II).
 - If the Treatment-by-Group interaction term is significant at the 0.1 level, data must not be pooled and Model III of the largest site applied.
 - Intra-subject contrasts for the estimation of the treatment effect (and hence, a PE and its CI) cannot be unbiased obtained from this model. It serves only as a decision tool.

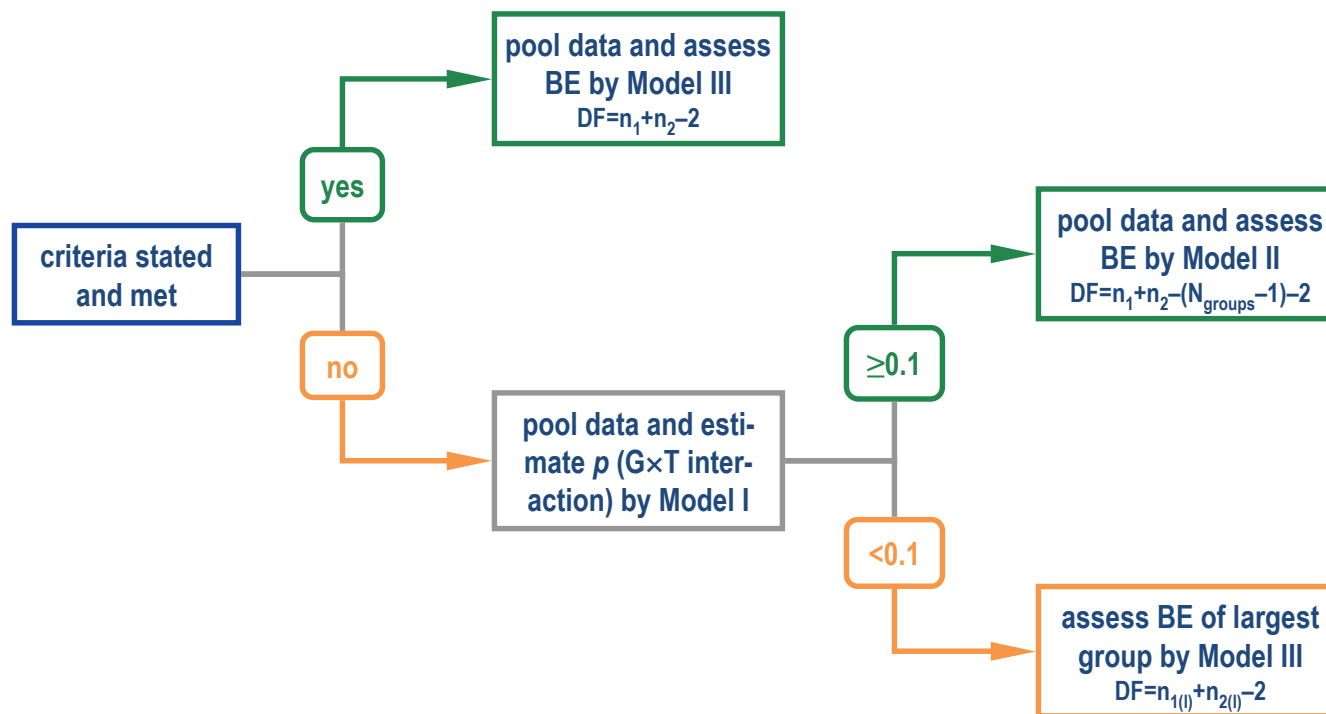
Group Effect

Models proposed by the FDA

- **Model II**
 - Fixed effects:
Group, Sequence, Treatment, Period(Group), Group×Sequence
 - Random effect:
Subject(Group×Sequence)
 - The model takes the multigroup nature of the study into account and is more conservative than the naïve pooled model (three degrees of freedom less than Model III).
- **Model III**
 - Fixed effects:
Sequence, Treatment, Period
 - Random effect:
Subject(Sequence)
 - This is the common model for 2×2×2 crossover studies.

Group Effect

Evaluation according to the FDA



Group Effect – a Modern Myth?

