



PK-NCA, PK based Design, Biostatistics Part II

Helmut Schütz
BEBAC

Outliers

- Russian GL (Section 8)

In bioequivalence studies one or several subjects may show some parameters or their ratios significantly deviating from the core group (“outliers”). Detection of the outliers can be performed in appropriate statistical tests. Such observations are illustrated by means of charts showing individual standard deviations (centered by the mean value and normalized by the standard deviation).

The outliers can be discarded in a bioequivalence study, in case their exclusion is justified.

Outliers

- Types

- Concordant outlier (Type I)

- The PK response for *both* test and reference deviates from the majority of the study sample

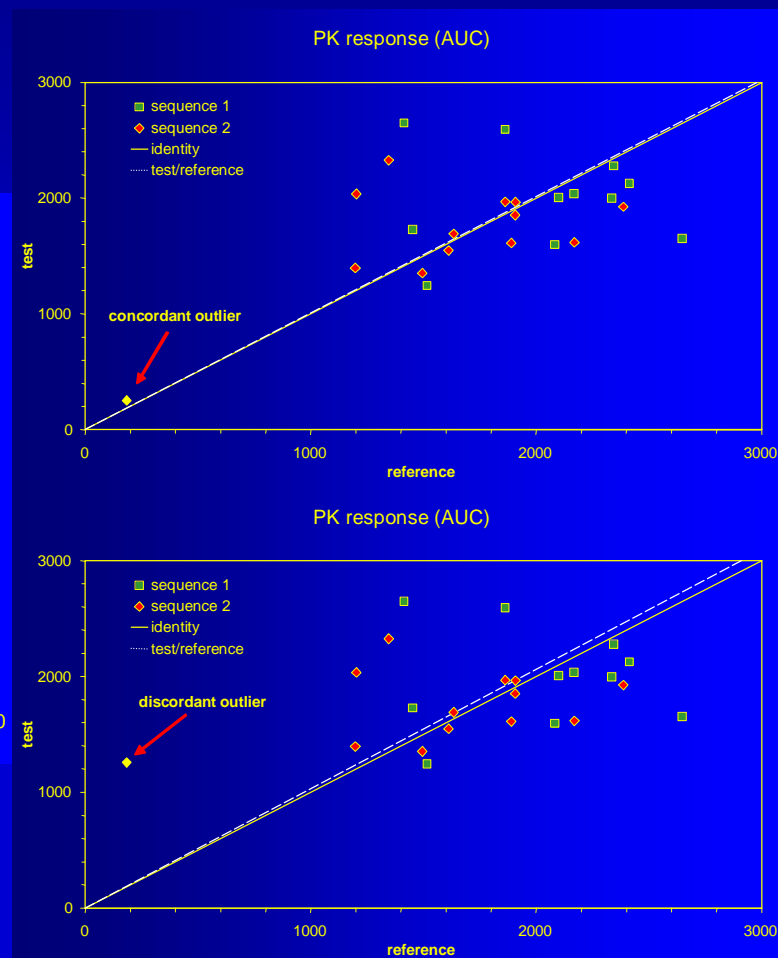
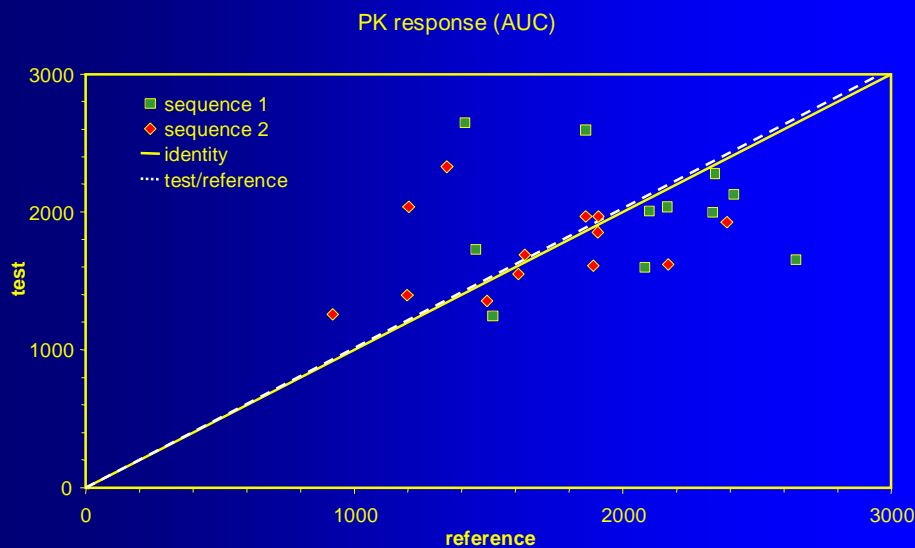
- Poor metabolizers may lead to high concentrations in 5–10% of subjects
 - Does not effect the BE-assessment
 - Should be discussed (polymorphism known?)

- Discordant outlier (Type II)

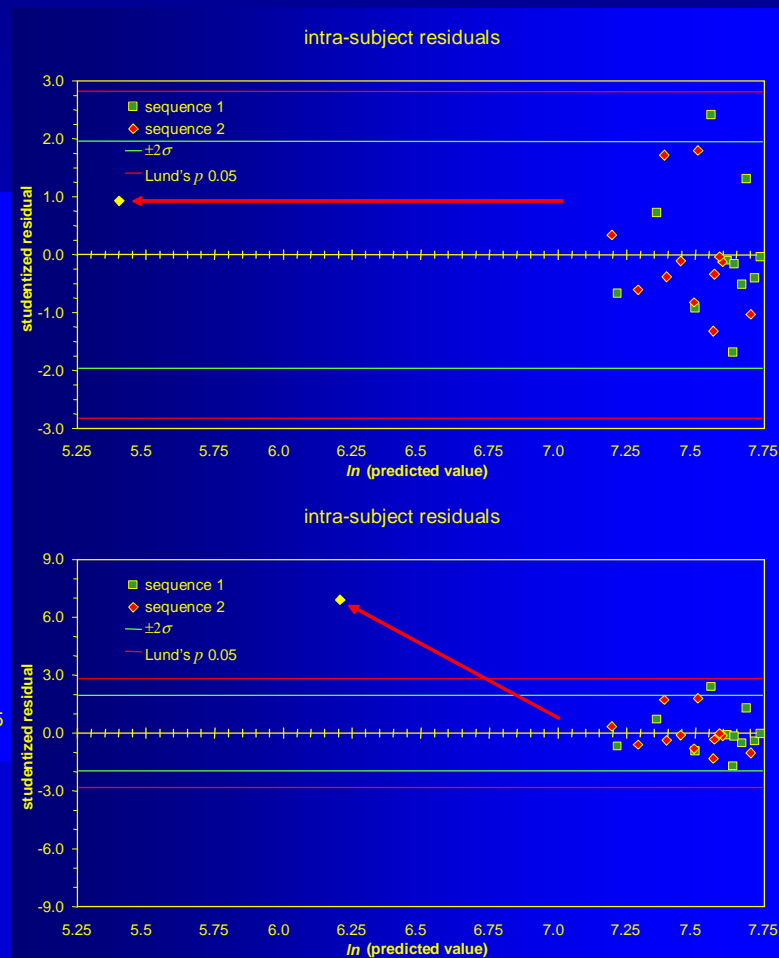
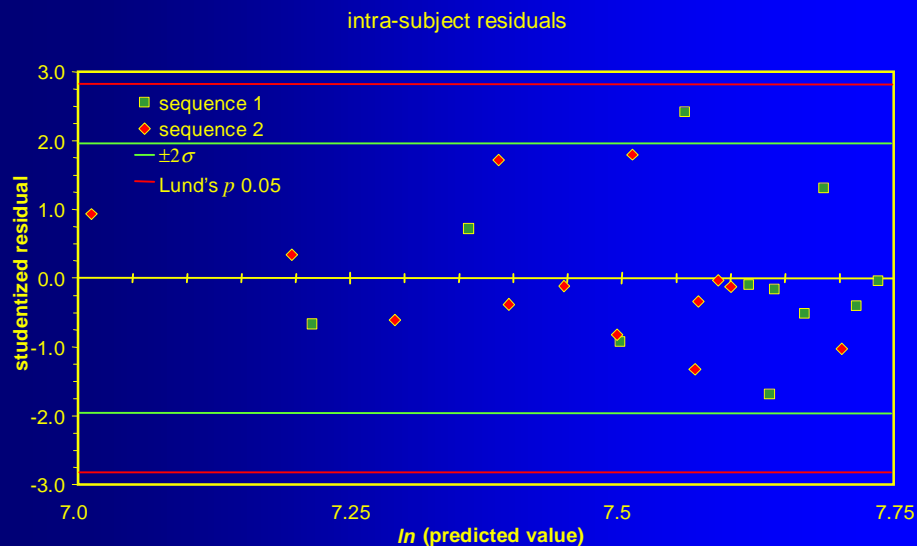
- The PK response of *either* test *or* reference deviates from the majority of the study sample

- Influences the BE-assessment to a great extent

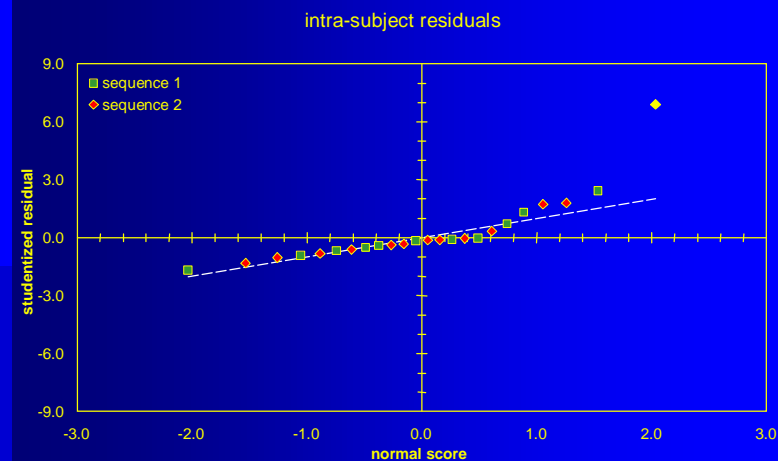
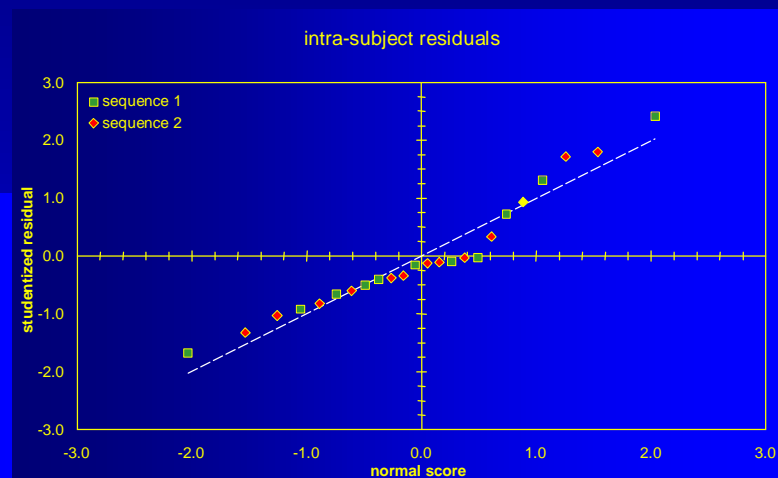
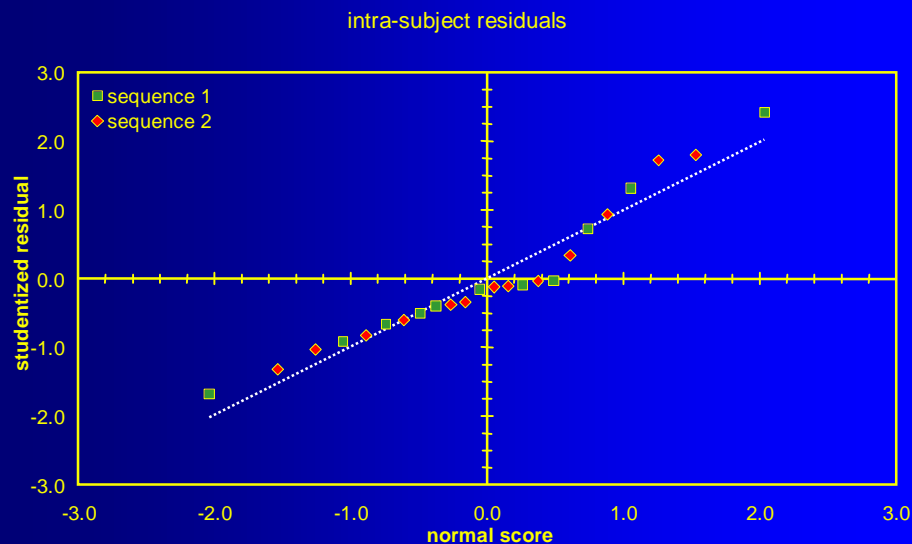
Outliers



Outliers



Outliers



Outliers

● Strategies / Solutions

- Be prepared to face the unexpected!
- Examples of drugs/formulations with documented product failures:
 - Drugs sensitive to low pH (gastric resistance!),
 - Monolithic MR products,
 - ...
- Include available information (PK, literature, previous studies) in the protocol
- Develop a statistical contingency plan

Outliers

● Solution I

- Since assumptions of the parametric statistical model are violated, you may apply a statistical method which does not rely on those!
- Drawback: Lacking regulatory acceptance of nonparametric methods in many countries...
 - 😊 WHO (Technical Report Series No. 937, Annex 9, Section 6.8, May 2006)
 - 😊 Japan NIHS (Bioequivalence Studies for Generic Products, Q&A Document, November 2006)
 - 😞 All other regulatory agencies

Outliers

● Solution II

- Stay with the parametric method, but
 - evaluate both the full data set and the reduced data set (outliers excluded) and discuss influence on the outcome of the study.
- In accordance with EMEA's 2006 Q&A #3:
 - Exceptional reasons may justify post-hoc data exclusion [...]. In such a case, the *applicant must demonstrate that the condition stated to cause the deviation is present in the outlier(s) only* and absence of this condition has been investigated using the same criteria for all other subjects.
 - Results of statistical analyses with and without the group of excluded subjects should be provided.

Practically impossible!

Re-testing of subjects

- If you suspect a product failure *of the reference formulation* consider re-testing:
 - The outlying subject should be re-tested with *both* the test and reference
 - Include ≥ 5 subjects, who showed a 'normal' response in the main study (*i.e.*, size of re-tested group ≥ 6 or 20 % of subjects, whichever is larger)
 - If the subject shows a 'normal' response, exclude the value from the main study
 - Although *sometimes* suggested by the FDA, not covered in any GL!

Add-on / Two-Stage Designs

- Sometimes properly planned studies fail due to
 - Pure chance (producer's risk hit)
 - False assumptions about variability and/or T/R-ratio
 - Poor study conduct (increasing variability)
 - 'True' bioinequivalence
- The patient's risk must be preserved
 - Already noticed at Bio-International Conferences (1989, 1992) and guidelines from the 1990s

History / early approaches

- *'The primary concern in bioequivalence assessment is to limit the risk of erroneously accepting bioequivalence. Only statistical procedures which do not exceed the nominal risk of 5% can be approved, and among them the one with the smallest risk of erroneously rejecting bioequivalence should be selected.'**
- Performing a second study and pooling data with the first's not acceptable
- Performing a (much larger) second study and base BE on this study *alone* was (and is) acceptable

* **CPMP Working Party**

Investigation of Bioavailability and Bioequivalence: Note for Guidance
Section 3.6 Data analysis, Document Ref. III/54/89-EN (1 May 1992)

History / early approaches

- However, naïve pooling (*without* α -adjustment) was performed in the past
 - Statistical model modified in order to include a formulation-by-study interaction factor
 - Test for homogeneity of error variances between studies
 - Pooling only acceptable if both tests not significant*

* **H Mellander**

Problems and Possibilities with the Add-On Subject Design, in:

Midha KK, Blume HH (eds.)

