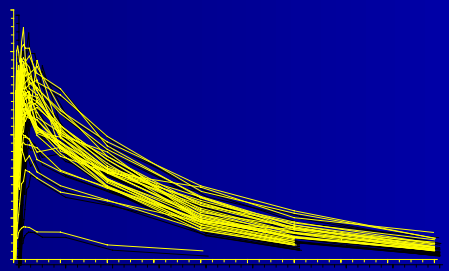
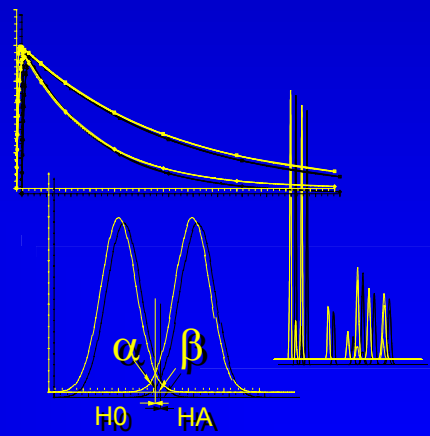
Statistical Evaluation of Bioequivalence Studies

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

Assumptions: General

World *'Reality'*





Model *'Data'*

Theory *'Truth'*

Assumptions: Pharmacokinetics

$$\frac{F_1 \cdot AUC_1}{D_1 \cdot CL_1}, \frac{F_2 \cdot AUC_2}{D_2 \cdot CL_2}$$

$$F_{rel}(BA) = \frac{AUC_1}{AUC_2}$$

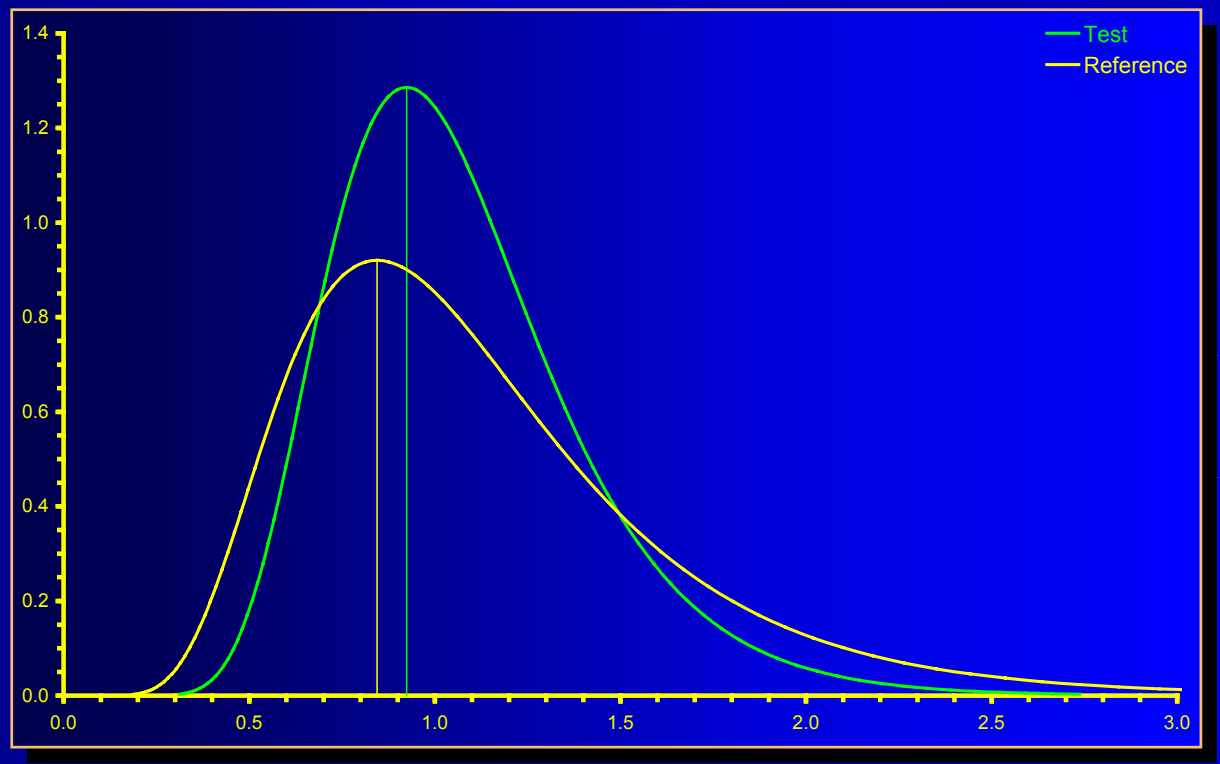
Assumption 1: $D_1 = D_2$ ($D_1/D_2 = 1^*$)