Testing for a $G \times T$ interaction by the FDA's Model I

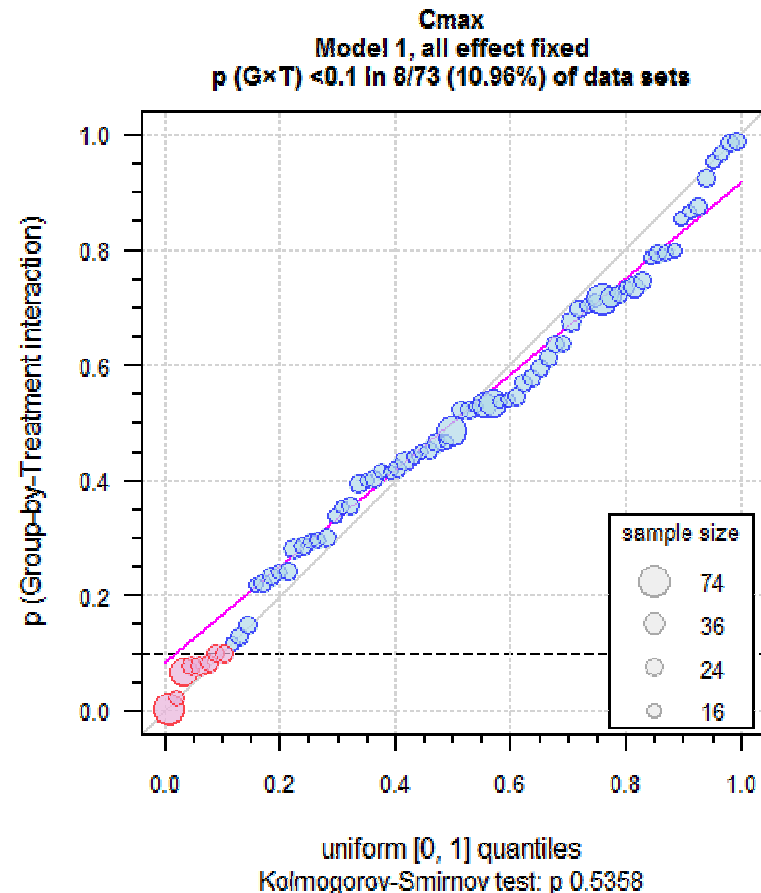
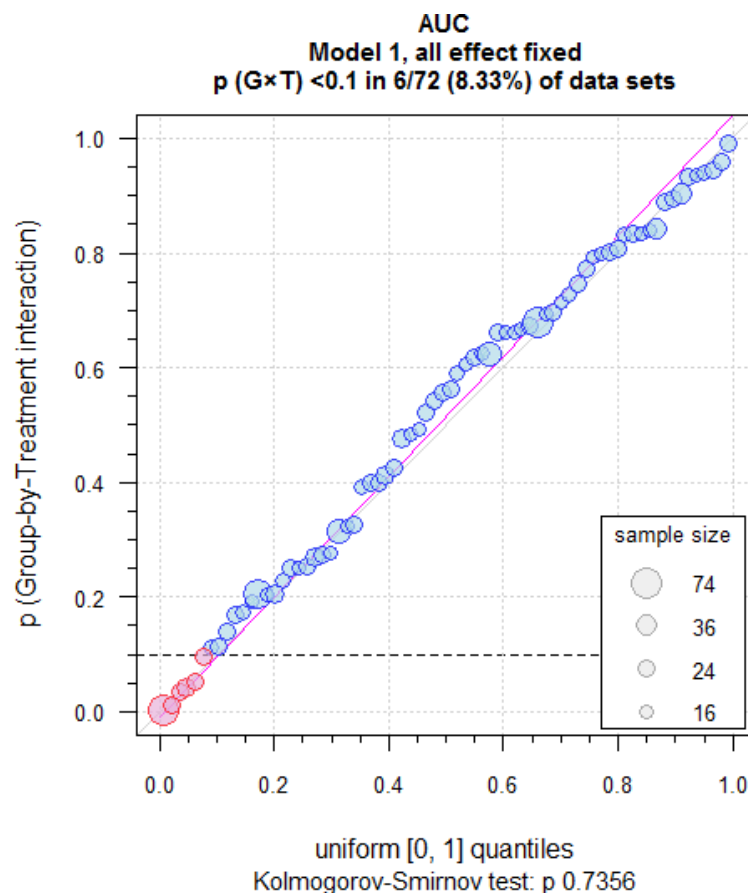
- Low sensitivity (between subject term).
Hence, testing at the 0.1 level is recommended.
- We can expect a false positive rate (*i.e.*, if there is *no true* $G \times T$) in ~10% of studies.
 - Consequences
 - Pooling of data not allowed.
 - Drop in power if BE has to be demonstrated in the largest group(s).

Retrospective evaluation of studies from our files

- 62 studies (single / multiple dose; mainly BE but also food-effect, DDI)
 - 50 analytes (72 data sets of AUC , 73 data sets of C_{max}).
 - Two to four groups.
 - Median interval between groups three days (range 1 – 18 days).

Group Effect – a Statistical Artifact?