Bio-International. Bioavailability, Bioequivalence and Pharmacokinetics
medpharm Scientific Publishers, Stuttgart, pp. 85–90 (1993)



Add-on Designs

- Example (acc. to Canada's 1992+ guidances)
 - Second part in at least 12 subjects
Pooling only allowed if both of two consistency tests not significant ($p > 0.05$)
 - Equality of residual mean squares (F -test) of the two parts. Smaller MSE must be used as the denominator.
Example:
0.01321 (1st part: $n=55$, df 53)
0.01718 (2nd part: $n=14$, df 12)

Add-on Designs

| Sum of Squares Hypothesis | DF | SSE | MSE | F_stat | P_value |
|---------------------------|----|-----------|-----------|----------|---------|
| Sequence | 1 | 0.0271658 | 0.0271658 | 0.173639 | 0.6786 |
| Sequence*Subject | 53 | 8.29185 | 0.15645 | 11.844 | <0.0001 |
| Treatment | 1 | 0.211196 | 0.211196 | 15.9885 | 0.0002 |
| Period | 1 | 0.0271536 | 0.0271536 | 2.05565 | 0.1575 |
| Error | 53 | 0.700088 | 0.0132092 | | |

| Sum of Squares Hypothesis | DF | SSE | MSE | F_stat | P_value |
|---------------------------|----|-----------|-----------|----------|---------|
| Sequence | 1 | 0.0489527 | 0.0489527 | 0.419204 | 0.5295 |
| Sequence*Subject | 12 | 1.4013 | 0.116775 | 6.79641 | 0.0012 |
| Treatment | 1 | 0.0349142 | 0.0349142 | 2.03203 | 0.1795 |
| Period | 1 | 0.0839476 | 0.0839476 | 4.88581 | 0.0472 |
| Error | 12 | 0.206183 | 0.0171819 | | |

$$\hat{F} = \frac{MSE_{large}}{MSE_{small}} = \frac{0.0171819}{0.0132092} = 1.30075$$

$$F_{1-0.05,12,53} = 1.940$$

$$p(\hat{F}) = 0.24595 > 0.05 \quad \checkmark$$

Add-on Designs

- Example (Canada cont'd)
 - Second part in at least 12 subjects. Pooling is only allowed if two consistency tests not significant ($p > 0.05$):
 - Since first test not significant (p 0.246), pool studies
 - Now test for study-by-formulation interaction

Add-on Designs

| Tests of Model Effects | | | | |
|------------------------|----------|----------|---------|---------|
| Hypothesis | Numer_DF | Denom_DF | F_stat | P_value |
| int | 1 | 56.5 | 2144.16 | <0.0001 |
| Study | 1 | 56.5 | 0.0007 | 0.9784 |
| Treatment | 1 | 64.6 | 9.9949 | 0.0024 |
| Treatment*Study | 1 | 64.6 | 0.1156 | 0.7349 |



Bioequivalence Statistics

$$F_{1-0.05,1,64.6} = 3.989$$

User-Specified Confidence Level for CI's = 95.0000
 Percent of Reference to Detect for 2-1 Tests = 20.0%
 A.H.Lower = 0.800 A.H.Upper = 1.250

$$p(0.1156) = 0.7349 > 0.05$$

Formulation variable: Treatment

| | | | | | | |
|--------------|---------|----------|-----|----------|---------|------------|
| Reference: R | LSMean= | 6.088010 | SE= | 0.132921 | GeoLSM= | 440.543718 |
| Test: T | LSMean= | 6.167145 | SE= | 0.132921 | GeoLSM= | 476.822902 |

Difference = 0.0791, Diff_SE= 0.0250, df= 64.6
 Ratio(%Ref) = 108.2351

CI 90% = (103.8061, 112.8531)

CI 95% = (102.9556, 113.7853)

Average bioequivalence shown for confidence=95.00 and percent=20.0.



Add-on Designs

- Example (Canada cont'd)
 - Formulation-by-study interaction not significant (p 0.7349), pooled analysis acceptable
 - No α -adjustment mentioned in 1992 guideline, but recommended in 2010's draft (Bonferroni: 95% CI)
 - Entirely removed from May 2012 guidance
 - 2012 guidance allows group sequential designs instead
 - Group sequential designs allow better control of patient's risk

Problems with α -inflation

- Patient's risk likely is not preserved
 - The probability to obtain at least one significant result with k independent (!) t -tests (at level α) is

$$P(k) = 1 - (1 - \alpha)^k$$

$$P(2) = 1 - (1 - 0.05)^2 = 0.0975$$

- Bonferroni-correction of two studies would mandate calculation of a 95% confidence interval

$$\alpha_{adj} = \alpha / k$$

$$P_{adj}(2) = 1 - (1 - 0.025)^2 = 0.04938 < 0.05$$

- Applicability doubtful since no *independent* tests!

Problems with α -inflation

- Patient's risk (cont'd)
 - For two repeated tests on accumulating data the overall level is $\sim 8\%$ ¹
 - In naïve pooling the variance will be underestimated²
 - Simulations of BE studies (sample sizes 24 – 48, CV_{intra} 19 – 37%, 1 – 3 interim looks) showed empirical α of up to 5.97%³

¹ **Armitage P, McPherson K, and BC Rowe**

Repeated significance tests on accumulating data

J R Statist Soc A 132, 235–44 (1969)

² **Wittes J, Schabenberger O, Zucker D, Brittain E, and M Proschan**

Internal pilot studies I: type I error rate of the naïve t-test

Statistics in Medicine 18, 3481–91 (1999)

³ **Hauck WW, Preston PE, and FY Bois**

A group sequential approach to crossover trials for average bioequivalence

Journal of Biopharmaceutical Statistics 7(1), 87–96 (1997)

Problems with α -inflation

- Patient's risk (cont'd)
 - Simulations of 1 Mio BE studies (12 subjects in 1st study, CV_{intra} 20%, sample size re-estimation based on PE 0.95 and CV_{intra} of 1st study) showed empirical α of 5.84%¹
 - With two repeated tests at 2.94% overall $\alpha \sim 5\%$ ²
 - Derived for tests assuming normally distributed data with known variances. Approximately valid if sample size not too small.

¹ **Potvin D, Diliberti CE, Hauck WW, Parr AF, Schuirmann DJ, and RA Smith**
Sequential design approaches for bioequivalence studies with crossover designs
Pharmaceut Statist 7/4, 245–62 (2008), DOI: 10.1002/pst.294
<http://www3.interscience.wiley.com/cgi-bin/abstract/115805765/ABSTRACT>

² **SJ Pocock**
Group sequential methods in the design and analysis of clinical trials
Biometrika 64, 191–9 (1977)

Recent developments

- Review of guidelines
 - New Zealand (Oct 2001)
 - Sequential Designs
 - Declared in the protocol
 - Maximum sample size *a priori* (≤ 40 !)
 - 'Appropriate statistical tests (e.g., sequential *t*-test)'
 - FDA
 - Sequential Designs: not mentioned in guidances but acceptable (pers. comm. Barbara Davit, Ljubljana, May 2010)
 - EMA (Jan 2010)
 - Sequential Designs: fairly detailed informations given

Two-Stage Design

- EMA GL on BE (2010)
 - Section 4.1.8
 - Initial group of subjects treated and data analysed.
 - If BE not been demonstrated an additional group can be recruited and the results from both groups combined in a final analysis.
 - Appropriate steps to preserve the overall type I error (patient's risk).
 - Stopping criteria should be defined *a priori*.
 - First stage data should be treated as an interim analysis.

Two-Stage Design

- EMA GL on BE (2010)
 - Section 4.1.8 (cont'd)
 - Both analyses conducted at adjusted significance levels (with the confidence intervals accordingly using an adjusted coverage probability which will be higher than 90%). [...] 94.12% confidence intervals for both the analysis of stage 1 and the combined data from stage 1 and stage 2 would be acceptable, but there are many acceptable alternatives and the choice of how much alpha to spend at the interim analysis is at the company's discretion.

Two-Stage Design

- EMA GL on BE (2010)
 - Section 4.1.8 (cont'd)
 - Plan to use a two-stage approach must be pre-specified in the protocol along with the adjusted significance levels to be used for each of the analyses.
 - When analysing the combined data from the two stages, a term for stage should be included in the ANOVA model.
- Russia (2012 draft?)

Sequential Designs

- Have a long and accepted tradition in clinical research (mainly phase III)
 - Based on work by Armitage *et al.* (1969), McPherson (1974), Pocock (1977), O'Brien and Fleming (1979), Lan & DeMets (1983), ...
 - First proposal by Gould (1995) in the area of BE did not get regulatory acceptance in Europe, but
 - stated in Canadian guidance (2012) and EMA's BE guideline (2010).

AL Gould

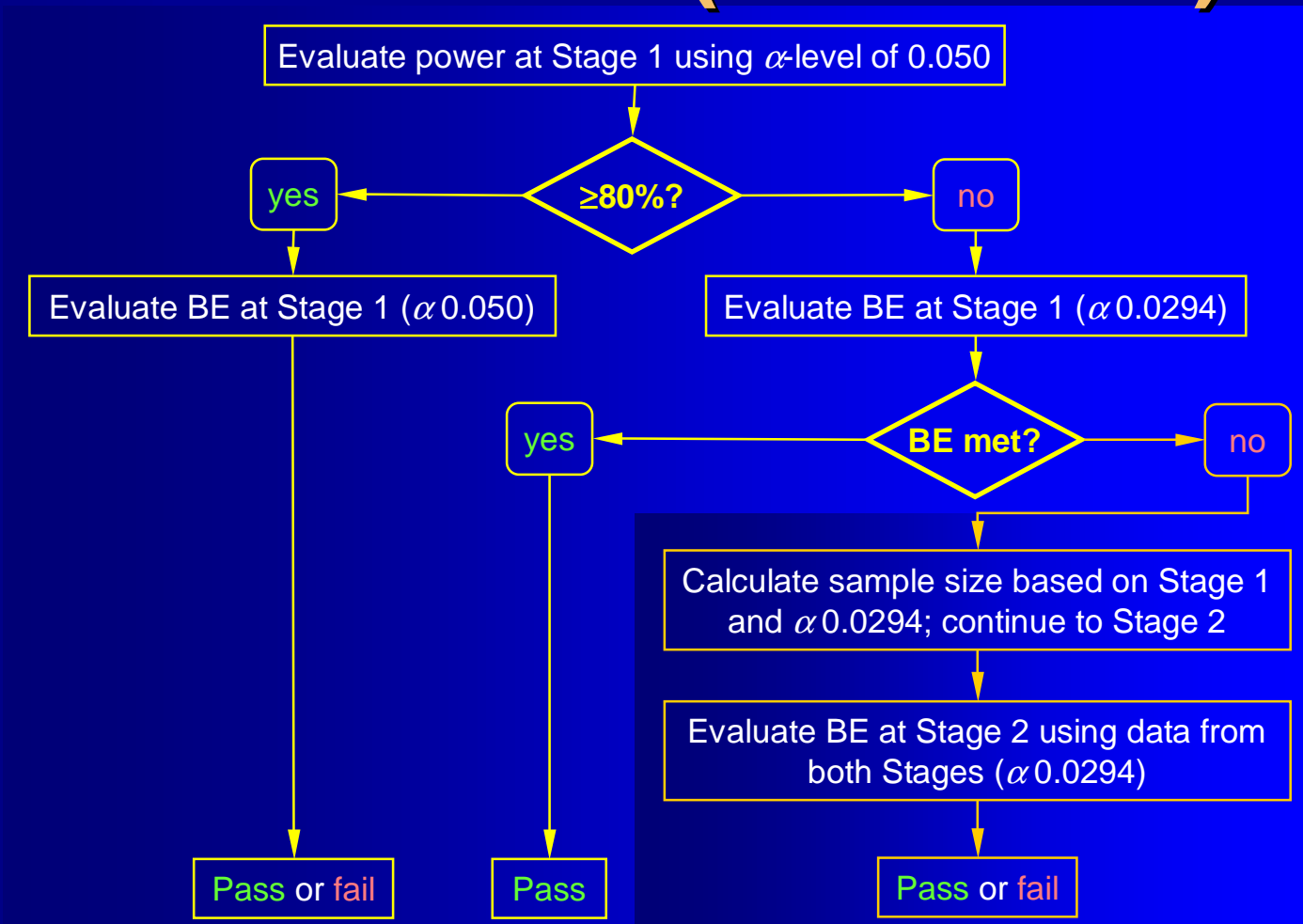
Group Sequential Extension of a Standard Bioequivalence Testing Procedure
J Pharmacokin Biopharm 23/1, 57–86 (1995)

Sequential Designs

- Methods by Potvin *et al.* (2008) promising
 - Supported by 'The Product Quality Research Institute' (members: FDA/CDER, Health Canada, USP, AAPS, PhRMA, ...)
 - Acceptable by US-FDA
 - Canada guidance (May 2012)
 - Acceptable as a Two-Stage Design in the EU
 - Three of BEBAC's protocols approved by German BfArM, one study accepted

Potvin D, Diliberti CE, Hauck WW, Parr AF, Schuirmann DJ, and RA Smith
Sequential design approaches for bioequivalence studies with crossover designs
Pharmaceut Statist 7/4, 245–62 (2008), DOI: 10.1002/pst.294
<http://www3.interscience.wiley.com/cgi-bin/abstract/115805765/ABSTRACT>

Potvin *et al.* (Method C)



Potvin *et al.* (Method C)

- Technical Aspects

- Only *one* Interim Analysis (after Stage 1)
- If possible, use software (too wide step sizes in Diletti's tables), preferable the exact method (avoid approximations)
- Should be termed 'Power Analysis' *not* 'Bioequivalence Assessment' in the protocol
- No *a-posteriori* Power – only a validated method in the decision tree
- No adjustment for the PE observed in Stage 1

Potvin *et al.* (Method C)

- Technical Aspects (cont'd)
 - No stop criterion (*'futility rule'*) preventing to go into Stage 2 with a very high sample size! Must be clearly stated in the protocol (unfamiliar to the IEC because common in Phase III).
 - If power <80% in Stage 1 or in the pooled analysis (data from Stages 1 + 2), Pocock's α 0.0294 is used (*i.e.*, the $1 - 2 \times \alpha = 94.12\%$ CI is calculated)
 - Overall patient's risk preserved at $\sim \leq 0.05$

Potvin *et al.* (Method C)

- Technical Aspects (cont'd)

- If the study is stopped after Stage 1, the (conventional) statistical model is:

```
fixed: sequence + period + treatment
random: subject(sequence)
```

- If the study continues to Stage 2, the model for the combined analysis is:

```
fixed: sequence + stage + period(stage) + treatment
random: subject(sequence × stage)
```

- No poolability criterion!

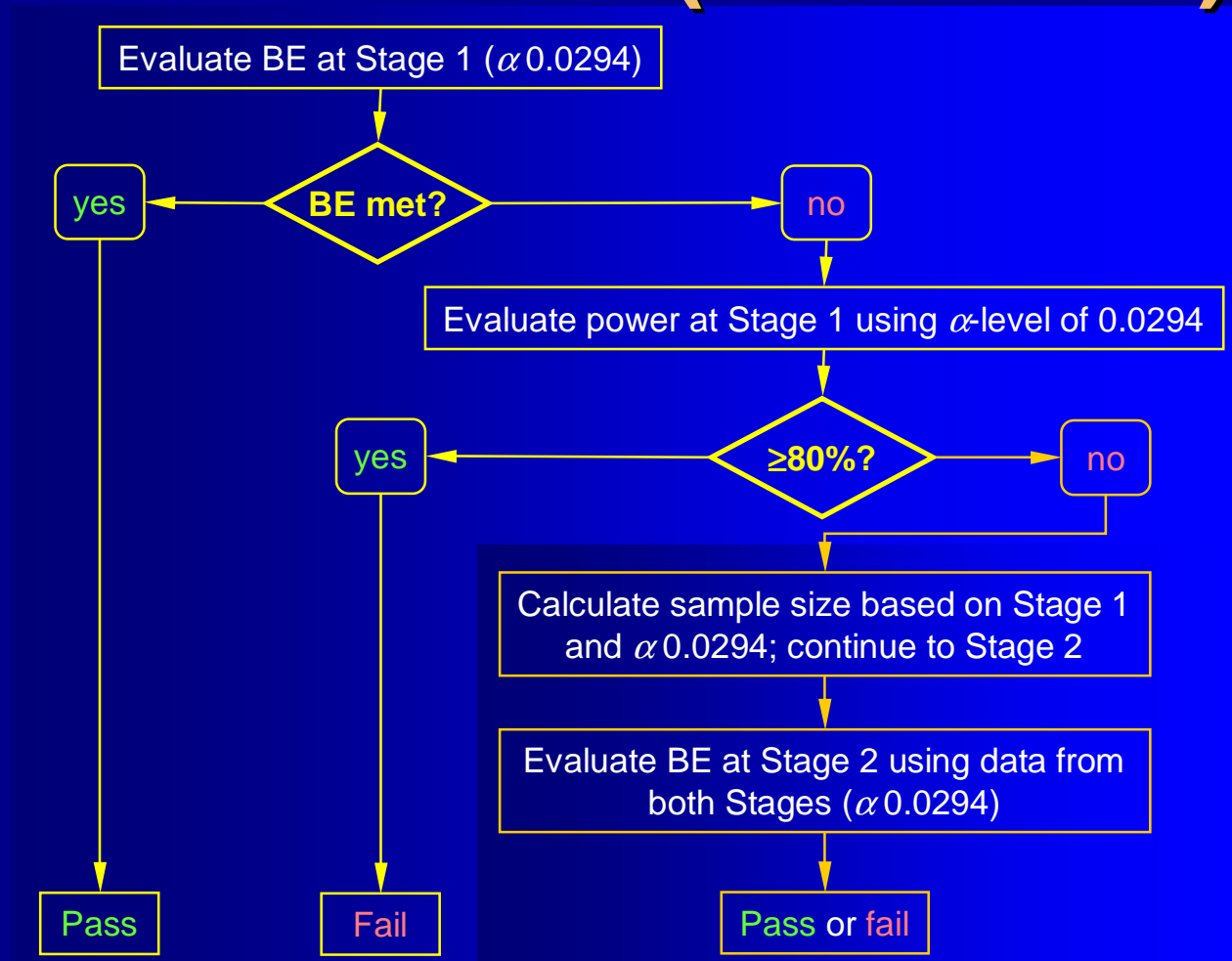
Combining is *always allowed* – even if a significant difference between Stages is observed

Potvin *et al.* (Method C)

- Technical Aspects (cont'd)
 - Potvin *et al.* used a simple approximative power estimation based on the shifted t -distribution (to increase speed in their simulations?)
 - If possible use the exact method (Owen; *R* package *PowerTOST* exact = 'TRUE') or at least the one based on the noncentral t -distribution (*PowerTOST* exact = 'FALSE')
 - Power obtained in Stage 1:

| method | power |
|---------------------------|--------|
| approx. (shifted t) | 64.94% |
| approx. (noncentral t) | 66.45% |
| exact | 66.47% |

Potvin *et al.* (Method B)



Potvin *et al.* (example B/C)

Model Specification and User Settings

Dependent variable : Response

Transform : LN

Fixed terms : int+Sequence+Treatment+Period

Random/repeated terms : Sequence*Subject

12 subjects in Stage 1,
conventional BE model

Final variance parameter estimates:

Var(Sequence*Subject) 0.408682

Var(Residual) 0.0326336

Intrasubject CV

0.182132

CV_{intra} 18.2%

Bioequivalence Statistics

User-Specified Confidence Level for CI's = 94.1200

Percent of Reference to Detect for 2-1 Tests = 20.0%

A.H.Lower = 0.800 A.H.Upper = 1.250

Reference: Reference LSMean= 0.954668 SE= 0.191772 GeoLSM= 2.597808

Test: Test LSMean= 1.038626 SE= 0.191772 GeoLSM= 2.825331

Difference = 0.0840, Diff_SE= 0.0737, df= 10.0

Ratio(%Ref) = 108.7583

α 0.0294
(if power <80%)

Classical

CI 90% = (95.1474, 124.3162)

CI User = (92.9291, 127.2838)

Failed 90% CI (if power \geq 80%)
and 94.12% CI (if power <80%)

Failed to show average bioequivalence for confidence=94.12 and percent=20.0.