Assumption 2: $CL_1 = CL_2$

Assumptions: Statistics

Distribution

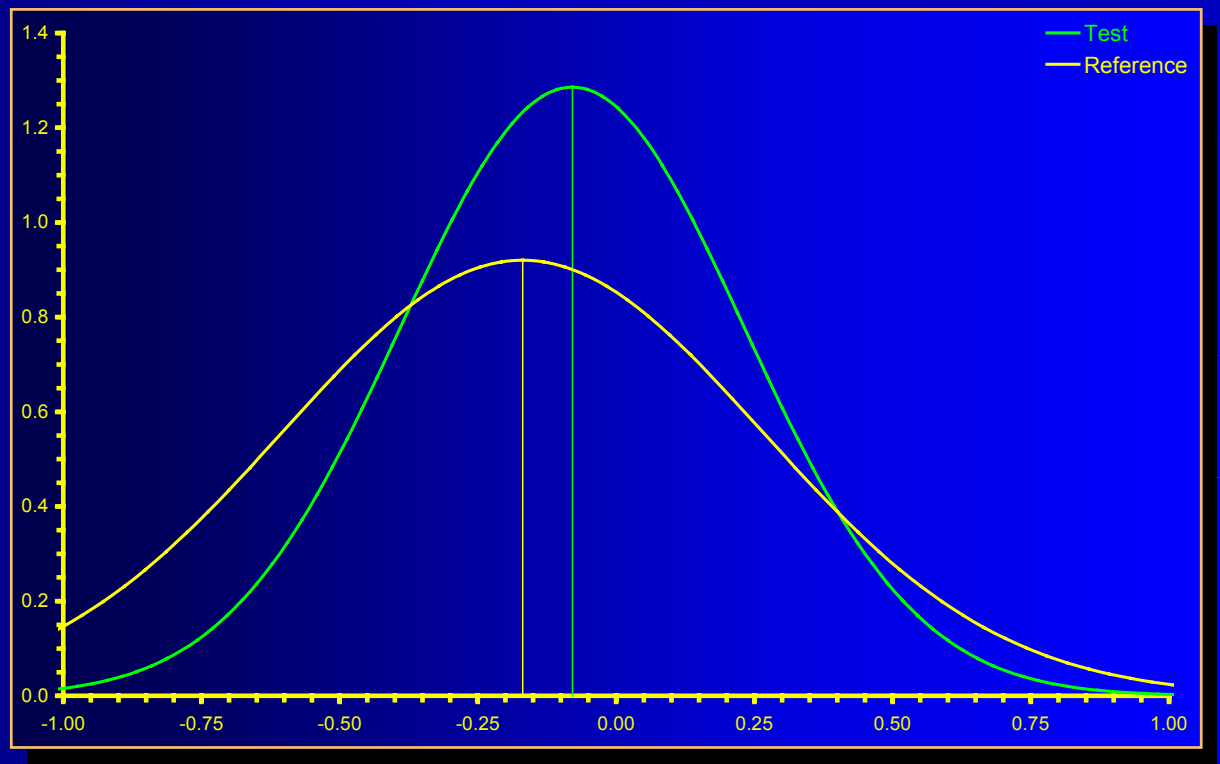
- IDD (Independent Identically Distribution)



Assumptions: Statistics

Multiplicative Model

- Log-Transformation (PK, Analytics)



Assumptions: Statistics

Multiplicative Model (without carryover)

$$X_{ijk} = \mu \cdot \pi_k \cdot \Phi_l \cdot s_{ik} \cdot e_{ijk}$$

X_{ijk}: *ln*-transformed response of j-th subject ($j=1,\ldots,n_i$) in i-th sequence ($i=1,2$) and k-th period ($k=1,2$), μ: global mean, μ_l: expected formulation means ($l=1,2$: $\mu_1=\mu_{test}$, $\mu_2=\mu_{ref.}$), π_k: fixed period effects, Φ_l: fixed formulation effects ($l=1,2$: $\Phi_1=\Phi_{test}$, $\Phi_2=\Phi_{ref.}$)

Assumptions: Statistics

Multiplicative Model (without carryover)

$$X_{ijk} = \mu \cdot \pi_k \cdot \Phi_l \cdot s_{ik} \cdot e_{ijk}$$

s_{ik}: random subject effect, e_{ijk}: random error

Main Assumptions:

- All $ln\{s_{ik}\}$ and $ln\{e_{ijk}\}$ are independently and normally distributed about unity with variances σ^2_s and σ^2_e.
- All observations made on different subjects are independent.