Testing for a G×T interaction by the FDA's Model I



Group Effect

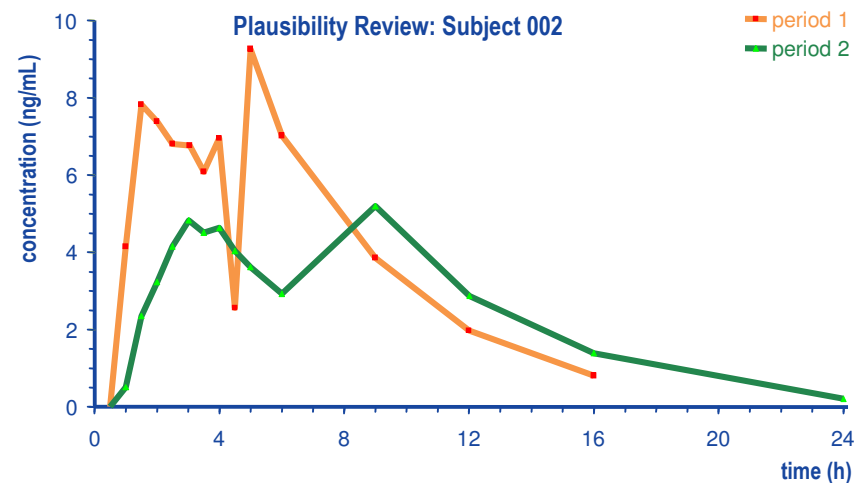
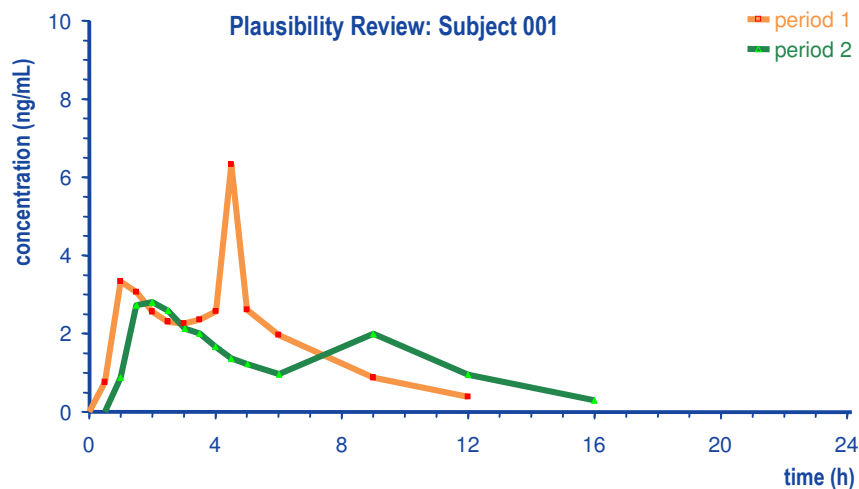
Recommendations and potential Alternatives

- EMA
 - I have never seen a single case where in the protocol anything was stated (*i.e.*, implicitly no impact on the treatment comparison was assumed). Pooled data and the model given in the GL was used.
 - In one deficiency letter (biosimilar 2016) the EMA requested Model II of pooled data (but *without* a pre-test).
- Russia / Eurasian Economic Area / MENA-States
 - Deficiency letter guaranteed if nothing about the group effect was stated in the protocol. *Post hoc* justification never accepted.
 - If applicable, state in the protocol that the criteria for pooling are met. Evaluate the study by the model given in the GL.
 - Alternatively, state in the protocol that the multigroup nature of the study will be taken into account.
Pool data and evaluate the study by Model II. Loss in power is negligible.

Pitfalls: Case Study 1

Sample mix-up.

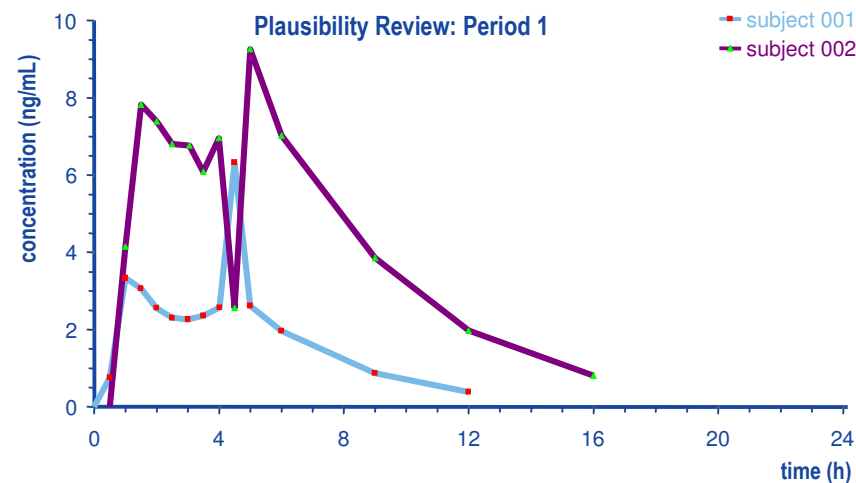
- Very large CRO (study performed in 2008). Common drug, biphasic MR formulations, pilot study (suboptimal sampling between 6 – 14 h).



Pitfalls: Case Study 1

Sample mix-up.