Potvin et al. (example B/C)

```
require(PowerTOST)
power.TOST(alpha=0.05, logscale=TRUE,
            theta1=0.8, theta2=1.25, theta0=0.95,
            cv=0.182132, n=12,
            design = "2x2", exact = TRUE)
```

α 0.05 (C), α 0.0294 (B), expected ratio 95% – *not* 108.76% obs. in stage 1! CV_{intra} 18.2%, 12 subjects in Stage 1

[1] 0.6646934

Power 66.5% – initiate Stage 2

```
sampleN.TOST(alpha=0.0294, targetpower=0.80, logscale=TRUE,
             theta1=0.8, theta2=1.25, theta0=0.95,
             cv=0.182132, design = "2x2", exact = TRUE,
             print = TRUE)
```

Calculate total sample size:
expected ratio 95%, CV_{intra} 18.2%,
80% power

```
+++++ Equivalence test - TOST +++++
      Sample size estimation
```

```
-----
Study design: 2x2 crossover
log-transformed data (multiplicative model)
```

```
alpha = 0.0294, target power = 0.8
BE margins      = 0.8 ... 1.25
Null (true) ratio = 0.95,  CV = 0.182132
```

```
Sample size
n      power
20     0.829160
```

Total sample size 20: include another 8 for Stage 2

Potvin *et al.* (example B/C)

Model Specification and User Settings

Dependent variable : Cmax (ng/mL)

Transform : LN

Fixed terms : int+Sequence+Stage+Period(Stage)+Treatment

Random/repeated terms : Sequence*Stage*Subject

8 subjects in Stage 2 (20 total),
modified model for pooled analysis

Final variance parameter estimates:

Var(Sequence*Stage*Subject) 0.518978

Var(Residual) 0.0458956

Intrasubject CV 0.216714

Bioequivalence Statistics

User-Specified Confidence Level for CI's = 94.1200

Percent of Reference to Detect for 2-1 Tests = 20.0%

A.H.Lower = 0.800 A.H.Upper = 1.250

Formulation variable: Treatment

Reference: Reference LSMean= 1.133431 SE= 0.171385 GeoLSM= 3.106297

Test: Test LSMean= 1.147870 SE= 0.171385 GeoLSM= 3.151473

Difference = 0.0144, Diff_SE= 0.0677, df= 17.0

Ratio(%Ref) = 101.4544

Classical

CI 90% = (90.1729, 114.1472)

CI User = (88.4422, 116.3810)

Average bioequivalence shown for confidence=94.12 and percent=20.0.

α 0.0294 in
pooled analysis

BE shown with 94.12% CI;
overall $\alpha \leq 0.05$!

Potvin *et al.* (B vs. C)

● Pros & cons

- Method C (*if power $\geq 80\%$!*) is a conventional BE study; no penalty in terms of α needs to be applied
- Method C goes to Stage 2 less often and has smaller average total sample sizes than Method B for cases where the initial sample size is reasonable for the *CV*
- If the size of Stage 1 is low for the actual *CV* both methods go to Stage 2 almost all the time; total sizes are similar
- Method B slightly more conservative than C

Potvin *et al.* (B vs. C)

● Recommendations

- Method C preferred due to slightly higher power than method B
- Plan the study *as if* the *CV* is known
 - If assumptions turn out to be true = no penalty
 - If lower power (CV_{intra} higher than expected), BE still possible in first stage (penalty; 94.12% CI) or continue to stage 2 as a 'safety net'.
- Don't jeopardize! Smaller sample sizes in the first stage than in a fixed design don't pay off. Total sample sizes are ~20% higher.

Sequential Designs

- Methods by Potvin *et al.* (2008) limited to point estimate of 0.95 and 80% power
 - Follow-up paper 2011
 - Slight inflation of patient's risk (α 0.0547) observed in Methods B/C if PE 0.90 instead of 0.95 was used
 - Method D (like C, but α 0.0280 instead of α 0.0294)
 - Might be usefull if PE 0.95 and power 90% as well; *not validated yet!*

Montague TH, Potvin D, DiLiberti CE, Hauck WW, Parr AF, and DJ Schuirmann
Additional results for 'Sequential design approaches for bioequivalence studies
with crossover designs'
Pharmaceut Statist 11/1, 8–13 (2011), DOI: [10.1002/pst.483](https://doi.org/10.1002/pst.483)

Sequential Designs

● Caveats

- Methods for 'classical' group-sequential designs derived based on
 - Test for differences (superiority, parallel groups)
 - Large samples (Z test of normal distributed data with known variance)
 - Fixed total sample size (interim analysis at N/k)
 - Balanced case (no drop outs)
- Don't apply any published procedure unquestioned (*i.e.*, if not validated for bioequivalence)
- *Simulations mandatory* to derive an empirical $\alpha (\leq 0.052)$!

Open Issues

- Feasibility / futility rules

- It would be desirable to stop a study after stage 1 under certain circumstances

- (1) BE is unlikely to be shown in even very high sample sizes (e.g., CI outside acceptance range)
→ reformulate
- (2) It turns out that the drug/formulation is highly variable
→ replicate design study in order to perform scaling required
- (3) The calculated sample size exceeds the budget of the project by far

Open Issues

- Feasibility / futility rules
 - These issues are not covered by Potvin *et al.* and Montague *et al.*
 - If you decide to include a rule for early stopping, this is not part of the statistical procedure any more
 - (1) and (2) are ethically justifiable
 - (3) Acceptance?

Open Issues

- Arbitrary PE and/or power
 - Simulations mandatory
 - Set desired PE and power
 - Define maximum α -inflation (≤ 0.052 ?)
 - Simulate sufficiently large number of studies (N)
 - Count number of studies accepted BE at 1.25 (n_1) and number of studies rejected BE at the desired PE (n_2)
 - Empirical $\alpha = n_1/N$
 - Empirical $\beta = n_2/N$; power = $1 - \beta$
 - Start with Pocock's nominal α 0.0294 and decrease stepwise if empirical α too high
 - Compiled language required (speed!)

Open Issues

- Adaption for stage 1's PE (full adaptive design)
 - If applied naïvely, α -inflation of up to 30%! ¹
 - Various methods for superiority trials; only one recent publication in the BE context ²
 - Simulations mandatory; no code in public domain (fast language required: Fortran, MATLAB, C/C++)

¹ **Cui L, Hung MJ, and S-J Wang**

Modification of sample size in group sequential clinical trials
Biometrics 55, 853–7 (1999)

² **A Fuglsang**

Controlling type I errors for two-stage bioequivalence study designs
Clinical Research and Regulatory Affairs 28(4), 100–5 (2011)



Open Issues

- Dropping a candidate formulation from a higher-order cross-over design

| Stage 1 | | | Stage 2 | |
|---------|-------|-------|---------|-------|
| I | II | III | I | II |
| T_1 | T_2 | R | R | T_2 |
| T_2 | R | T_1 | T_2 | R |
| R | T_1 | T_2 | ... | ... |
| T_1 | R | T_2 | | |
| T_2 | T_1 | R | | |
| R | T_2 | T_1 | | |
| ... | ... | ... | | |

How to
decide *which*
formulation to drop?

- Statistical model of BE assumes IID (common σ^2)

- Let's assume to continue with T_2
- If $\sigma^2_{T_1} > \sigma^2_{T_2}$ and/or σ^2_R , the pooled variance in Stage 1 will be inflated. The estimated total sample size will be too high. Expensive, but no influence on α expected.
- If $\sigma^2_{T_1} < \sigma^2_{T_2}$ and/or σ^2_R , power will be lower – increasing the producer's risk only.

Don't try this at home!

- 6x3 dose proportionality study
R 20 mg, T_1 30 mg, T_2 40 mg; CV_{intra} 8.76%
- $T_2 \div 2$, all effects fixed (EMA), Method D_B , PE 90%, α 0.028

| Stage 1 | | | | | |
|---------|--------|-------|--------|-------|--------|
| I | | II | | III | |
| R | 162.28 | T_2 | 153.44 | T_1 | 235.62 |
| T_1 | 75.34 | R | 72.04 | T_2 | 43.82 |
| T_2 | 64.38 | T_1 | 78.18 | R | 71.28 |
| T_1 | 124.06 | T_2 | 49.06 | R | 86.42 |
| T_2 | 100.22 | R | 97.72 | T_1 | 121.36 |
| R | 32.30 | T_1 | 63.87 | T_2 | 70.33 |
| T_1 | 118.74 | R | 42.25 | T_2 | 65.97 |
| T_2 | 66.07 | T_1 | 69.52 | R | 38.30 |

| Stage 2 | | | |
|---------|-------|-------|-------|
| I | | II | |
| R | 80.23 | T_2 | 64.26 |
| T_2 | 65.72 | R | 67.91 |
| T_2 | 19.61 | R | 20.81 |
| R | 32.55 | T_2 | 29.35 |

Extremely imbalanced due to arbitrary cut of original dataset!
N=6 (single balanced block) would have zero df for sequences.

Don't try this at home!

Model Specification and User Settings

Dependent variable : Response

Transform : LN

Fixed terms : int+sequence+treatment+period+subject(sequence)

8 subjects in Stage 1,
all effects fixed (EMA)

Final variance parameter estimates:

Var(Residual) 0.0811756

CV_{intra} 8.13%

Bioequivalence Statistics

User-Specified Confidence Level for CI's = 94.4000

α 0.028 (Method B/D)

Percent of Reference to Detect for 2-1 Tests = 20.0%

A.H.Lower = 0.800 A.H.Upper = 1.250

Reference: Reference LSMean= 4.263887 SE= 0.103103 GeoLSM= 71.085745

Test: Test 1 LSMean= 4.686177 SE= 0.103103 GeoLSM= 108.437840

Difference = 0.4223, Diff_SE= 0.1436, df= 12.0

Ratio(%Ref) = 152.5451

CI User = (112.5795, 206.6985)

Failed to show average bioequivalence for confidence=94.40 and percent=20.0.

Test: Test 2 LSMean= 4.318248 SE= 0.103103 GeoLSM= 75.056997

Difference = 0.0544, Diff_SE= 0.1436, df= 12.0

Ratio(%Ref) = 105.5866

CI User = (77.9237, 143.0697)

Failed to show average bioequivalence for confidence=94.40 and percent=20.0.

Don't try this at home!

```
require(PowerTOST)
power.TOST(alpha=0.0280, logscale=TRUE,
            theta1=0.8, theta2=1.25, theta0=0.90,
            cv=se2cv(0.0811756), n=8,
            design="3x6x3", exact=TRUE)
```

α 0.028, expected ratio 90%,
MSE 0.08118 (CV_{intra} 8.13%),
8 subjects in Stage 1, 6x3 design

[1] 0.7776753

Power 77.8% <80% – initiate Stage 2

```
sampleN.TOST(alpha=0.0280, targetpower=0.80, logscale=TRUE,
             theta1=0.8, theta2=1.25, theta0=0.90,
             cv=se2cv(0.0811756), design="3x6x3", exact=TRUE,
             print=TRUE)
```

Calculate total sample size:
expected ratio 90%, CV_{intra} 8.13%,
80% power, keeping 6x3 design

```
+++++ Equivalence test - TOST +++++
      Sample size estimation
```

```
-----
Study design: 3x6x3 crossover
log-transformed data (multiplicative model)
```

```
alpha = 0.0294, target power = 0.8
BE margins      = 0.8 ... 1.25
Null (true) ratio = 0.9,  CV = 0.08130951
```

```
Sample size
n      power
12     0.930078
```

Total sample size 12: include another 4 for Stage 2

Don't try this at home!

Model Specification and User Settings

Dependent variable : Response

Transform : LN

Fixed terms : int+Sequence+Stage+Period(Stage)+Treatment

Random/repeated terms : Sequence*Stage*Subject

Final variance parameter estimates:

Var(Residual) 0.0985763

4 subjects in Stage 2 (12 total),
modified model for pooled analysis

Bioequivalence Statistics

User-Specified Confidence Level for CI's = 94.4000

Percent of Reference to Detect for 2-1 Tests = 20.0%

A.H.Lower = 0.800 A.H.Upper = 1.250

Reference: Reference LSMean= 3.888945 SE= 0.216489 GeoLSM= 48.859311

Test: Test 1 LSMean= 4.284496 SE= 0.229396 GeoLSM= 72.565947

Difference = 0.3956, Diff_SE= 0.1256, df= 14.825

Ratio(%Ref) = 148.5202

CI User = (114.4688, 192.7011)

Failed to show average bioequivalence for confidence=94.40 and percent=20.0.

Test: Test 2 LSMean= 3.889827 SE= 0.216489 GeoLSM= 48.902424

Difference = 0.0009, Diff_SE= 0.1069, df= 14.825

Ratio(%Ref) = 100.0882

CI User = (80.1937, 124.9182)

Average bioequivalence shown for confidence=94.40 and percent=20.0.

Don't try this at home!

- Lessons learned, open questions
 - Not validated! Don't think about using it at all!
 - Note that due to the massive imbalance the LSM of Test 1 (although not included in Stage 2) changed from Stage 1 in the pooled analysis!
 - Stage 1: 108.44
 - Pooled: 72.57
 - Drug has low CV_{intra} , but high CV_{inter} – Apples and oranges?

| CV% | T ₁ | T ₂ | R | model |
|---------|----------------|----------------|-------|--------|
| Stage 1 | 28.61 | 41.30 | 70.66 | period |
| Stage 2 | – | 82.50 | 85.95 | period |
| Pooled | 28.61 | 56.91 | 65.87 | period |

Don't try this at home!

- Lessons learned, open questions
 - Must use software in the power calculation which can handle the degrees of freedom of a Williams' design in Stage 1 correctly (e.g., *PowerTOST*)
 - Obvious which formulation to drop in this example, but what if formulations are similar in PEs?
Keep the one with smaller CV_{inter} ?
 - Design in the sample size estimation of Stage 2?
 - 3x6 (block size 6 → 12)
 - 2x2 (block size 2 → 10)
 - The latter would have failed in the example

Don't try this at home!

- Lessons learned, open questions
 - Although an tempting idea, not recommended until a statistical decision tree is developed and suitable simulations have shown that the patient's risk is not inflated

Open Issues

- Replicated designs (HVDs/HVDPs)
 - Nothing published yet!
 - Statistical model?
 - Although EMA assumes equal variances of formulations (Q&A document Jan 2010) that does not reflect the 'real world' (quite often $\sigma^2_{WR} > \sigma^2_{WT}$)
 - If you set up simulations allow for different variances of test and reference

Hierarchy of Designs

- The more 'sophisticated' a design is, the more information (in terms of σ^2) we may obtain.