Global Harmonization?

Transformations (e.g. […], logarithm) should be specified in the protocol and a rationale provided […]. The general principles guiding the use of transformations to ensure that the assumptions underlying the statistical methods are met are to be found in standard texts […]. In the choice of statistical methods due attention should be paid to the statistical distribution […]. When making this choice (for example between parametric and non-parametric methods) it is important to bear in mind the need to provide statistical estimates of the size of treatment effects together with confidence intervals […].

Anonymous [International Conference on Harmonisation];
Topic E 9: Statistical Principles for Clinical Trials. (5 February 1998)

Global Harmonization?

No analysis is complete until the assumptions that have been made in the modeling have been checked. Among the assumptions are that the repeated measurements on each subject are independent, normally distributed random variables with equal variances. Perhaps the most important advantage of formally fitting a linear model is that diagnostic information on the validity of the assumed model can be obtained. These assumptions can be most easily checked by analyzing the residuals.

Jones, B. and M.G. Kenward; Design and Analysis of Cross-Over Trials.
2nd Edition, Chapman & Hall, Boca Raton, London, New York, Washington, D.C. (2003)

Nonparametrics

The limited sample size in a typical BE study precludes a reliable determination of the distribution of the data set. Sponsors and/or applicants are not encouraged to test for normality of error distribution after log-transformation […].

Anonymous [FDA, Center for Drug Evaluation and Research (CDER)];
Guidance for Industry: Statistical Approaches to Establishing Bioequivalence. (January 2001)

Acceptable in:
Turkey (MOH, November 2005)
Saudia Arabia (SFDA, May 2005)

Nonparametrics

5. In which cases may a non-parametric statistical model be used?

The NfG states under 3.6.1–Statistical analysis: "*AUC and C_{max} should be analysed using ANOVA after log transformation.*"

The reasons for this request are the following:

a) the AUC and C_{max} values as biological parameters are usually not normally distributed;

b) a multiplicative model may be plausible;

c) after log transformation the distribution may allow a parametric analysis.

Comments:

a) – true b) – true c) – maybe, but may also terribly fail

Anonymous [EMEA/CHMP/EWP/40326/2006];
Questions & Answers on the BA and BE Guideline (27 July 2006)