- Barcode-system out of order in the first period of the study.
 - No bail-out procedure (e.g., four eyes principle).
 - Suspected sample mix-up at 4.5 h.
 - Concentrations confirmed.
 - No deviation documented in clinical phase.
 - Drug has very low intra-subject CV ($AUC \leq 10\%$, C_{max} 10–15%) and high inter-subject CV (>50%) due to polymorphism.
- Pivotal studies are generally performed in only 14 subjects.
- A single mixed-up sample close to t_{max} could ruin an entire study.



Pitfalls: Case Study 1

Sample mix-up.

- We tried to confirm the mix-up by comparing lab-values of the suspect samples (and each of the two neighbouring ones in each profile).
- Anticoagulant was citrate for GC/MS.
- With this anticoagulant the analyzer was validated only for γ -GT and albumine.

subject	time (h)	analyte (ng/mL)	γ -GT (U/L)	albumine (g/dL)
001	4.0	2.572	13	3.8
001	4.5	6.330	9	3.5
001	5.0	2.615	14	3.9
002	4.0	6.956	9	3.4
002	4.5	2.561	14	4.0
002	5.0	9.262	8	3.4

- γ -GT and albumine showed a similar pattern like the analyte.
- Mean values of γ -GT in the pre- and post-study lab exams were 14 U/L (# 001) and 9 U/L (# 002). Means of albumine were 3.9 g/dL (# 001) and 3.4 g/dL (# 002).
- Luckily subjects differed in their values. The pilot study was only supportive...

Pitfalls: Case Study 1

Sample mix-up.

- Before the current EMA GLs a blinded plausibility review was acceptable (and still is in many regulations like the FDA).
- According to the current EMA GLs re-analyzing of samples is not permitted.
 - Gerald Beuerle of TEVA/ratiopharm (joint EGA/EMA workshop, London 2010) presented an example where due to a single mix-up a study would *pass*.
 - » The study would *fail* to show BE if the results were exchanged.
 - » The study would *fail* to show BE if the two subjects were excluded.
 - » Panelists of the EMA's PKWP confirmed that either procedure is not acceptable and the values have to be used as *the are* (i.e., the study would *pass*).
 - Helmut Schütz: *'The EMA is a Serious Risk to Public Health!'*
- At the 2nd International Conference of the Global Bioequivalence Harmonization Initiative (Rockville, 15 – 16 September 2016) Session IV was devoted to the issue (*Exclusion of PK Data in the Assessment of IR and MR Products*).

Pitfalls: Case Study 1

Sample mix-up.

- Lessons learned:
 - The most critical phase is the transfer from centrifuged blood sample tubes to the vials containing the sample matrix used in bioanalytics.
 - When we installed a barcode-system in 1991, the rate of sample mix-ups dropped from 0.2% to zero.
 - A bail-out procedure must be in place (four eyes principle), an SOP at hand and followed by the personnel!
 - I once audited a CRO where the SOP mandated that the centrifuged samples and vials are scanned one after the other – immediately after the transfer.
 - » The technician took four Eppendorf vials (centrifuged blood samples) in his left hand and scanned them.
 - » Then he scanned four empty sample vials.
 - » Next he pipetted the four samples one after the other.
 - » *'Why are you do this in such a way?' –
'It saves time, and four vials fit nicely in my hand.'*