- Hierarchy of designs:

Full replicate (TRTR | RTRT) ↗

Partial replicate (TRR | RTR | RRT) ↗

Standard 2x2 cross-over (RT | RT) ↗

Parallel (R | T)

- Variances which can be estimated:

Parallel: total variance (between + within)

2x2 Xover: + between, within subjects ↗

Partial replicate: + within subjects (reference) ↗

Full replicate: + within subjects (reference, test) ↗



Replicate designs

- Any replicate design can be evaluated according to 'classical' (unscaled) Average Bioequivalence (ABE)
- ABE mandatory if scaling not allowed
 - FDA: $s_{WR} < 0.294$ ($CV_{WR} < 30\%$); use mixed effects model (e.g., SAS Proc MIXED)
 - EMA: $CV_{WR} \leq 30\%$; use all fixed effects model according to 2011's Q&A-document (e.g., SAS Proc GLM)
 - Even if scaling is not intended, replicate design give more informations about formulation(s)

Replicate designs

■ Designs

- Two-sequence three-period

T R T

R T R

Sample size to obtain the same power as a 2x2x2 study: ~75%

- Two-sequence four-period

T R T R

R T R T

Sample size to obtain the same power as a 2x2x2 study: ~50%

- **and many others...** (FDA: TRR|RTR|RRT aka 'partial replicate')

- The statistical model is quite complicated – and dependent on the actual design!

$$X_{ijkl} = \mu \cdot \pi_k \cdot \Phi_l \cdot s_{ij} \cdot e_{ijkl}$$

Application: HVDs/HVDPs

- Highly Variable Drugs / Drug Products ($CV_{WR} > 30\%$)
 - ✓ USA Recommended in product specific guidances. GMR 0.80 – 1.25. Minimum sample size 24.
 - ✓ CAN 2010 draft GL. Scaling for AUC only. No restriction on GMR.
 - ± EU Widening of acceptance range (for C_{max} only: to maximum 69.84% – 143.19%), if CV_{WR} in the study $> 30\%$. GMR 0.80 – 1.25. Demonstration that $CV_{WR} > 30\%$ is not caused by outliers.



Application: HVDs/HVDPs

- All (!) ANDAs submitted to FDA/OGD 2003 – 2005 (1010 studies, 180 drugs)
 - 31% (57/180) highly variable ($CV \geq 30\%$)
 - of these HVDs/HVDPs,
 - 60% due to PK (e.g., first pass metabol.)
 - 20% formulation performance
 - 20% unclear

Davit BM, Conner DP, Fabian-Fritsch B, Haidar SH, Jiang X, Patel DT, Seo PR, Suh K, Thompson CL, and LX Yu

Highly Variable Drugs: Observations from Bioequivalence Data Submitted to the FDA for New Generic Drug Applications

The AAPS Journal 10/1, 148–56 (2008)

<http://www.springerlink.com/content/51162107w327883r/fulltext.pdf>

HVDPs (US/EU)

- Advisory Committee for Pharmaceutical Sciences (ACPS) to FDA (10/2006) on HVDs
- Follow-up papers in 2008 (ref. in API-GLs)
 - Replicate study design [TRR|RTR|RRT]
 - Reference Scaled Average Bioequivalence (RSABE)
 - Minimum sample size 24 subjects
 - GMR restricted to [0.80,1.25]

Haidar SH, Davit B, Chen M-L, Conner D, Lee LM, Li QH, Lionberger R, Makhlouf F, Patel D, Schuirmann DJ, and LX Yu

Bioequivalence Approaches for Highly Variable Drugs and Drug Products

Pharmaceutical Research 25/1, 237–41 (2008)

<http://www.springerlink.com/content/u503p62056413677/fulltext.pdf>

Haidar SH, Makhlouf F, Schuirmann DJ, Hyslop T, Davit B, Conner D, and LX Yu

Evaluation of a Scaling Approach for the Bioequivalence of Highly Variable Drugs

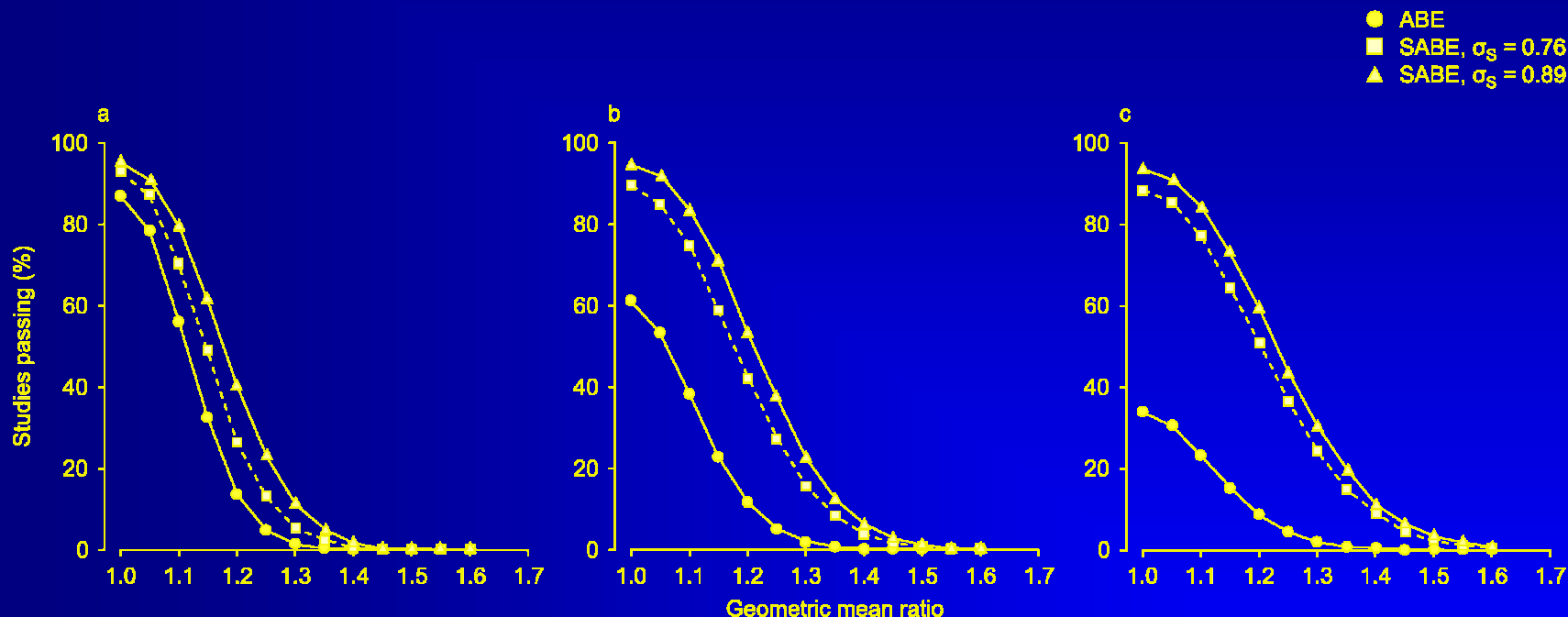
The AAPS Journal, 10/3, (2008) DOI: [10.1208/s12248-008-9053-4](https://doi.org/10.1208/s12248-008-9053-4)

HVDs/HVDPs

- Replicate designs

- 4-period replicate designs:
sample size = $\sim \frac{1}{2}$ of 2×2 study's sample size
- 3-period replicate designs:
sample size = $\sim \frac{3}{4}$ of 2×2 study's sample size
- Reminder: number of treatments (and biosamples)
 \sim conventional 2×2 cross-over
- Allow for a safety margin – expect a higher number
of drop-outs due to the additional period(s)
- Consider increased blood loss (ethics!)
Eventually improved bioanalytics required

HVDPs (US/EU)



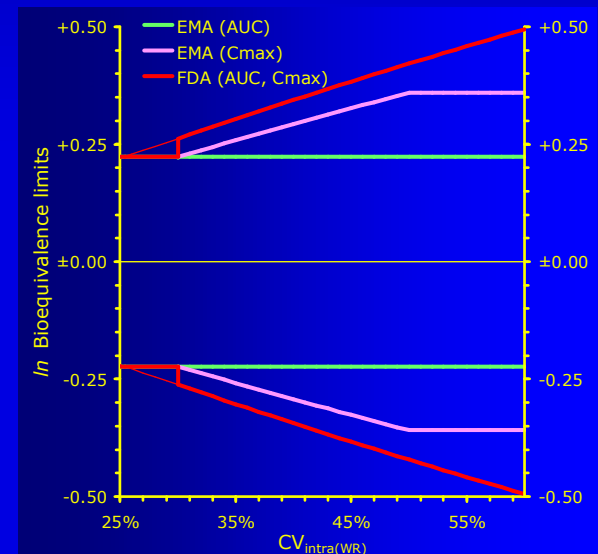
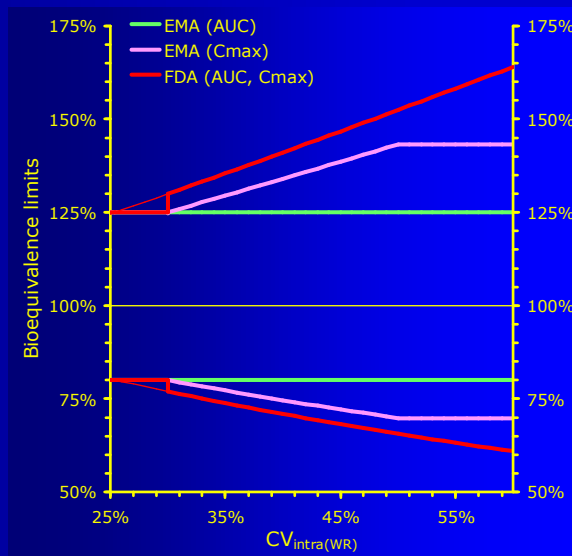
Tóthfalusi *et al.* (2009), Fig. 3

Simulated ($n=10000$) three-period replicate design studies (TRT-RTR) in 36 subjects; GMR restriction 0.80–1.25. (a) CV=35%, (b) CV=45%, (c) CV=55%.

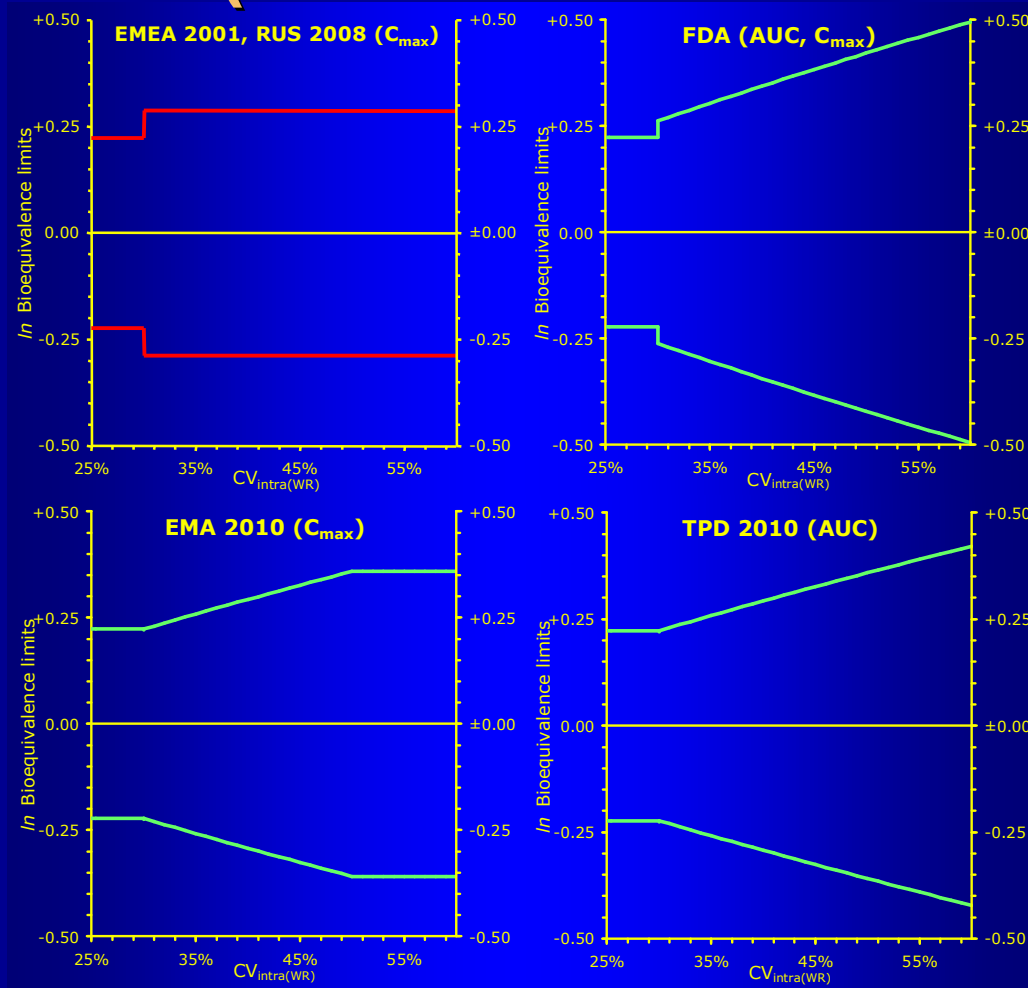
ABE: Conventional Average Bioequivalence, SABE: Scaled Average Bioequivalence, 0.76: EU criterion, 0.89: FDA criterion.

HVDPPs (US/EU)

- FDA's and EMA's approaches differ; FDA's leads to a discontinuity of the acceptance range at $CV = 30\%$, because FDA's scaling CV is 25.83% ($\sigma_{WR} 0.294$) – but to be *applied* at $CV \geq 30\%$.



HVDPs (No Global Harmonization!)



HVDs/HVDPs (Reg. models)

- Common to FDA and EMA

ABE model

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

SABE model

$$-\theta_S \leq \frac{\mu_T - \mu_R}{\sigma_W} \leq +\theta_S$$

Regulatory regulatory switching condition θ_S is derived from the regulatory standardized variation σ_0 (proportionality between acceptance limits in ln-scale and σ_W in the highly variable region).

Tóthfalusi *et al.* (2009)

HVDs/HVDPs (Reg. models)

- Differences between FDA and EMA

FDA: Regulatory regulatory switching condition θ_s is set to 0.893, which would translate into

$$CV_{WR} = 100 \sqrt{e^{\left(\frac{\ln(1.25)}{0.893}\right)^2} - 1} \approx 25.83\%$$

RSABE is allowed only if $CV_{WR} \geq 30\%$ ($s_{WR} \geq 0.294$), which explains to the discontinuity at 30%.

HVDs/HVDPs (Reg. models)

- Differences between FDA and EMA

EMA: Regulatory regulatory switching condition θ_s avoids the discontinuity.

$$CV_W = 0.30$$

$$\sigma_0 = \sqrt{\ln(CV_W^2 + 1)} = 0.2935603792085 \dots$$

$$\theta_s = \frac{\ln(1.25)}{\sigma_0} = -\frac{\ln(0.80)}{\sigma_0} \approx 0.760$$

HVDs/HVDPs (FDA)

- Haidar *et al.* (2008), progesterone guid. (2010)

Starting from the SABE model

$$-\theta_S \leq \frac{\mu_T - \mu_R}{\sigma_W} \leq +\theta_S$$

Rearrangement leads to a linear form

$$(\mu_T - \mu_R)^2 - \theta_S^2 \cdot \sigma_W^2 \leq 0$$

Since we don't have the true parameters, we use estimates

$$E_m = (\mu_T - \mu_R)^2$$

$$E_s = \theta_S^2 \cdot \sigma_W^2$$

HVDs/HVDPs (FDA)

- Haidar *et al.* (2008), progesterone guid. (2010)

Distributions of E_m and E_s are known and their upper confidence limits can be calculated

$$C_m = \left(|m_T - m_R| + t_{\alpha, N-S} \cdot SE \right)^2$$

$$C_s = \frac{\theta_s^2 \cdot (N - S) \cdot s_W^2}{\chi_{\alpha, N-S}^2}$$

t and χ^2 are the inverse cumulative distribution functions at $\alpha 0.05$ and $N - S$ degrees of freedom (N subjects, S sequences). SE is the standard error of the difference between means.

HVDs/HVDPs (FDA)

- Haidar *et al.* (2008), progesterone guid. (2010)
Howe method gets the CL from individual CIs

$$L_m = (C_m - E_m)^2$$

$$L_s = (C_s - E_s)^2$$

$$CL = E_m - E_s + \sqrt{L_m + L_s}$$

The CL of the rearranged SABE criterion ([slide 67](#)) is evaluated at the 95% level. If the upper 95% is positive, RSABE is rejected, and accepted otherwise.

HVDs/HVDPs (EMA)

- EU GL on BE (2010)
 - Average Bioequivalence (ABE) with Expanding Limits (ABEL)
 - The regulatory switching condition θ_S at CV_{WR} 30% would be 0.7601228297680...
 - According to the GL (2010) and the Q&A document (2011, 2012) use k ($\equiv \theta_S$) with 0.760 (*not* the exact value).