Nonparametrics

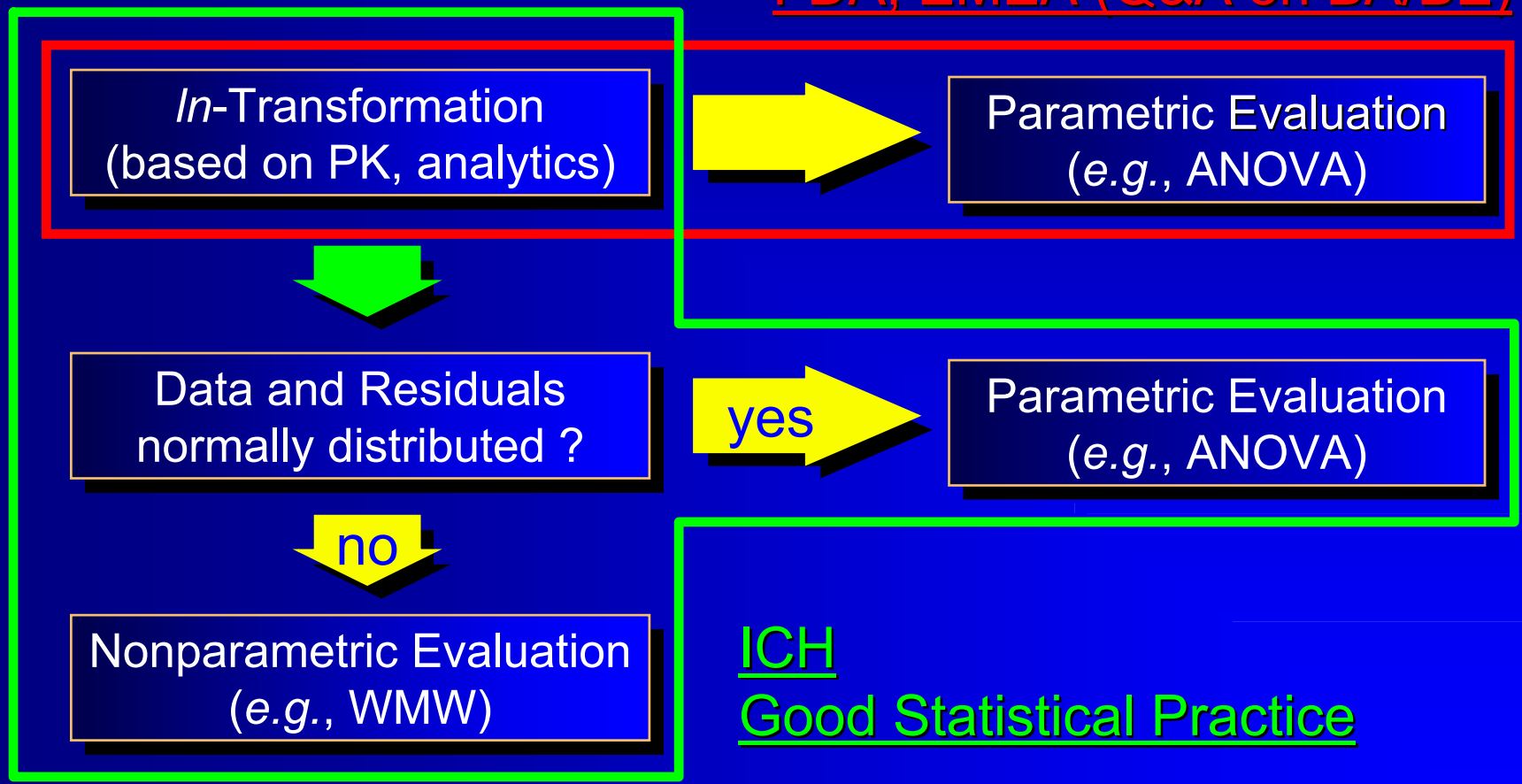
5. In which cases may a non-parametric statistical model be used?

However, the true distribution in a pharmacokinetic data set usually cannot be characterised due to the small sample size, so it is not recommended to have the analysis strategy depend on a pre-test for normality. Parametric testing using ANOVA on log-transformed data should be the rule. Results from non-parametric statistical methods or other statistical approaches are nevertheless welcome as sensitivity analyses. Such analyses can provide reassurance that conclusions from the experiment are robust against violations of the assumptions underlying the analysis strategy.

Comment: It is well known that the efficiency of *e.g.*, the Wilcoxon-Mann-Whitney test for normal distributed data is $3/\pi \approx 95.5$ %; for *not normal distributed data* the efficiency is > 100 %!

Global Harmonization?

FDA, EMEA (Q&A on BA/BE)



ICH
Good Statistical Practice

Global Harmonization?

- In almost all regulations two metrics are necessary to demonstrate BE, namely
  - extent (*e.g.*, AUC_t, AUC_∞, Ae), and
  - rate (*e.g.*, C_{max}, PTF) of exposure.
- One exception: US-FDA (where AUC_∞ *and* AUC_t must demonstrate extent of BE)
  - Although stated in the Guideline, such a requirement is statistically flawed.
    - Multiplicity issues (what is the patient's risk?)
    - Impossible α-adjustment (interdependence)

There can be only one!