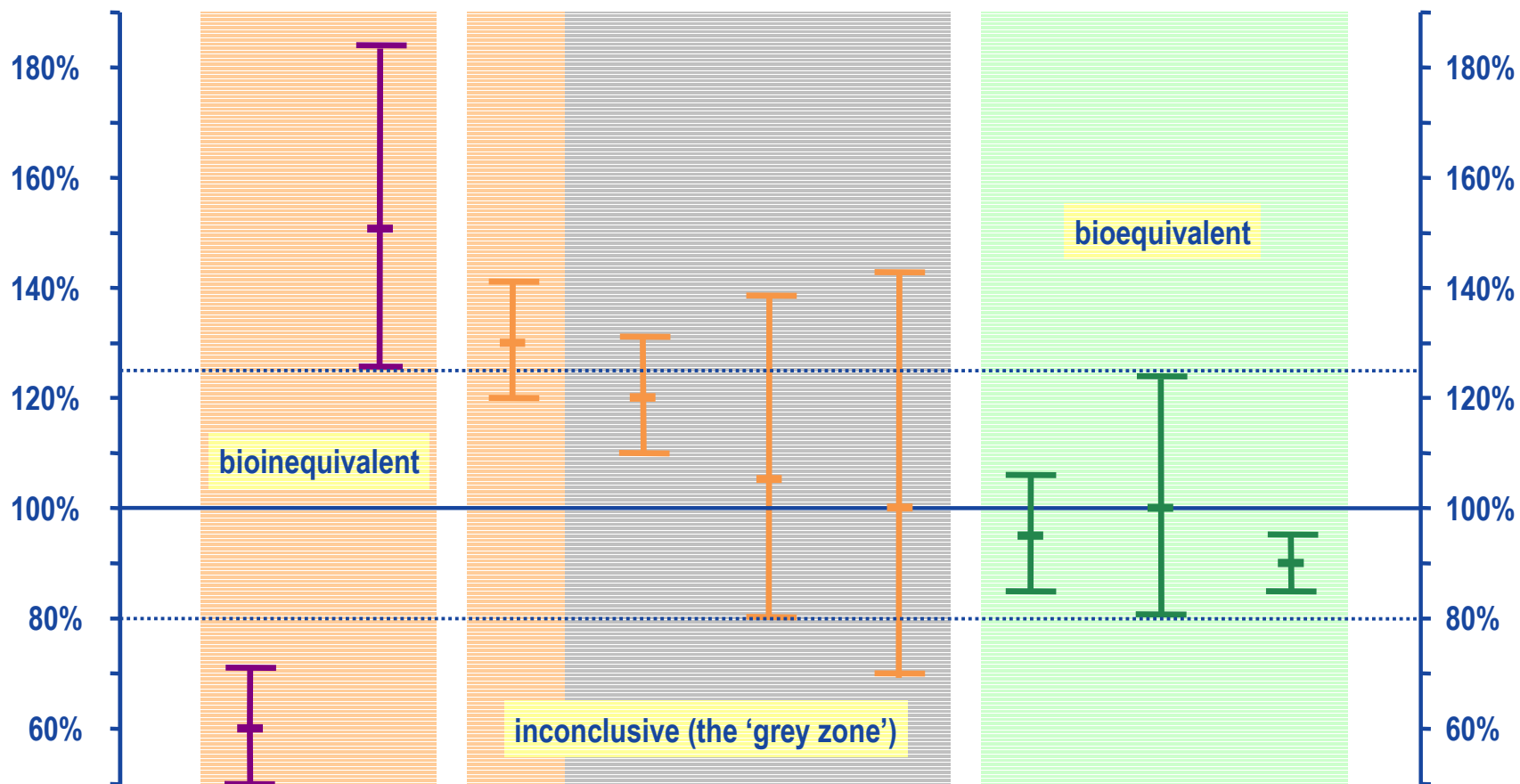
'Lack' of Statistical Power

Sometimes a properly planned study fails by pure chance.

- Power is *fixed* by design (in the sample size estimation)!
- It is unavoidable, that the producer's risk (probability of Type I Error, where $\beta = 1 - \text{power}$) hits in actual studies.
 - If studies are planned with 80% power,
one out of five studies will fail –
even if products *are* bioequivalent.
 - *Post hoc* (aka *a posteriori*) power is a flawed statistical concept.
 - Reporting *post hoc* power is a bad habit and should be abandoned.
 - **Either** a study has demonstrated bioequivalence **or not**.
 - » As 'high' power does not further *support* the claim of already demonstrated BE,
 - » 'low' power does not *invalidate* the conclusion of BE!
- The only *realistic* remedy for a failed study is to repeat it in a larger sample size – *if the PE is promising*.

'Lack' of Statistical Power?

Some studies: Point estimates and their 90% CIs.



Are Add-on Studies acceptable?

Add-on Designs

- In an Add-on Design (AOD) an initial group of subjects is treated and – if the result is inclusive (*i.e.*, although the point estimate is within the BE-limits, the CI is not) –
 - an additional group of subjects can be recruited and
 - the assessment of bioequivalence repeated in the pooled dataset.
- General conditions:
 - The intention to perform an AOD has to be stated in the protocol.
 - The same batches of products and the same clinical and bioanalytical methods have to be employed in both groups.
 - Additional requirements were stated in some jurisdictions.
- Somewhat popular in the 1990s and reflected in regulatory documents (HC 1992, NZ 1997) – and later abandoned. Currently still in Argentina (2006), Korea (2008), Japan (2012), Mexico (2013).

Are Add-on Studies acceptable?

Add-on Designs

- Statistically questionable
 - Repeated testing without adjusting the level of the tests will inflate the Type I Error (patient's risk).
 - If k repeated test are performed at $\alpha 0.05$, the TIE will approach $1 - (1 - \alpha)^k$ or 9.75% for two tests.
 - In naïve pooling of data, both the variance will be underestimated and the nominal level of the test will be exceeded.
 - Inflation of the TIE demonstrated in simulations (Potvin *et al.* 2008, Wonne-
mann *et al.* 2015, Schütz 2015).
- Preserving the consumer risk
 - Bonferroni correction (for two tests $\alpha 0.025$ or a 95% CI) keeps the TIE at $\leq 4.94\%$.
 - Sample size penalty compared to a fixed-sample design (20–30% more subjects).
 - n_2 should be $\geq n_1$ (Birkett and Day 1994).