HVDs/HVDPs (EMA)

- EU GL on BE (2010)

- If you have σ_{WR} (the *intra*-subject standard deviation of the reference formulation) go to the next step; if not, calculate it from CV_{WR}

$$\sigma_{WR} = \sqrt{\ln(CV_{WR}^2 + 1)}$$

- Calculate the scaled acceptance range based on the regulatory constant k ($\theta_s=0.760$)


$$[L, U] = e^{\mp k \cdot \sigma_{WR}}$$

HVDs/HVDPs (EMA)

- Q&A document (March 2011)
 - Two methods proposed (Method A preferred)
 - **Method A:** All effects fixed; assumes equal variances of test and reference, and no subject-by-formulation interaction; only a common within (*intra*-) subject variance is estimated
 - **Method B:** Similar to A, but random effects for subjects. Common within (*intra*-) subject variance and between (*inter*-) subject variance are estimated.
 - Outliers: Boxplots (of model residuals?) suggested.

*Questions & Answers on the Revised EMA Bioequivalence Guideline
Summary of the discussions held at the 3rd EGA Symposium on Bioequivalence
June 2010, London
http://www.egagenerics.com/doc/EGA_BEQ_Q&A_WEB_QA_1_32.pdf*

HVDs/HVDPs (EMA)

- At higher CVs the GMR is of increasing importance!
- $CV_{WR} > 50\%$ still requires large sample sizes
- No commercial software for sample size estimation can handle the GMR restriction
- Recently sample size tables based on simulations were published
- Expect a solution from the  community

L Tóthfalusi and L Endrenyi

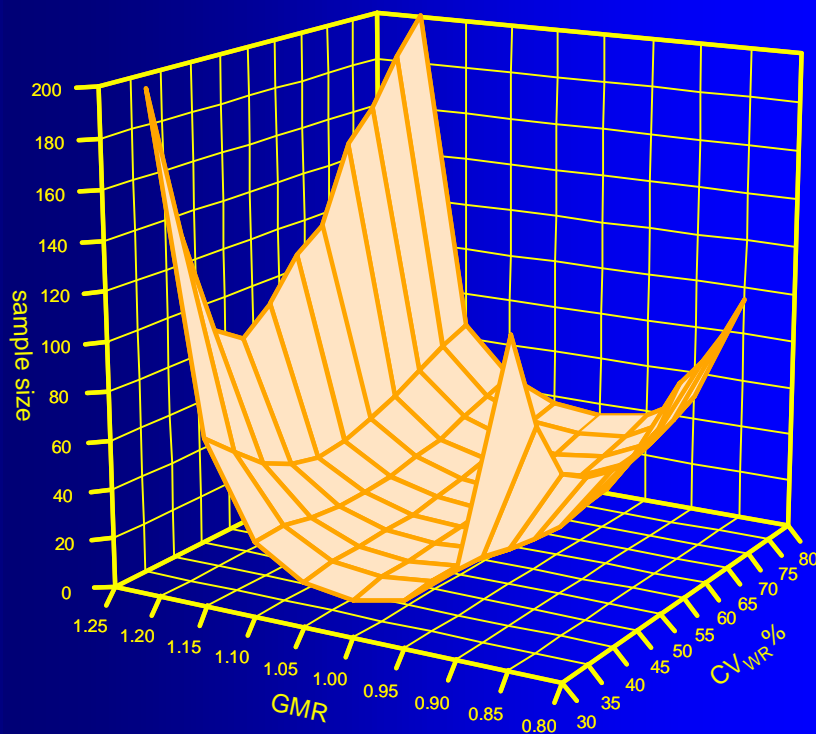
Sample Sizes for Designing Bioequivalence Studies for Highly Variable Drugs

J Pharm Pharmaceut Sci 15(1), 73–84 (2011)

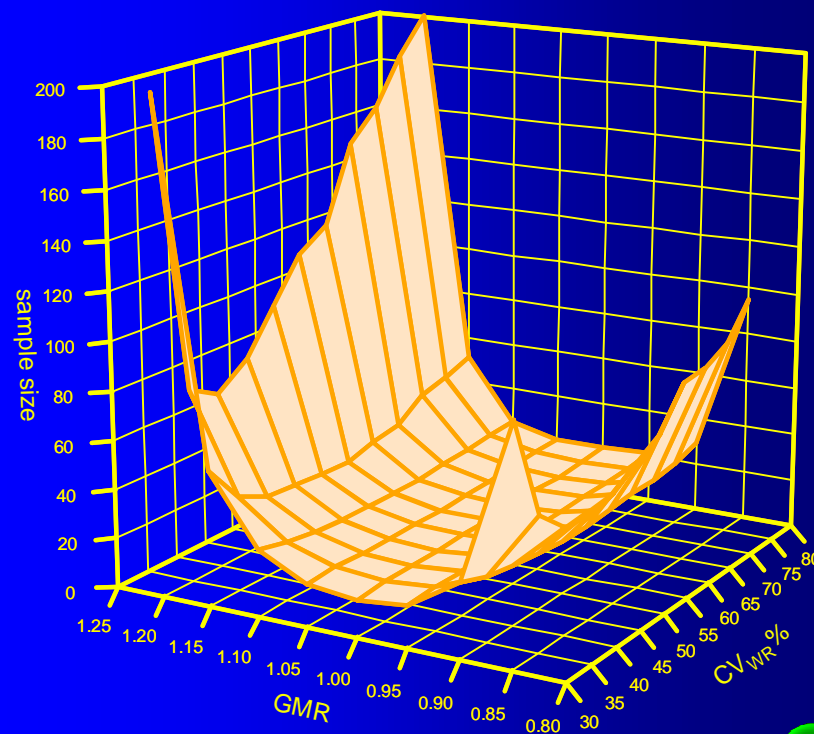
<http://ejournals.library.ualberta.ca/index.php/JPPS/article/download/11612/9489>

HVDPs (US/EU; sample sizes)

EMA-ABEL: Full replicate, 80% power



FDA-RSABE: Full replicate, 80% power



Example datasets (EMA)

- Q&A document (March 2011)

- Data set I

RTTR | TRTR full replicate, 77 subjects, imbalanced, incomplete

- FDA

$s_{WR} 0.446 \geq 0.294 \rightarrow$ apply RSABE (CV_{WR} 46.96%)

a. critbound $-0.0921 \leq 0$ and

b. $80.00\% \leq \text{pointest } 115.46\% \leq 125.00\%$ ✓

- EMA

➤ CV_{WR} 46.96% \rightarrow apply RSABE ($> 30\%$)

➤ Scaled Acceptance Range: 71.23% – 140.40%

➤ A: $71.23\% \leq 107.11\% - 124.89\% \leq 140.40\%$, PE 115.66% ✓

➤ B: $71.23\% \leq 107.17\% - 124.97\% \leq 140.40\%$, PE 115.73% ✓

Example datasets (EMA)

- Q&A document (March 2011)

- Data set II

TRR | RTR | RRT partial replicate, 24 subjects,
balanced, complete

- FDA

s_{WR} $0.114 < 0.294 \rightarrow$ apply ABE (CV_{WR} 11.43%)
 $80.00\% \leq 97.05 - 107.76 \leq 125.00\%$ (CV_{intra} 11.55%) ✓

- EMA

- CV_{WR} 11.17% \rightarrow apply ABE ($\leq 30\%$)
- A: 90% CI 97.32% – 107.46%, PE 102.26% ✓
- B: 90% CI 97.32% – 107.46%, PE 102.26% ✓
- A/B: CV_{intra} 11.86%

Outliers (EMA)

- EU GL on BE (2010), Section 4.1.10
 - The applicant should justify that the calculated intra-subject variability is a reliable estimate and that it is not the result of outliers.
- EGA/EMA Q&A (2010)
 - Q: How should a company proceed if outlier values are observed for the reference product in a replicate design study for a Highly Variable Drug Product (HVDP)?

Outliers (EMA)

- EGA/EMA Q&A (2010)
 - A: The outlier cannot be removed from evaluation [...] but should not be taken into account for calculation of within-subject variability and extension of the acceptance range.
An outlier test is not an expectation of the medicines agencies but outliers could be shown by a box plot. This would allow the medicines agencies to compare the data between them.

Outliers (EMA)

- Data set I (full replicate)

- CV_{WR} 46.96%

- ABEL 71.23% – 140.40%

- Method A: 107.11% – 124.89%

- Method B: 107.17% – 124.97%

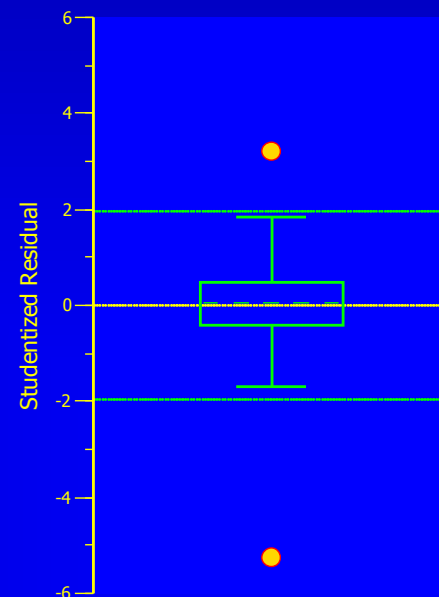
- But there *are* two outliers!

- Excluding subjects 45 and 52

- CV_{WR} drops to 32.16%.

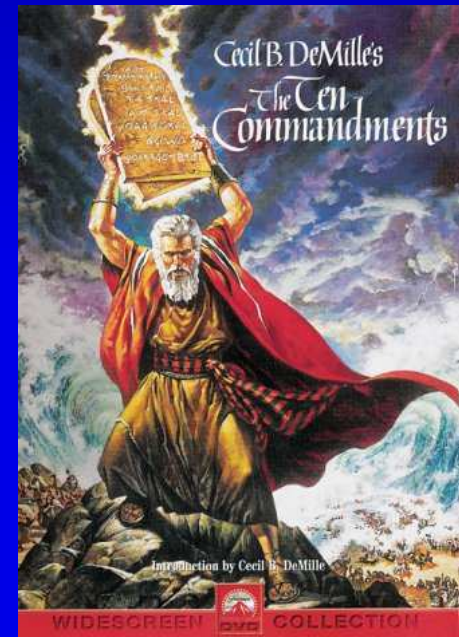
- ABEL 78.79% – 126.93%

- Almost no more gain compared to conventional limits.



Sample Size Estimation

- The estimated *CV* has a certain degree of uncertainty (in the pivotal study it is more likely that we will be able to reproduce the PE, than the *CV*)
 - The smaller the size of the pilot, the more uncertain the outcome
 - The more formulations we have tested, lesser degrees of freedom will result in worse estimates
 - Remember: CV is an *estimate* – *not set in stone!*



Pilot Studies: Sample Size

- Small pilot studies (sample size <12)
 - Are useful in checking the sampling schedule and
 - the appropriateness of the analytical method, but
 - are not suitable for the purpose of sample size planning!
 - Sample sizes (T/R 0.95, power $\geq 80\%$) based on a n=10 pilot study

```
require(PowerTOST)
expsampleN.TOST(alpha=0.05,
  targetpower=0.80, theta1=0.80,
  theta2=1.25, theta0=0.95, CV=0.40,
  dfCV=24-2, alpha2=0.05, design="2x2")
```

| CV% | CV | | ratio |
|-----|-------|-----------|---------------|
| | fixed | uncertain | uncert./fixed |
| 20 | 20 | 24 | 1.200 |
| 25 | 28 | 36 | 1.286 |
| 30 | 40 | 52 | 1.300 |
| 35 | 52 | 68 | 1.308 |
| 40 | 66 | 86 | 1.303 |

If pilot n=24:
n=72, ratio 1.091

Pilot Studies: Sample Size

- Moderate sized pilot studies (sample size ~12–24) lead to more consistent results (both *CV* and PE)
 - If we stated a procedure in your protocol, even BE may be claimed in the pilot study, and no further study will be necessary (US-FDA)
 - If we have some previous hints of high intra-subject variability (>30%), a pilot study size of *at least* 24 subjects is reasonable
 - A Sequential Design may also avoid an unnecessarily large pivotal study

Justification

- Good Scientific Practice!
 - Every influential factor can be *tested* in a pilot study.
 - Sampling schedule: matching C_{max} , lag-time (first point C_{max} problem), reliable estimate of λ_z
 - Bioanalytical method: LLOQ, ULOQ, linear range, metabolite interferences, ICSR
 - Food, posture, ...
 - Variability of PK metrics
 - Location of PE



Justification

- Best description by the FDA (2003)
 - The study can be used to validate analytical methodology, assess variability, optimize sample collection time intervals, and provide other information. For example, for conventional immediate-release products, careful timing of initial samples may avoid a subsequent finding in a full-scale study that the first sample collection occurs after the plasma concentration peak. For modified-release products, a pilot study can help determine the sampling schedule to assess lag time and dose dumping.

Application

- Most common to assess *CV* and PE needed in sample size estimation for a pivotal BE study
 - To select between candidate test formulations compared to one reference
 - To find a suitable reference
 - If design issues (clinical performance, bioanalytics) are already known, a two-stage sequential design would be a better alternative!

Solutions

- *Do not* use the pilot study's *CV*, but calculate an upper confidence interval!
 - Gould recommends a 75% confidence interval (*i.e.*, a producer's risk of 25%).
 - Unless you are under time pressure, a Two-Stage design will help in dealing with the uncertain estimate from the pilot.

LA Gould

Group Sequential Extension of a Standard Bioequivalence Testing Procedure
J Pharmacokin Biopharm 23/1, 57–86 (1995)

Published data

- Literature search for CV
 - Preferably other BE studies (the bigger, the better!)
 - PK interaction studies (Cave: mainly in steady state! Generally lower CV than after SD)
 - Food studies (CV higher/lower than fasted!)
 - If CV_{intra} is not given (quite often!), a little algebra helps. All you need is the 90% geometric confidence interval and the sample size.



Algebra...

● Calculation of CV_{intra} from CI

- Point estimate (PE) from the Confidence Interval

$$PE = \sqrt{CL_{lo} \cdot CL_{hi}}$$

- Estimate the number of subjects / sequence (example 2x2 cross-over)

- If total sample size (N) is an even number, assume (!)

$$n_1 = n_2 = \frac{1}{2}N$$

- If N is an odd number, assume (!)

$$n_1 = \frac{1}{2}N + \frac{1}{2}, n_2 = \frac{1}{2}N - \frac{1}{2} \text{ (not } n_1 = n_2 = \frac{1}{2}N\text{!)}$$

- Difference between one CL and the PE in log-scale; use the CL which is given with more significant digits

$$\Delta_{CL} = \ln PE - \ln CL_{lo} \quad \text{or} \quad \Delta_{CL} = \ln CL_{hi} - \ln PE$$

Algebra...

- Calculation of CV_{intra} from CI (cont'd)
 - Calculate the Mean Square Error (MSE)

$$MSE = 2 \left(\frac{\Delta_{CL}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \cdot t_{1-\alpha, n_1+n_2-2}}} \right)^2$$

- CV_{intra} from MSE as usual

$$CV_{intra} \% = 100 \cdot \sqrt{e^{MSE} - 1}$$

Algebra...

● Calculation of CV_{intra} from CI (cont'd)

- Example: 90% CI [0.91 – 1.15], N 21 ($n_1 = 11$, $n_2 = 10$)

$$PE = \sqrt{0.91 \cdot 1.15} = 1.023$$

$$\Delta_{CL} = \ln 1.15 - \ln 1.023 = 0.11702$$

$$MSE = 2 \left(\frac{0.11702}{\sqrt{\left(\frac{1}{11} + \frac{1}{10} \right) \times 1.729}} \right)^2 = 0.04798$$

$$CV_{intra} \% = 100 \times \sqrt{e^{0.04798} - 1} = 22.2\%$$

Algebra...