Acceptance range for C_{max}

- Wider acceptance range for C_{max} (*e.g.*, 0.75–1.33), if
  - justified based on safety and efficacy grounds, and
  - specified in the study protocol
  - ✓ EU, WHO, Australia, NZ, Turkey, Saudia Arabia, Malaysia, Taiwan, ASEAN States, Argentina
  - ✓ RSA Standard for all drugs (no justification)
  - ✓ Japan, Switzerland (even for AUC)
  - – FDA, Brazil, India

Acceptance range for C_{max}

2. **Assessment of C_{max} in bioequivalence studies. In which cases is it allowed to use a wider acceptance range for the ratio of C_{max}?**

The NfG states under 3.6.2 that "*With respect to the ratio of C_{max} the 90% confidence interval for this measure of relative bioavailability should lie within an acceptance range of 0.80 – 1.25. In specific cases, such as a narrow therapeutic range, the acceptance interval may need to be tightened.*"

The NfG also states that "*In certain cases a wider interval may be acceptable. The interval must be prospectively defined, e.g. 0.75 – 1.33, and justified addressing in particular any safety or efficacy concerns for patients switched between formulations".*

Anonymous [EMEA/CHMP/EWP/40326/2006];
Questions & Answers on the BA and BE Guideline (27 July 2006)

Acceptance range for C_{max}

The possibility offered here by the guideline to widen the acceptance range of 0.80 – 1.25 for the ratio of C_{max} (not for AUC) should be considered exceptional and limited to a small widening (0.75 – 1.33).

Restricted to products for which at least one of the following criteria applies:

1) Data on PK/PD relationships (safety and efficacy) adequate to demonstrate that PD is not affected in a clinically significant way.
2) If PK/PD data are inconclusive or not available, clinical safety and efficacy data may be used, but specific for the compound and persuasive.
3) Reference product is a HVDP. See #8 of the Q&A document.

Comment: In a silent side-step widening of the acceptance range for AUC (NfG: „*AUC-ratio: […] In rare cases a wider acceptance range may be acceptable if it is based on sound clinical justification.*“) was entirely eliminated.

Outliers

- Problems
  - Parametric methods (ANOVA, GLM) are very sensitive to outliers
    - A single outlier may underpower a properly sized study.
    - Exclusion of outliers only possible if procedure stated in the protocol, and reason is justfied, *e.g.*,
      - Lacking compliance (subject did not take the medication),
      - Vomiting (up to 2 × t_{max} for IR, at all times for MR),
      - Analytical problems (*e.g.*, interferences in chromato-graphy);
      - Not acceptable if only based on statistical grounds.

Outliers

- Solution I
  - Since assumptions are violated, you may apply a statistical method which does not rely on those!
  - Drawback: Regulatory acceptance?

Outliers

- Solution II
  - Stay with the parametric method, but
    - evaluation of both the full (original) data set, and the reduced data set (outliers exluded), and
    - discuss influence on the outcome of the study.
  - In accordance with EMEA's Q&A #3:
    - Exceptional reasons may justify post-hoc data exclusion but this should be considered with utmost care. 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.

Re-testing of subjects

- If you suspect a product failure of the reference formulation, you may 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).
  - Expect questions anyway (although *sometimes* suggested by the FDA, not covered in any guideline; statistical evaluation not trivial…)

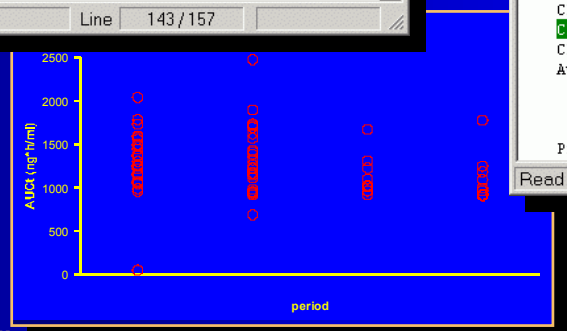
Re-testing of subjects

n=24: 83.3%–131.1% ⇒ +n=6: 86.7%–122.5%

Bioequivalence Text - [Untitled11]

```
Bioequivalence Statistics

User-Specified Confidence Level for CI's and Power = 90.0000
Percent of Reference to Detect for 2-1 Tests and Power = 20.0%
A.H.Lower =  0.800   A.H.Upper =  1.250

Formulation variable: formulation
Reference: R   LSMean=    7.085698  SE=    0.104041  GeoLSM= 1194.756651
--------------------------------------------------------------------------

Test:      T   LSMean=    7.129902  SE=    0.104041  GeoLSM= 1248.754167

    Difference =     0.0442,  Diff_SE=    0.1321,  df= 22.0
    Ratio(%Ref) =   104.5195

                        Classical               Westlake
    CI  80% = (   87.7732,  124.4610)  (   80.9169,  119.0831)
    CI  90% = (   83.2969,  131.1493)  (   74.5939,  125.4061)
    CI  95% = (   79.4604,  137.4815)  (   68.4657,  131.5343)
    Failed to show average bioequivalence for confidence=90.00 and percent=20.0.

                       Two One-Sided T-tests

    Prob(< 80%)=0.0277  Prob(> 125%)=0.0947  Max=0.0947  Total=0.1224
```