Are Add-on Studies acceptable?

Add-on Designs

- Only if unavoidable!
 - If you apply in Argentina, Korea, Japan, or Mexico – aim for a scientific advice suggesting a Two-Stage Design (Session 4, part I) instead.
 - If you do not succeed:
 - Employ Bonferroni's adjustment (95% confidence interval).
 - Adjust the sample size accordingly.

Failing a fed or fasting part of the Study

MR products (EMA 2014) and some product-specific guidance by the FDA

- Fasting and fed in the *same* study in the EMA's approaches 1 and 2.
- Fasting and fed in *separate* studies (fasting, fed) in the EMA's approach 3 and recommended by the FDA.
- Suggestions
 - Educated guess whether the study failed *only* by lacking power (too small sample size) or a 'bad' point estimate (slides 11–12).
 - If the PE is promising, repeat the study in a larger sample size.
 - » If fasting/fed was nested in a design (EMA #1 and #2) it will be difficult. If you repeat the entire study due to pure chance the respective other comparison may fail this time due to pure chance.
 - » For EMA #3 and the FDA repeat the respective study.

Failing a fed or fasting part of the Study

MR products (EMA 2014) and some product-specific guidance by the FDA

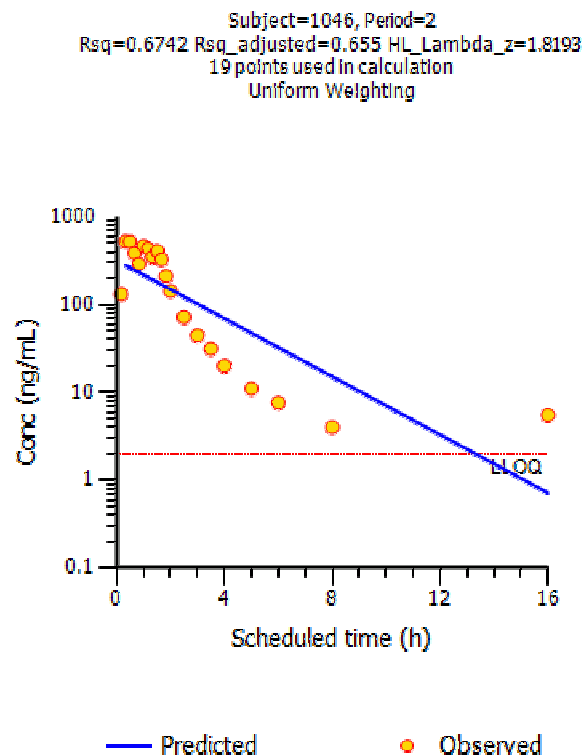
- **Suggestions**

- If products are inequivalent (CI completely outside the BE-limits) or if the PE is not promising (e.g., close to or even outside the BE-limits) modify the formulation.
 - » If you did not do that before, consult with an expert in IVIVC and explore new dissolution methods (maybe biorelevant).
 - » Development of candidate formulations with different release characteristics.
 - » Pilot *in vivo* studies and development of a discriminatory dissolution method which allows selection of a test formulation which matches the reference *in vitro*.
 - » Repeat the entire pivotal BE-program.

Pitfalls: Case Study 2

NCA (estimating λ_z).

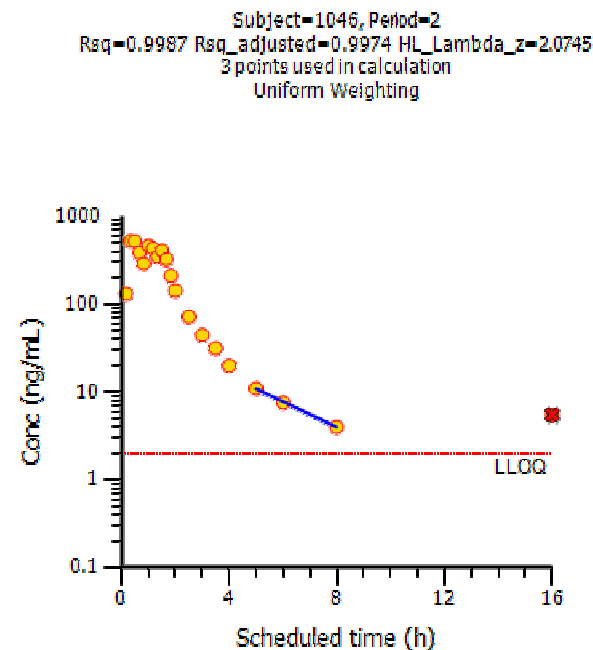
- Large CRO (study performed in 2013). 4-period full replicate; the double peak is specific for the formulation.
 - In four cases the last concentration was *increasing*. The CRO followed EMA's GLs and did not re-analyze samples (PK reason alone not sufficient). Obviously the CRO tried to 'save' the profiles by including more data points...
 - To the right the most extreme case.
 - Two samples (at 10 & 12 h) were BLQ.
 - 5.47 ng/mL ($\sim 2.7 \times$ LLOQ) at 16 h.
 - The first time point for the estimation of λ_z was t_{max} .