● Proof: CI from calculated values

- Example: 90% CI [0.91 – 1.15], N 21 ($n_1 = 11$, $n_2 = 10$)

$$\ln PE = \ln \sqrt{CL_{lo} \cdot CL_{hi}} = \ln \sqrt{0.91 \times 1.15} = 0.02274$$

$$SE_{\Delta} = \sqrt{\frac{2 \cdot MSE}{N}} = \sqrt{\frac{2 \times 0.04798}{21}} = 0.067598$$

$$CI = e^{\ln PE \pm t \cdot SE_{\Delta}} = e^{0.02274 \pm 1.729 \times 0.067598}$$

$$CI_{lo} = e^{0.02274 - 1.729 \times 0.067598} = 0.91$$

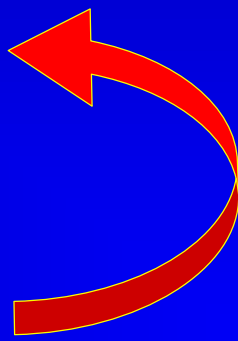
$$CI_{hi} = e^{0.02274 + 1.729 \times 0.067598} = 1.15$$



Sensitivity to Imbalance

- If the study was more imbalanced than assumed, the estimated *CV* is conservative
 - Example: 90% CI [0.89 – 1.15], N 24 ($n_1 = 16$, $n_2 = 8$, but not reported as such); *CV* 24.74% in the study

| Balanced Sequences assumed... | n_1 | n_2 | CV% |
|----------------------------------|-------|-------|-------|
| | 12 | 12 | 26.29 |
| | 13 | 11 | 26.20 |
| | 14 | 10 | 25.91 |
| | 15 | 9 | 25.43 |
| Sequences in study | 16 | 8 | 24.74 |

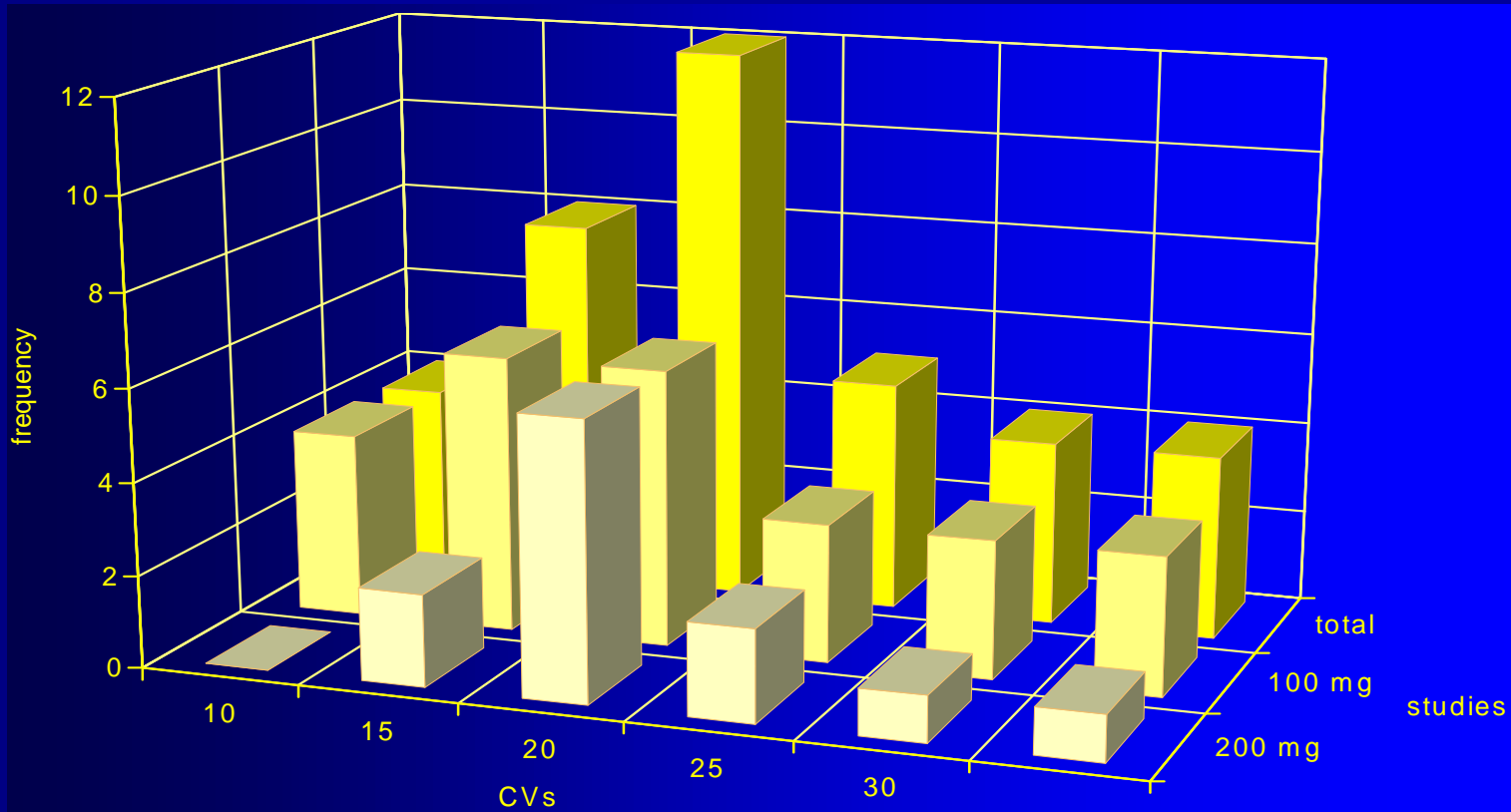


No Algebra...

- Implemented in R-package *PowerTOST*, function *CVfromCI* (not only 2x2 cross-over, but also parallel groups, higher order cross-overs, replicate designs). Previous example:

```
require(PowerTost)
CVfromCI(lower=0.91, upper=1.15, n=21, design="2x2", alpha=0.05)
[1] 0.2219886
```

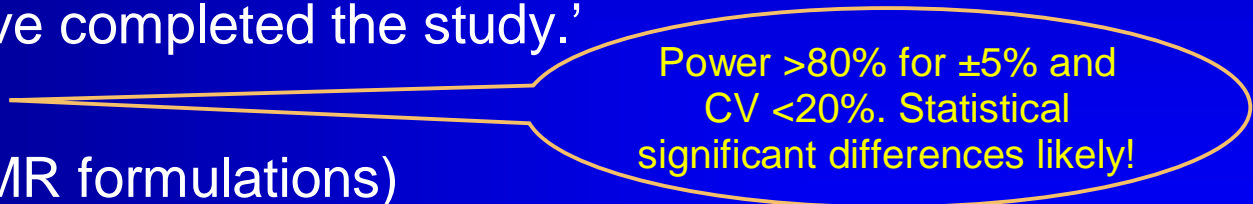
Literature data



Doxycycline (37 studies from **Blume/Mutschler**, *Bioäquivalenz: Qualitätsbewertung wirkstoffgleicher Fertigarzneimittel*, GOVI-Verlag, Frankfurt am Main/Eschborn, 1989-1996)

Sample Size (Limits)

● Minimum

- 12 WHO, EU, CAN, NZ, AUS, AR, MZ, ASEAN States, RSA
- 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** 

Power >80% for $\pm 5\%$ and CV <20%. Statistical significant differences likely!
- 20 RSA (MR formulations)
- 24 Saudia Arabia (12 to 24 if statistically justifiable)
- 24 Brazil
- 'Sufficient number' Japan

Sample Size (Limits)

- Maximum

- NZ: If the calculated number of subjects appears to be higher than is ethically justifiable, it may be necessary to accept a statistical power which is less than desirable. Normally it is not practical to use more than about 40 subjects in a bioavailability study.
- All others: Not specified (judged by IEC/IRB or local Authorities).
ICH E9, Section 3.5 applies: 'The number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed.'

EU

- NfG on the Investigation of BA/BE (2001)
 - The number of subjects required is determined by
 - the error variance associated with the primary characteristic to be studied as estimated from
 - a pilot experiment,
 - previous studies, or
 - published data,
 - the significance level desired,
 - the expected deviation (Δ) from the reference product compatible with BE and,
 - the required power.

EU

● NfG on the Investigation of BA/BE (2001)

■ Problems/solutions

■ ... the error variance associated with the *primary characteristic* to be studied ...

- Since BE must be shown **both** for *AUC* and C_{max} , and,
- if you plan your sample size only for the 'primary characteristic' (e.g., *AUC*), in many cases you will fail for the secondary parameter (e.g., C_{max}), which most likely shows higher variability – your study will be 'underpowered'.
- Based on the assumption, that CV is identical for test and reference (what if only the reference formulation has high variability, e.g., some formulations of PPIs?).

EU

● NfG on the Investigation of BA/BE (2001)

■ Problems/solutions

■ ... as estimated from

- a *pilot experiment*,
- *previous studies*, or
- *published data*,

■ The correct order should read:

1. previous studies →
 2. pilot study →
 3. published data
- Only in the first case you 'know' all constraints resulting in variability
 - Pilot studies are often too small to get *reliable* estimates of variability
 - Advisable only if you have data from a couple of studies

EU

● NfG on the Investigation of BA/BE (2001)

■ Problems/solutions

■ ... the *significance level desired* ...

- Throughout the NfG the significance level (α , error type I: patient's risk to be treated with a bio*ine*quivalent drug) is fixed to 5% (corresponding to a 90% confidence interval)
- You may *desire* a higher significance level, but such a procedure is not considered acceptable
- In special cases (e.g., dose proportionality testing), a correction for multiplicity may be necessary
- In some legislations (e.g., Brazil's ANVISA), α must be tightened to 2.5% for NTIDs (95% confidence interval)

EU

● NfG on the Investigation of BA/BE (2001)

■ Problems/solutions

■ ... the *required power*.

- Generally the power is set to at least 80 % (β , error type II: producers's risk to get no approval for a bioequivalent drug; power = $1 - \beta$).
- If you plan for power of less than 70 %, problems with the ethics committee are likely (ICH E9).
- If you plan for power of more than 90 % (especially with low variability drugs), problems with the regulator are possible ('forced bioequivalence').
- Add subjects ('alternates') according to the expected drop-out rate!

EU

● NfG on the Investigation of BA/BE (2001)

■ Problems/solutions

■ ... the *expected deviation (Δ) from the reference* ...

- Reliable estimate only from a previous full-sized study
- If you are using data from a pilot study, allow for a safety margin
- If no data are available, commonly a GMR (geometric test/reference-ratio) of 0.95 ($\Delta = 5\%$) is used
- If more than $\Delta = 10\%$ is expected, questions from the ethics committee are likely
- **BE GL (2010) batches must not differ more than 5%.**

EU

- EMA BE Guideline (2010)

- The number of subjects to be included in the study should be based on an *appropriate* sample size calculation.

Cookbook?

Hierarchy of Designs

- The more 'sophisticated' a design is, the more information can be extracted.

- Hierarchy of designs:

Full replicate (TRTR | RTRT) ↗

Partial replicate (TRR | RTR | RRT) ↗

Standard 2x2 cross-over (RT | RT) ↗

Parallel (R | T)

- Variances which can be estimated:

Parallel: total variance (between + within)

2x2 Xover: + between, within subjects ↗

Partial replicate: + within subjects (reference) ↗

Full replicate: + within subjects (reference, test) ↗



Coefficient(s) of Variation

- From any design one gets variances of *lower* design levels (only!)
 - Example: Total **CV%** from a 2×2 cross-over used in planning a parallel design study
 - Intra-subject **CV%** (**W**ithin) $\longrightarrow CV_{intra} \% = 100 \cdot \sqrt{e^{MSE_W} - 1}$
 - Inter-subject **CV%** (**B**etween) $\longrightarrow CV_{inter} \% = 100 \cdot \sqrt{e^{\frac{MSE_B - MSE_W}{2}} - 1}$
 - Total **CV%** (**P**ooled) \downarrow $CV_{total} \% = 100 \cdot \sqrt{e^{\frac{MSE_B + MSE_W}{2}} - 1}$

Hauschke D, Steinijans VW and E Diletti

Presentation of the intrasubject coefficient of variation for sample size planning in bioequivalence studies
Int J Clin Pharmacol Ther 32/7, 376-378 (1994)



Coefficient(s) of Variation

- CV_s of *higher* design levels not available.
 - If only mean \pm SD of reference available...
 - Avoid 'rule of thumb' $CV_{intra} = 60\%$ of CV_{total}
 - Don't plan a cross-over based on CV_{total}
 - Examples (cross-over studies)

| drug, formulation | design | n | metric | CV_{intra} | CV_{inter} | CV_{total} | % _{intra/total} |
|--------------------|--------|----|------------|--------------|--------------|--------------|--------------------------|
| methylphenidate MR | SD | 12 | AUC_t | 7.00 | 19.1 | 20.4 | 34.3 |
| paroxetine MR | MD | 32 | AUC_τ | 25.2 | 55.1 | 62.1 | 40.6 |
| lansoprazole DR | SD | 47 | C_{max} | 47.0 | 25.1 | 54.6 | 86.0 |

- ... pilot study unavoidable

Pooling of CV%

- Intra-subject *CV* from different studies can be pooled (LA Gould 1995, Patterson and Jones 2006)
 - In the parametric model of log-transformed data, additivity of variances (not of *CVs*!) apply
 - Do not use the arithmetic mean (or the geometric mean either) of *CVs*
 - Before pooling variances must be weighted according to the studies' sample size and sequences
 - Larger studies are more influential than smaller ones
 - More sequences (with the same n) give higher *CV*



Pooling of CV%

- Intra-subject **CV** from different Xover studies

- Calculate the variance from **CV**

$$\sigma_w^2 = \ln(CV_{\text{intra}}^2 + 1)$$

- Calculate the total variance weighted by *df*

$$\sum \sigma_w^2 df$$

- Calculate the pooled **CV** from total variance

$$CV = \sqrt{e^{\sum \sigma_w^2 df / \sum df} - 1}$$

- Optionally calculate an upper $(1-\alpha)$ % confidence limit on the pooled **CV** (recommended $\alpha = 0.25$)

$$CL_{CV} = \sqrt{e^{\sum \sigma_w^2 df / \chi_{\alpha, \sum df}^2} - 1}$$

Pooling of CV%

- Degrees of freedom of various Xover designs

| Name | df | Name in PowerTOST |
|------------------------------|----------|-------------------|
| 2x2x2 cross over | $n - 2$ | 2x2 |
| 3x3 Latin Squares | $2n - 4$ | 3x3 |
| 6 sequence Williams' design | $2n - 4$ | 3x6x3 |
| 4x4 Latin Squares, Williams' | $3n - 6$ | 4x4 |
| 2x2x3 replicate design | $2n - 3$ | 2x2x3 |
| 2x2x4 replicate design | $3n - 4$ | 2x2x4 |
| 2x4x4 replicate design | $3n - 4$ | 2x4x4 |
| 2x3x3 partial replicate | $3n - 4$ | 2x3x2 |

Pooling of CV%

- Example: 3 studies, different Xover designs

| CV_{intra} | n | seq. | df | σ_W | σ^2_W | $\sigma^2_W \times df$ | | |
|--------------|-----|----------|----|------------|--------------|------------------------|-------------------|---------------------|
| 15% | 12 | 6 | 20 | 0.149 | 0.0223 | 0.4450 | | |
| 25% | 16 | 2 | 14 | 0.246 | 0.0606 | 0.8487 | | |
| 20% | 24 | 2 | 22 | 0.198 | 0.0392 | 0.8629 | σ_{pooled} | σ^2_{pooled} |
| N | 52 | Σ | 56 | | Σ | 2.1566 | 0.196 | 0.0385 |

$$\sqrt{2.1566/56}$$

$$2 \times n - 4$$

$$n - 2$$

$$100\sqrt{e^{0.0385} - 1}$$

| CV_{pooled} | $CV_{g.mean}$ |
|---------------|-------------------|
| 19.81% | 19.57% |

$$100\sqrt{e^{56 \times 0.0385 / 48.546} - 1}$$

| α | $1 - \alpha$ | $\chi^2_{(\alpha, df)}$ | | |
|----------|--------------|-------------------------|--------|-------|
| 0.25 | 0.75 | 48.546 | 21.31% | +7.6% |

Pooling of CV%

- R package *PowerTost* function *CVpooled*, example's data

```
require(PowerTOST)
CVs <- ("
  PKmetric | CV | n | design | source
    AUC    | 0.15 | 12 | 3x6x3 | study 1
    AUC    | 0.25 | 16 | 2x2   | study 2
    AUC    | 0.20 | 24 | 2x2   | study 3
")
txtcon <- textConnection(CVs)
CVdata <- read.table(txtcon, header=TRUE, sep="|",
                     strip.white=TRUE, as.is=TRUE)
close(txtcon)
CVsAUC <- subset(CVdata, PKmetric=="AUC")
print(CVpooled(CVsAUC, alpha=0.25), digits=4, verbose=TRUE)
```

Poolled CV = 0.1981 with 56 degrees of freedom

Upper 75% confidence limit of CV = 0.2131

Pooling of CV%

- Or we may combine pooling with an estimated sample size based on uncertain CV_s (we will see later what that means)

R package *PowerTost* function *expsampleN.TOST*,
data of last example

CV_s and degrees of freedom must be stated as
vectors:

$CV=c(0.15,0.25,0.2)$, $dfCV=c(20,14,22)$

Pooling of CV%

```
require(PowerTOST)
expsampleN.TOST(alpha=0.05,
  targetpower=0.8, theta0=0.95,
  CV=c(0.15,0.25,0.2),
  dfCV=c(20,14,22),
  alpha2=0.25, design="2x2",
  print=TRUE, details=TRUE)
```

```
+++++++ Equivalence test - TOST +++++++
      Sample size est. with uncertain CV
-----
```

```
Study design:  2x2 crossover
Design characteristics:
df = n-2, design const. = 2, step = 2
log-transformed data (multiplicative model)
alpha = 0.05, target power = 0.8
BE margins      = 0.8 ... 1.25
Null (true) ratio = 0.95
Variability data
```

```
      CV df
      0.15 20
      0.25 14
      0.20 22
```

```
CV(pooled)          = 0.1981467 with 56 df
one-sided upper CL = 0.2131329 (level = 75%)
```

```
Sample size search
```

```
      n      exp. power
16     0.733033
18     0.788859
20     0.832028
```

Pooling of CV%

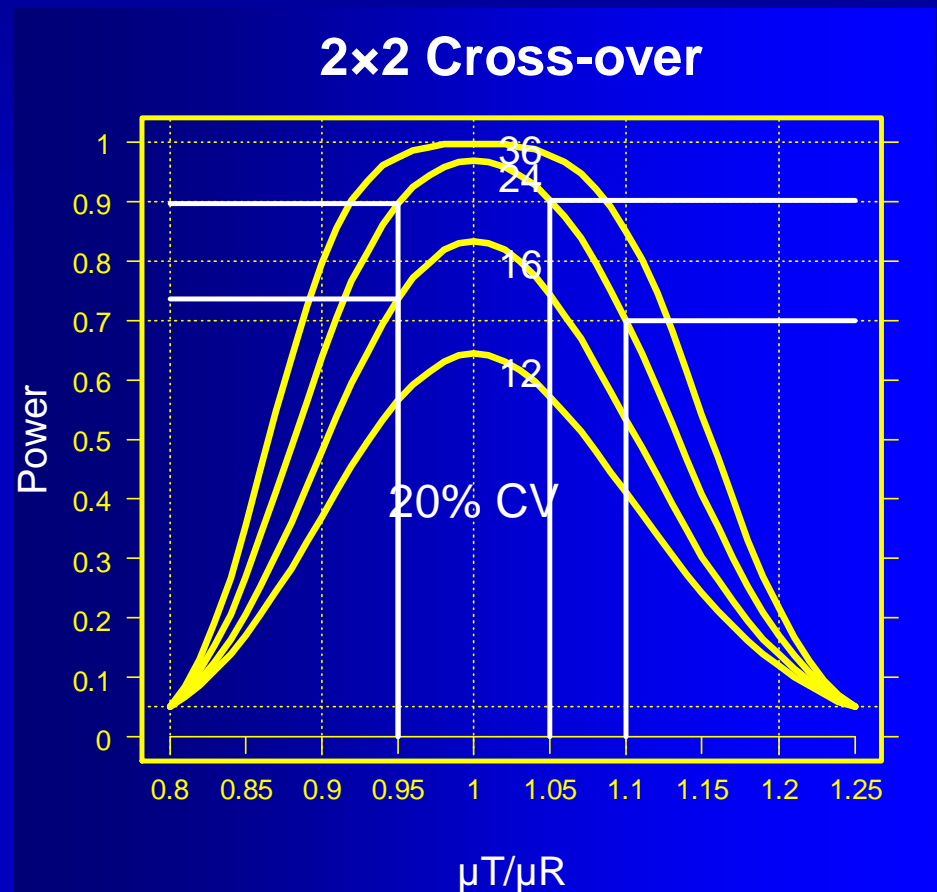
- ‘*Doing the maths*’ is just *part* of the job!
 - Does it make sense to pool studies of different origin and sometimes unknown quality?
 - The reference product may have been subjected to many (*minor only?*) changes from the formulation used in early publications
 - Different bioanalytical methods are applied. Newer (e.g. LC/MS-MS) methods are not *necessarily* better in terms of *CV* (matrix effects!)
 - Generally we have insufficient information about the clinical setup (e.g., posture control)
 - Review studies critically; don’t try to mix oil with water

Power Curves

Power to show BE
with 12 – 36
subjects for
 CV_{intra} 20%

n 24 ↓ 16:
power 0.896 → 0.735

μ_T/μ_R 1.05 ↓ 1.10:
power 0.903 → 0.700



Tools

- Sample Size Tables (Phillips, Diletti, Hauschke, Chow, Julious, ...)
- Approximations (Diletti, Chow, Julious, ...)
- General purpose (SAS, S+, R, StaTable, ...)
- Specialized Software (nQuery Advisor, PASS, FARTSSIE, StudySize, ...)
- Exact method (Owen – implemented in R-package *PowerTOST*)*

* Thanks to Detlew Labes!



Background

- Reminder: Sample Size is not directly obtained – only power
- Solution given by DB Owen (1965) as a difference of two bivariate noncentral t -distributions
 - Definite integrals cannot be solved in closed form
 - ‘Exact’ methods rely on numerical methods (currently the most advanced is AS 243 of RV Lenth; implemented in R, FARTSSIE, EFG). nQuery uses an earlier version (AS 184).

Background

● Power calculations...

- 'Brute force' methods (also called 'resampling' or 'Monte Carlo') converge asymptotically to the true power; need a good random number generator (e.g., Mersenne Twister) and may be time-consuming
- 'Asymptotic' methods use large sample approximations
- Approximations provide algorithms which should converge to the desired power based on the t -distribution

Comparison

CV%

| original values | Method | Algorithm | 5 | 7.5 | 10 | 12 | ### | 14 | 15 | 16 | ### | 18 | 20 | 22 |
|-------------------------------|--------------------|-----------|----|-----|----|----|-----|----|----|----|-----|----|----|----|
| PowerTOST 0.9-8 (2012) | exact | Owen's Q | 4 | 6 | 8 | 8 | 10 | 12 | 12 | 14 | 16 | 16 | 20 | 22 |
| Patterson & Jones (2006) | noncentr. <i>t</i> | AS 243 | 4 | 5 | 7 | 8 | 9 | 11 | 12 | 13 | 15 | 16 | 19 | 22 |
| Diletti <i>et al.</i> (1991) | noncentr. <i>t</i> | Owen's Q | 4 | 5 | 7 | NA | 9 | NA | 12 | NA | 15 | NA | 19 | NA |
| nQuery Advisor 7 (2007) | noncentr. <i>t</i> | AS 184 | 4 | 6 | 8 | 8 | 10 | 12 | 12 | 14 | 16 | 16 | 20 | 22 |
| FARTSSIE 1.6 (2008) | noncentr. <i>t</i> | AS 243 | 4 | 5 | 7 | 8 | 9 | 11 | 12 | 13 | 15 | 16 | 19 | 22 |
| EFG 2.01 (2009) | noncentr. <i>t</i> | AS 243 | 4 | 5 | 7 | 8 | 9 | 11 | 12 | 13 | 15 | 16 | 19 | 22 |
| | brute force | ElMaestro | 4 | 5 | 7 | 8 | 9 | 11 | 12 | 13 | 15 | 16 | 19 | 22 |
| StudySize 2.0.1 (2006) | central <i>t</i> | ? | NA | 5 | 7 | 8 | 9 | 11 | 12 | 13 | 15 | 16 | 19 | 22 |
| Hauschke <i>et al.</i> (1992) | approx. <i>t</i> | | NA | NA | 8 | 8 | 10 | 12 | 12 | 14 | 16 | 16 | 20 | 22 |
| Chow & Wang (2001) | approx. <i>t</i> | | NA | 6 | 6 | 8 | 8 | 10 | 12 | 12 | 14 | 16 | 18 | 22 |
| Kieser & Hauschke (1999) | approx. <i>t</i> | | 2 | NA | 6 | 8 | NA | 10 | 12 | 14 | NA | 16 | 20 | 24 |

CV%

| original values | Method | Algorithm | ### | 24 | 25 | 26 | ### | 28 | 30 | 32 | 34 | 36 | 38 | 40 |
|-------------------------------|--------------------|-----------|-----|----|----|----|-----|----|----|----|----|----|----|----|
| PowerTOST 0.9-8 (2012) | exact | Owen's Q | 24 | 26 | 28 | 30 | 34 | 34 | 40 | 44 | 50 | 54 | 60 | 66 |
| Patterson & Jones (2006) | noncentr. <i>t</i> | AS 243 | 23 | 26 | 28 | 30 | 33 | 34 | 39 | 44 | 49 | 54 | 60 | 66 |
| Diletti <i>et al.</i> (1991) | noncentr. <i>t</i> | Owen's Q | 23 | NA | 28 | NA | 33 | NA | 39 | NA | NA | NA | NA | NA |
| nQuery Advisor 7 (2007) | noncentr. <i>t</i> | AS 184 | 24 | 26 | 28 | 30 | 34 | 34 | 40 | 44 | 50 | 54 | 60 | 66 |
| FARTSSIE 1.6 (2008) | noncentr. <i>t</i> | AS 243 | 23 | 26 | 28 | 30 | 33 | 34 | 39 | 44 | 49 | 54 | 60 | 66 |
| EFG 2.01 (2009) | noncentr. <i>t</i> | AS 243 | 23 | 26 | 28 | 30 | 33 | 34 | 39 | 44 | 49 | 54 | 60 | 66 |
| | brute force | ElMaestro | 23 | 26 | 28 | 30 | 33 | 34 | 39 | 44 | 49 | 54 | 60 | 66 |
| StudySize 2.0.1 (2006) | central <i>t</i> | ? | 23 | 26 | 28 | 30 | 33 | 34 | 39 | 44 | 49 | 54 | 60 | 66 |
| Hauschke <i>et al.</i> (1992) | approx. <i>t</i> | | 24 | 26 | 28 | 30 | 34 | 36 | 40 | 46 | 50 | 56 | 64 | 70 |
| Chow & Wang (2001) | approx. <i>t</i> | | 24 | 26 | 28 | 30 | 34 | 34 | 38 | 44 | 50 | 56 | 62 | 68 |
| Kieser & Hauschke (1999) | approx. <i>t</i> | | NA | 28 | 30 | 32 | NA | 38 | 42 | 48 | 54 | 60 | 66 | 74 |

Approximations

Hauschke *et al.* (1992)

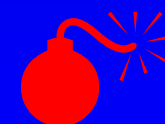
Patient's risk α 0.05, Power 80% (Producer's risk β 0.2), AR [0.80 - 1.25], CV 0.2 (20%), T/R 0.95

1. $\Delta = \ln(0.8) - \ln(T/R) = -0.1719$
2. Start with e.g. $n=8/\text{sequence}$
 1. $df = n \cdot 2 - 1 = 8 \times 2 - 1 = 14$
 2. $t_{\alpha, df} = 1.7613$
 3. $t_{\beta, df} = 0.8681$
 4. new $n = [(t_{\alpha, df} + t_{\beta, df})^2 \cdot (CV/\Delta)]^2 = (1.7613 + 0.8681)^2 \times (-0.2/0.1719)^2 = 9.3580$
3. Continue with $n=9.3580/\text{sequence}$ ($N=18.716 \rightarrow 19$)
 1. $df = 16.716$; roundup to the next integer 17
 2. $t_{\alpha, df} = 1.7396$
 3. $t_{\beta, df} = 0.8633$
 4. new $n = [(t_{\alpha, df} + t_{\beta, df})^2 \cdot (CV/\Delta)]^2 = (1.7396 + 0.8633)^2 \times (-0.2/0.1719)^2 = 9.1711$
4. Continue with $n=9.1711/\text{sequence}$ ($N=18.3422 \rightarrow 19$)
 1. $df = 17.342$; roundup to the next integer 18
 2. $t_{\alpha, df} = 1.7341$
 3. $t_{\beta, df} = 0.8620$
 4. new $n = [(t_{\alpha, df} + t_{\beta, df})^2 \cdot (CV/\Delta)]^2 = (1.7341 + 0.8620)^2 \times (-0.2/0.1719)^2 = 9.1233$
5. Convergence reached ($N=18.2466 \rightarrow 19$):
Use 10 subjects/sequence (20 total)

S-C Chow and H Wang (2001)

Patient's risk α 0.05, Power 80% (Producer's risk β 0.2), AR [0.80 - 1.25], CV 0.2 (20%), T/R 0.95

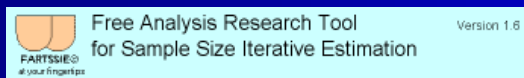
1. $\Delta = \ln(T/R) - \ln(1.25) = 0.1719$
2. Start with e.g. $n=8/\text{sequence}$
 1. $df_{\alpha} = \text{roundup}(2 \cdot n - 2) \cdot 2 - 2 = (2 \times 8 - 2) \times 2 - 2 = 26$
 2. $df_{\beta} = \text{roundup}(4 \cdot n - 2) = 4 \times 8 - 2 = 30$
 3. $t_{\alpha, df} = 1.7056$
 4. $t_{\beta/2, df} = 0.8538$
 5. new $n = \beta^2 \cdot [(t_{\alpha, df} + t_{\beta/2, df})^2 / \Delta^2] = 0.2^2 \times (1.7056 + 0.8538)^2 / 0.1719^2 = 8.8723$
3. Continue with $n=8.8723/\text{sequence}$ ($N=17.7446 \rightarrow 18$)
 1. $df_{\alpha} = \text{roundup}(2 \cdot n - 2) \cdot 2 - 2 = (2 \times 8.8723 - 2) \times 2 - 2 = 30$
 2. $df_{\beta} = \text{roundup}(4 \cdot n - 2) = 4 \times 8.8723 - 2 = 34$
 3. $t_{\alpha, df} = 1.6973$
 4. $t_{\beta/2, df} = 0.8523$
 5. new $n = \beta^2 \cdot [(t_{\alpha, df} + t_{\beta/2, df})^2 / \Delta^2] = 0.2^2 \times (1.6973 + 0.8523)^2 / 0.1719^2 = 8.8045$
4. Convergence reached ($N=17.6090 \rightarrow 18$):
Use 9 subjects/sequence (18 total)



| sample size | 18 | 19 | 20 |
|-------------|--------|--------|--------|
| power % | 79.124 | 81.428 | 83.468 |

Approximations obsolete

- Exact sample size tables still useful in checking the plausibility of software's results
- Approximations based on noncentral t (FARTSSIE17)



<http://individual.utoronto.ca/ddubins/FARTSSIE17.xls>

or  / S+ →

- Exact method (Owen) in R-package *PowerTOST*

<http://cran.r-project.org/web/packages/PowerTOST/>

```
require(PowerTOST)
sampleN.TOST(alpha=0.05,
targetpower=0.80, logscale=TRUE,
theta1=0.80, diff=0.95, cv=0.30,
design="2x2", exact=TRUE)
```

```
alpha <- 0.05      # alpha
CV <- 0.30         # intra-subject CV
theta1 <- 0.80     # lower acceptance limit
theta2 <- 1/theta1 # upper acceptance limit
ratio <- 0.95      # expected ratio T/R
PwrNeed <- 0.80    # minimum power
Limit <- 1000      # Upper Limit for search
n <- 4             # start value of sample size search
s <- sqrt(2)*sqrt(log(CV^2+1))
repeat{
  t <- qt(1-alpha,n-2)
  nc1 <- sqrt(n)*(log(ratio)-log(theta1))/s
  nc2 <- sqrt(n)*(log(ratio)-log(theta2))/s
  prob1 <- pt(+t,n-2,nc1); prob2 <- pt(-t,n-2,nc2)
  power <- prob2-prob1
  n <- n+2 # increment sample size
  if(power >= PwrNeed | (n-2) >= Limit) break }
Total <- n-2
if(Total == Limit){
  cat("Search stopped at Limit",Limit,
      " obtained Power",power*100,"%\n")
} else
  cat("Sample Size",Total,"(Power",power*100,"%)\n")
```

Sensitivity Analysis

- ICH E9 (1998)

- Section 3.5 Sample Size, paragraph 3

- The method by which the sample size is calculated should be given in the protocol [...]. The basis of these estimates should also be given.
 - It is important to investigate the sensitivity of the sample size estimate to a variety of deviations from these assumptions and this may be facilitated by providing a range of sample sizes appropriate for a reasonable range of deviations from assumptions.
 - In confirmatory trials, assumptions should normally be based on published data or on the results of earlier trials.