Pitfalls: Case Study 2

NCA (estimating λ_z).

- What I would do (if an SOP allows that). Two options:
 - Exclude the doubtful value from the estimation of λ_z . Justifications:
 - » The estimated half-life of 2.07 h is consistent with the ones of the same subject in the other periods (2.12, 2.00, 2.16 h).
 - » Two values before the doubtful value were BLQ – which agrees with the predicted λ_z .
 - Drop the profile from the *AUC* comparison, but keep C_{max} (higher variability anyway and reference-scaling intended in the protocol).



Pitfalls: Case Study 2

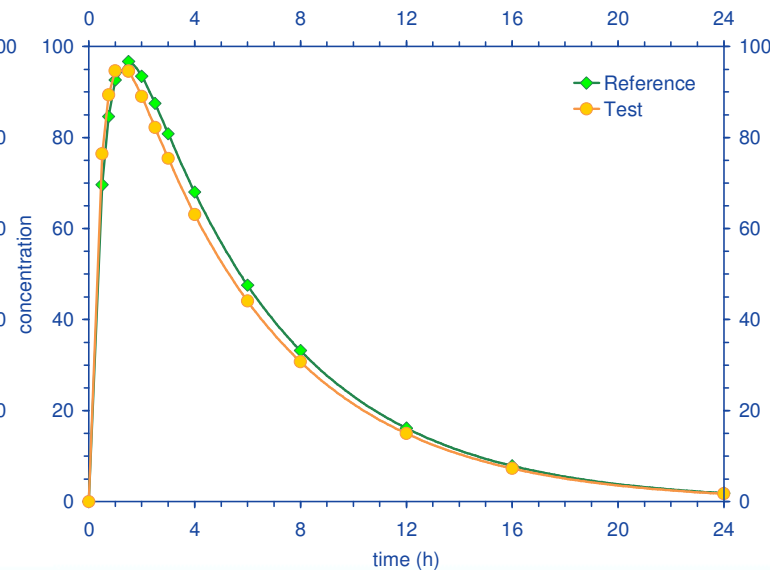
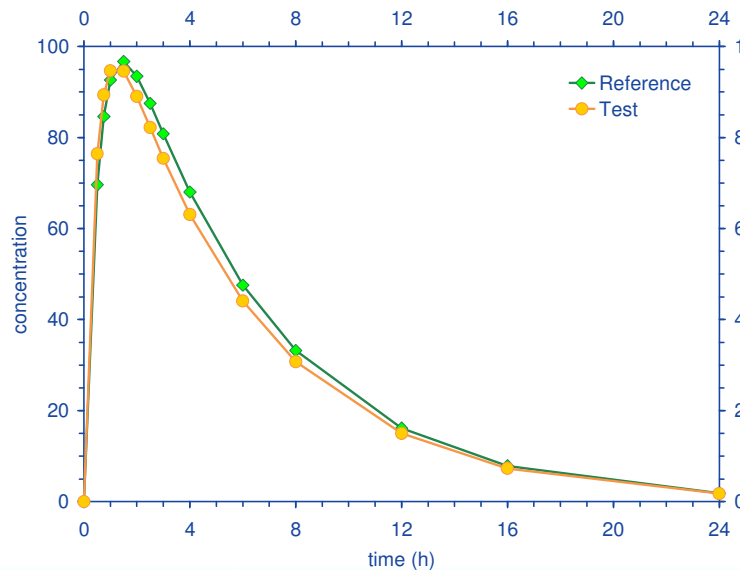
NCA (estimating λ_z).

- Lessons learned:
 - Never solely rely on automatic methods (maximum R^2_{adj}) implemented in software.
 - Visual inspection of the fit (and correction if necessary) recommended (Hauscke *et al.* 2007, Scheerans *et al.* 2008).
 - For IR products absorption is essentially complete after two times t_{max} . Hence, $\geq 2 \times t_{max}$ is good starting point to get an unbiased estimate of λ_z (not substantially contaminated by absorption).
 - In WinNonlin 5.3 (Pharsight) and Kinetica 5.0 (Thermo Scientific) t_{max} can be included by the automatic method. Update the software (Phoenix/WinNonlin ≥ 6.0) or rule it out in an SOP.
 - Have an SOP in place which allows
 - » visual inspection of fits / correction (mandatory),
 - » exclusion of a subject from the AUC comparison if no reliable fit can be established (good) or
 - » exclusion of data points (much better).

Pitfalls: Case Study 3

NCA (trapezoidal methods).

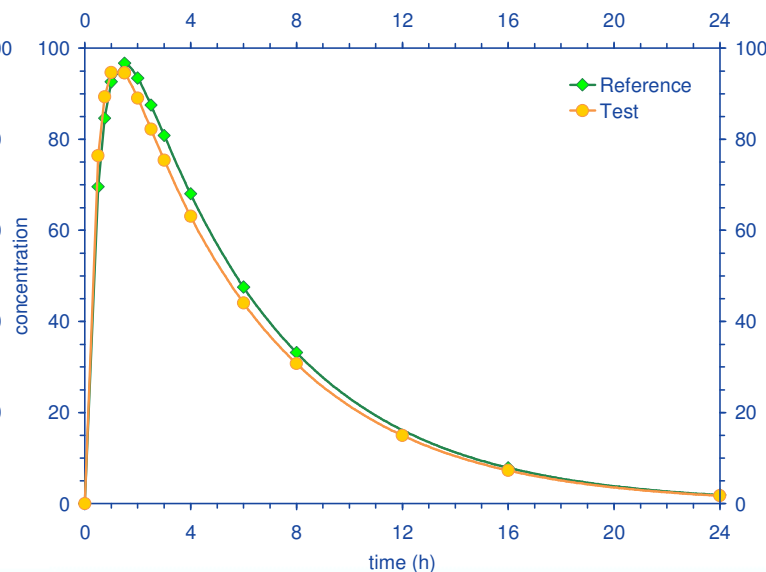
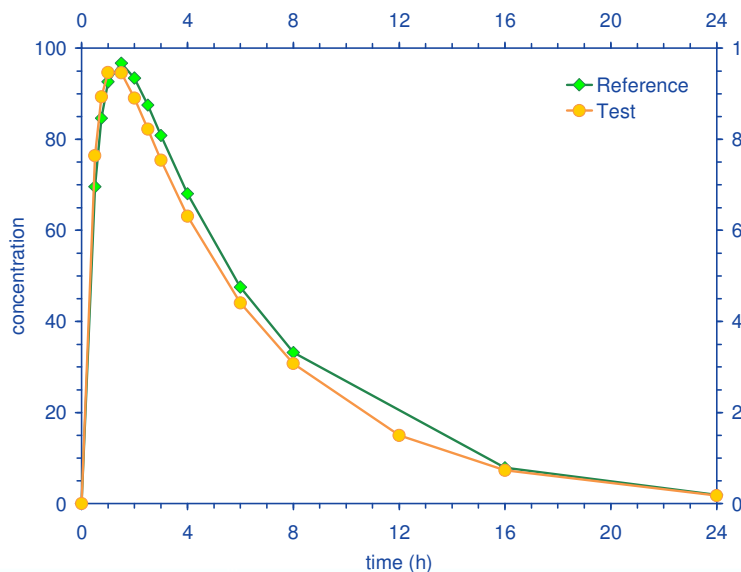
- If all samples are available, there is practically no difference between algorithms.
 - Simulated data. AUC_{∞} 697.8 (Reference), 662.9 (Test), true *GMR* 95.00%.
 - Linear trapezoidal: 707.6 (R), 670.9 (T); *GMR* 94.85% (bias **-0.20%**).
 - Lin-up / log-down trapezoidal: 693.7 (R), 658.0 (T); *GMR* 94.89% (bias **-0.16%**).



Pitfalls: Case Study 3

NCA (trapezoidal methods).

- If a sample is missing (e.g., vial broken in centrifugation), the chosen algorithm matters. 12 h sample (R) removed.
 - Simulated data. AUC_{∞} 697.8 (Reference), 662.9 (Test), true *GMR* 95.00%.
 - Linear trapezoidal: 725.1 (R), 670.9 (T); *GMR* 92.53% (bias **-2.60%**).
 - Lin-up / log-down trapezoidal: 693.7 (R), 658.0 (T); *GMR* 94.89% (bias **-0.15%**).



Pitfalls: Case Study 3

NCA (trapezoidal methods).

- Lessons learned:
 - Trapezoidal methods
 - The linear trapezoidal method goes back to the times were we drew profiles on millimeter paper, clipped them, and weighed them on an analytical balance.
 - I never saw anybody using a curve template in order to approximate an exponential decrease. Connecting data points by straight lines was state-of-the-art.
 - With a few exceptions (ethanol, Michaelis-Menten PK) we know [*sic*] that concentrations decrease exponentially.
Therefore, the most suitable NCA-method for calculating the AUC is the lin-up / log-down trapezoidal method.
 - Missing samples are not uncommon.
 - Only with the lin-up / log-down trapezoidal method we get unbiased estimates of the AUC.
 - The linear trapezoidal method should be abandoned.

General Hurdles and Pitfalls in BE Studies

Thank You!
Open Questions?



Helmut Schütz
BEBAC

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