Sensitivity Analysis

● Example

nQuery Advisor: $\sigma_w = \sqrt{\ln(CV_{intra}^2 + 1)}; \sqrt{\ln(0.2^2 + 1)} = 0.198042$

nQuery Advisor - [MTE2co-1.nqa]

File Edit View Options Assistants Randomize Plot Window Help

t-tests (TOST) of equivalence in ratio of means for crossover design (natural log scale)

| | 90% power | 25% CV | 4 drop outs | 25% CV + d.o. | PE 90% | worst case |
|---|-----------|----------|-------------|---------------|----------|------------|
| Test significance levels, α (one-sided) | 0.050 | 0.050 | 0.050 | 0.050 | 0.050 | 0.050 |
| Lower equivalence limit for $\mu_T / \mu_S, \Delta_L$ | 0.800 | 0.800 | 0.800 | 0.800 | 0.800 | 0.800 |
| Upper equivalence limit for $\mu_T / \mu_S, \Delta_U$ | 1.250 | 1.250 | 1.250 | 1.250 | 1.250 | 1.250 |
| Expected ratio, μ_T / μ_S | 0.950 | 0.950 | 0.950 | 0.950 | 0.900 | 0.900 |
| Crossover ANOVA, $\sqrt{\text{MSE}}$ (ln scale) | 0.198042 | 0.246221 | 0.198042 | 0.246221 | 0.198042 | 0.246221 |
| SD differences, σ_d (ln scale) | 0.280074 | 0.348209 | 0.280074 | 0.348209 | 0.280074 | 0.348209 |
| Power (%) | 90.00 | 77.60 | 86.88 | 69.53 | 66.94 | 45.09 |
| n per sequence group | 13 | 13 | 11 | 11 | 13 | 11 |

20% CV:
n=26

25% CV:
power 90% → **78%**

20% CV, 4 drop outs:
power 90% → **87%**

25% CV, 4 drop outs:
power 90% → **70%**

20% CV, PE 90%:
power 90% → **67%**

Sensitivity Analysis

● Example

PowerTOST, function *sampleN.TOST*

```
require(PowerTost)
sampleN.TOST(alpha=0.05, targetpower=0.9, logscale=TRUE,
             theta1=0.8, theta2=1.25, theta0=0.95, cv=0.2,
             design="2x2", exact=TRUE, print=TRUE)
```

```
+++++++ Equivalence test - TOST ++++++
          Sample size estimation
```

```
-----
Study design:  2x2 crossover
log-transformed data (multiplicative model)
alpha = 0.05, target power = 0.9
BE margins      = 0.8 ... 1.25
Null (true) ratio = 0.95,  CV = 0.2
Sample size
  n      power
26    0.917633
```

Sensitivity Analysis

- To calculate Power for a given sample size, use function *power.TOST*

```
require(PowerTost)
power.TOST(alpha=0.05, logscale=TRUE, theta1=0.8, theta2=1.25,
            theta0=0.95, CV=0.25, n=26, design="2x2", exact=TRUE)
[1] 0.7760553
power.TOST(alpha=0.05, logscale=TRUE, theta1=0.8, theta2=1.25,
            theta0=0.95, CV=0.20, n=22, design="2x2", exact=TRUE)
[1] 0.8688866
power.TOST(alpha=0.05, logscale=TRUE, theta1=0.8, theta2=1.25,
            theta0=0.95, CV=0.25, n=22, design="2x2", exact=TRUE)
[1] 0.6953401
power.TOST(alpha=0.05, logscale=TRUE, theta1=0.8, theta2=1.25,
            theta0=0.90, CV=0.20, n=26, design="2x2", exact=TRUE)
[1] 0.6694514
power.TOST(alpha=0.05, logscale=TRUE, theta1=0.8, theta2=1.25,
            theta0=0.90, CV=0.25, n=22, design="2x2", exact=TRUE)
[1] 0.4509864
```

Sensitivity Analysis

- Must be done *before* the study (*a priori*)
- The Myth of retrospective (*a posteriori* or *post hoc*) Power...
 - High values do not further support the claim of already demonstrated bioequivalence.
 - Low values do not invalidate a bioequivalent formulation.
 - Further reader:
 - RV Lenth
Two Sample-Size Practices that I don't recommend (2000)
 - JM Hoenig and DM Heisey
The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis (2001)
 - P Bacchetti
Current sample size conventions: Flaws, harms, and alternatives (2010)



The Myth of Power

There is simple intuition behind results like these: If my car made it to the top of the hill, then it is powerful enough to climb that hill; if it didn't, then it obviously isn't powerful enough. Retrospective power is an obvious answer to a rather uninteresting question. A more meaningful question is to ask whether the car is powerful enough to climb a particular hill never climbed before; or whether a different car can climb that new hill. Such questions are prospective, not retrospective.

The fact that retrospective power adds no new information is harmless in its own right. However, in typical practice, it is used to exaggerate the validity of a significant result ("not only is it significant, but the test is really powerful!"), or to make excuses for a nonsignificant one ("well, P is .38, but that's only because the test isn't very powerful"). The latter case is like blaming the messenger.



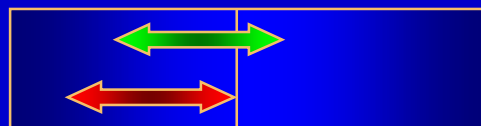
R/V Lenth

Two Sample-Size Practices that I don't recommend

<http://www.math.uiowa.edu/~rlenth/Power/2badHabits.pdf>

Low Variability

- Drugs / Drug Products with $CV_{intra} < 10-15\%$
 - No specific statements in any guideline.
 - Problems may arise according to significant treatment effects in ANOVA (*i.e.*, although the 90% CI is within the acceptance range – 100% is not included) – even for the (Russian) minimum sample size of 18.



- Denmark
 - DKMA considers that the 90% CI for the ratio test versus reference should include 100% [...].
 - Deviations are usually accepted if it can be adequately proved that the deviation has no clinically relevant impact on the efficacy and safety of the medicinal product.

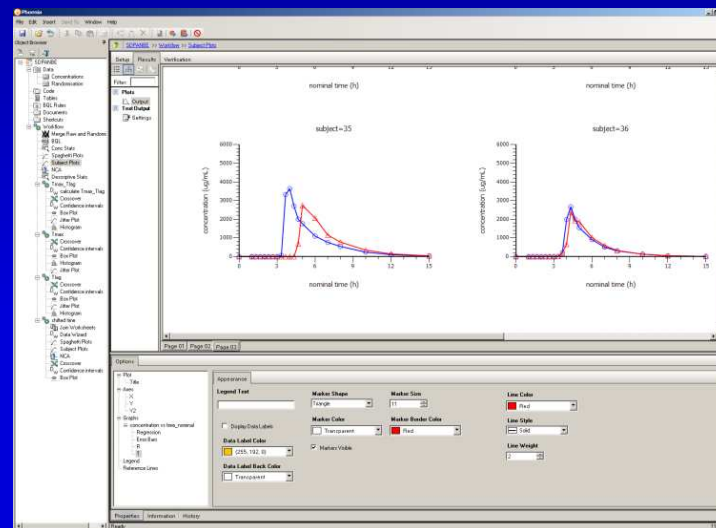
Danish Medicines Agency (DKMA)

Bioequivalence and labelling of medicinal products with regard to generic substitution (13 Jul 2011)

<http://www.dkma.dk/1024/visUKLSArtikel.asp?artikelID=6437>

PK and Statistical Softwares

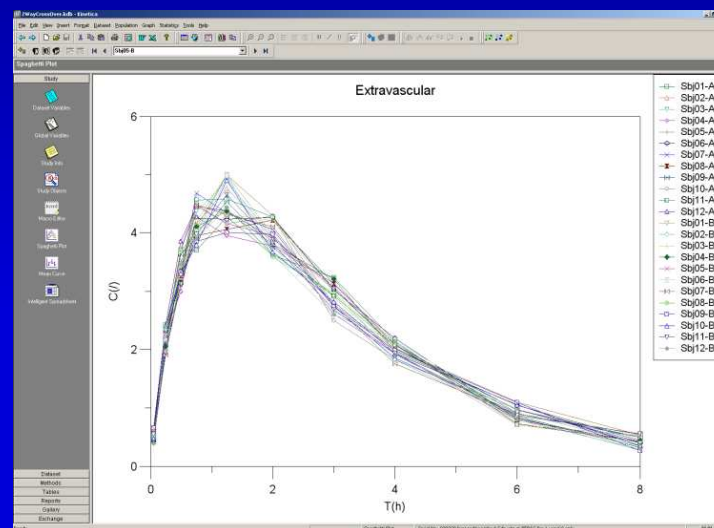
- Phoenix/WinNonlin 1.3 (Pharsight 2012)
 - Supported data formats: WNL, ASCII, XLS, XPT
 - NCA, classical PK/PD modeling, BE (2+ Xover, replicate, nonparametrics), basic deconvolution
 - With additional licenses: Population PK (NLME), IVIVC, remote execution of NONMEM, SPlus, R, SigmaPlot



PK and Statistical Softwares

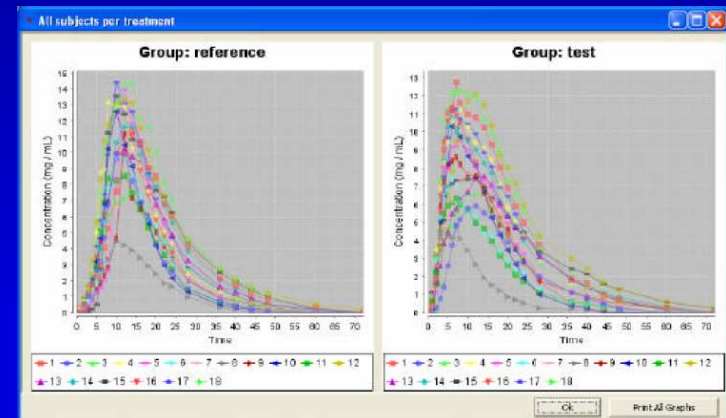
● Kinetica 5 (Thermo Scientific 2007)

- Supported data formats:
ASCII, XLS, ODBC
- NCA, classical PK/PD modeling, BE (only 2way Xover), deconvolution, Population PK, IVIVC
- Integration with other Thermo-products (Watson LIMS)



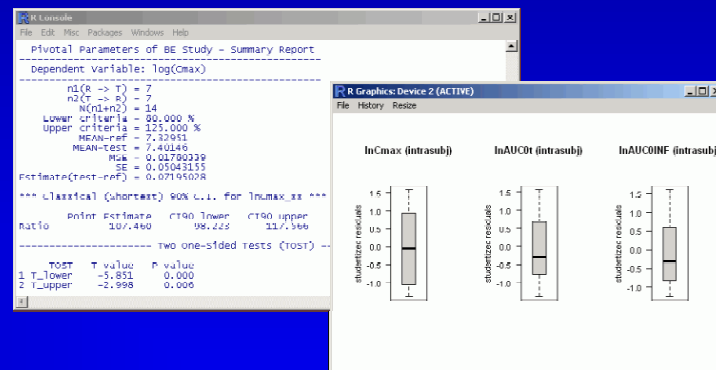
PK and Statistical Softwares

- EquivTest PK (Statistical Solutions 2009)
 - Supported data formats: ASCII, XLS
 - NCA, BE (only 2way Xover), nonparametrics



PK and Statistical Softwares

- Package *bear* 2.5.3 for *R* (2010 Hsin-ya Lee and Yung-jin Lee)
 - Supported data formats: ASCII, XLS, RData
 - NCA, BE (2way Xover, parallel, replicate), exhaustive outlier statistics
 - Basic sample size estimation



Thank You!

**PK–NCA, PK based Design,
Biostatistics**

Open Questions?



Helmut Schütz
BEBAC

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

To bear in Remembrance...

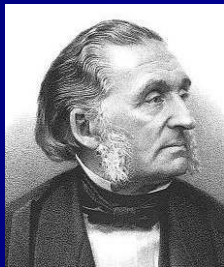
Power. That which statisticians are always calculating but never have.

Power: That which is wielded by the priesthood of clinical trials, the statisticians, and a stick which they use to beta their colleagues.



Power Calculation – A guess masquerading as mathematics.

Stephen Senn



You should treat as many patients as possible with the new drugs while they still have the power to heal.

Armand Trousseau



SAS code (EMA)

Method A

```
proc glm data=replicate;  
  class formulation subject period sequence;  
  model logDATA= sequence subject(sequence) period formulation;  
  estimate "test-ref" formulation -1+1;  
  test h=sequence e=subject(sequence);  
  lsmeans formulation / adjust=t pdiff=control("R") CL alpha=0.10;  
run;
```

Method B

```
proc mixed data=replicate;  
  class formulation subject period sequence;  
  model logDATA= sequence period formulation;  
  random subject(sequence);  
  estimate "test-ref" formulation -1 1 / CL alpha=0.10;  
run;
```

CV_{WR} (both methods)

```
data var;  
  set replicate;  
  if formulation='R';  
run;  
proc glm data=var;  
  class subject period sequence;  
  model logDATA= sequence subject(sequence) period;  
run;
```

SAS code (FDA)

Partial reference-replicated 3-way design

```
data test;
  set pk;
  if trt='T';
  latt=lauct;
run;

data ref1;
  set ref;
  if (seq=1 and per=2) or (seq=2 and per=1) or (seq=3 and per=1);
  lat1r=lauct;
run;

data ref2;
  set ref;
  if (seq=1 and per=3) or (seq=2 and per=3) or (seq=3 and per=2);
  lat2r=lauct;
run;

data ref2;
  set ref;
  if (seq=1 and per=3) or (seq=2 and per=3) or (seq=3 and per=2);
  lat2r=lauct;
run;
```

SAS code (FDA)

Partial reference-replicated 3-way design (cont'd)

```
proc glm data=scavbe;
  class seq;
  model ilat=seq/clparm alpha=0.1;
  estimate 'average' intercept 1 seq 0.3333333333 0.3333333333 0.3333333333;
  ods output overallanova=iglm1;
  ods output Estimates=iglm2;
  ods output NObs=iglm3;
  title1 'scaled average BE';
run;

pointest=exp(estimate);
x=estimate**2-stderr**2;
boundx=(max((abs(LowerCL)), (abs(UpperCL))))**2;

proc glm data=scavbe;
  class seq;
  model dlat=seq;
  ods output overallanova=dglm1;
  ods output NObs=dglm3;
  title1 'scaled average BE';
run;

dfd=df;
s2wr=ms/2;
```

SAS code (FDA)

Partial reference-replicated 3-way design (cont'd)

```
theta=((log(1.25))/0.25)**2;  
y=-theta*s2wr;  
→ boundy=y*dfd/cinv(0.95,dfd);  
SWR=sqrt(s2wr);  
critbound=(x+y)+sqrt(((boundx-x)**2)+((boundy-y)**2));
```

Apply RSABE if $SWR \geq 0.294$

RSABE if

a. $critbound \leq 0$ and

b. $0.8000 \leq pointest \leq 1.2500$

If $SWR < 0.294$, apply conventional (unscaled ABE), mixed effects model.

ABE if 90% CI within 0.8000 and 1.2500.

SAS code (FDA)

Fully replicated 4-way design

```
data test1;  
  set test;  
  if (seq=1 and per=1) or (seq=2 and per=2);  
  lat1t=lauct;  
run;
```

```
data test2;  
  set test;  
  if (seq=1 and per=3) or (seq=2 and per=4);  
  lat2t=lauct;  
run;
```

```
data ref1;  
  set ref;  
  if (seq=1 and per=2) or (seq=2 and per=1);  
  lat1r=lauct;  
run;
```

```
data ref2;  
  set ref;  
  if (seq=1 and per=4) or (seq=2 and per=3);  
  lat2r=lauct;  
run;
```

SAS code (FDA)

Fully replicated 4-way design (cont'd)

```
data scavbe;
  merge test1 test2 ref1 ref2;
  by seq subj;
  → ilat=0.5*(lat1t+lat2t-lat1r-lat2r);
  dlat=lat1r-lat2r;
run;

proc mixed data=scavbe;
  class seq;
  model ilat =seq/ddfm=satterth;
  estimate 'average' intercept 1 seq 0.5 0.5/e c1 alpha=0.1;
  ods output CovParms=iout1;
  ods output Estimates=iout2;
  ods output NObs=iout3;
  title1 'scaled average BE';
  title2 'intermediate analysis - ilat, mixed';
run;

pointest=exp(estimate);
x=estimate**2-stderr**2;
boundx=(max((abs(lower)), (abs(upper))))**2;
```

SAS code (FDA)

Fully replicated 4-way design (cont'd)

```
proc mixed data=scavbe;  
  class seq;  
  model dlat=seq/ddfm=satterth;  
  estimate 'average' intercept 1 seq 0.5 0.5/e c1 alpha=0.1;  
  ods output CovParms=dout1;  
  ods output Estimates=dout2;  
  ods output NObs=dout3;  
  title1 'scaled average BE';  
  title2 'intermediate analysis - dlat, mixed';  
run;  
  
s2wr=estimate/2;  
dfd=df;  
  
theta=((log(1.25))/0.25)**2;  
y=-theta*s2wr;  
boundy=y*dfd/cinv(0.95,dfd);  
swr=sqrt(s2wr);  
→ critbound=(x+y)+sqrt(((boundx-x)**2)+((boundy-y)**2));
```

SAS code (FDA)

Unscaled 90% BE confidence intervals (applicable if critbound>0)

PROC MIXED

data=pk;
CLASSES SEQ SUBJ PER TRT;
MODEL LAUCT = SEQ PER TRT/ DDFM=SATTERTH;
→ RANDOM TRT/TYPE=FA0(2) SUB=SUBJ G;
REPEATED/GRP=TRT SUB=SUBJ;
ESTIMATE 'T vs. R' TRT 1 -1/CL ALPHA=0.1;
ods output Estimates=unsc1;
title1 'unscaled BE 90% CI - guidance version';
title2 'AUCt';

run;

data unsc1;
set unsc1;
unscabe_lower=exp(lower);
unscabe_upper=exp(upper);
run;

→ Note: Lines marked with an arrow are missing in FDA's code!

Example datasets (EMA)

- Q&A document (March 2011)
 - Data set I
4-period 2-sequence (RTRT | TRTR) full replicate, imbalanced (77 subjects), incomplete (missing periods: two periods in two cases, one period in six cases).
 - Data set II
3-period 3-sequence (TRR | RTR | RRT) partial replicate, balanced (24 subjects), complete (all periods).
 - Download in Excel 2000 format:
[http://bebac.at/downloads/Validation Replicate Design EMA.xls](http://bebac.at/downloads/Validation%20Replicate%20Design%20EMA.xls)