Read Only | Line 143 / 157



Bioequivalence Text - [Untitled10]

```
Bioequivalence Statistics

User-Specified Confidence Level for CI's and Power = 90.0000
Percent of Reference to Detect for 2-1 Tests and Power = 20.0%
A.H.Lower =  0.800   A.H.Upper =  1.250

Formulation variable: formulation
Reference: R   LSMean=    7.106774  SE=    0.109442  GeoLSM= 1220.204475
--------------------------------------------------------------------------

Test:      T   LSMean=    7.137202  SE=    0.052715  GeoLSM= 1257.903871

    Difference =     0.0304,  Diff_SE=    0.1017,  df= 28.7
    Ratio(%Ref) =   103.0896

                        Classical               Westlake
    CI  80% = (   90.2172,  117.7987)  (   85.7333,  114.2667)
    CI  90% = (   86.7263,  122.5403)  (   81.2107,  118.7893)
    CI  95% = (   83.7240,  126.9345)  (   76.9552,  123.0448)
    Average bioequivalence shown for confidence=90.00 and percent=20.0.

                       Two One-Sided T-tests

    Prob(< 80%)=0.0093  Prob(> 125%)=0.0341  Max=0.0341  Total=0.0434
```

Read Only | Line 211 / 225

NTI Drugs

- USA, Japan — No difference to other drugs
- WHO, EU, NZ, India — 90 % CI, acceptance range may be tightened
- Denmark — 90 % CI within 0.90–1.11 for *some* drugs http://www.dkma.dk/1024/visUKLSArtikel.asp?artikelID=6437 (17 Jan 2006)
- Brazil — 95 % CI within 0.80–1.25
- Canada — Common procedure; considering
  - AUC: 90 % CI within 0.90–1.12
  - C_{max}: 90 % CI within 0.80–1.25

  http://www.hc-sc.gc.ca/dhp-mps/alt_formats/hpfb-gpsa/pdf/prodpharma/crit_dose_e.pdf (5 Jul 2005)

Add-on Design

- Reasonable,
  - if uncertain sample size estimate,
  - for ethical reasons.

✓	Canada	If BE not shown, additional subjects are included; *F*-test (equality of variances), pooled analysis. No α-adjustment.
✓	Japan	2nd part with sample size $\geq$ 1st part / 2
✓	RSA	max. sample size must be stated *a-priori*
✓	NZ	Group sequential design (with α-adjustment)
–	USA	No way
±	EU	Evaluation of first part by an independent statistician (CV only!). Not covered in NfG.

Group Sequential Design

- Not mentioned in any Guideline, but
  - are standard in clincial research.
  - Although discussed at BioInternationals ’89 to ’96, no concensus was reached.

  ± EU
    - Personal Experience:
      - A proposed method *) was not accepted in the planning phase (3 cases Germany).

        *) L.A. Gould;
        Group Sequential Extension of a Standard Bioequivalence Testing Procedure.
        J. Pharmacokin. Biopharm. 32(1), 57-86 (1995)

Group Sequential Design

± EU

- Personal Experience:
  - Evaluation of first part by an independent statistician (CV only!), performance of a second part, evaluation of pooled data without α-adjustment – 90 % CI (2 cases Germany, 1 case France).
  - May be a reasonable approach, because Add on Designs are in practice in Canada (since 1991), and Japan (since at least 1997).

HVDs/HVDPs

- Highly Variable Drugs / Drug Products (intra-subject variability >30 %)
  - ✓ USA Replicate Design recommended.
  - ± EU […] under certain circumstances […] alternative well-established designs could be considered such as […] replicate designs for substances with highly variable disposition.
  - ± NZ […] studies in which treatments are replicated within each subject, may improve discriminatory power for highly variable medicines.
  - ? Reference Scaled Average Bioequivalence (only stated in South African Guidelines).

Studies of >2 formulations

- Advantages
  - Allows to choose between two ore more candidate test formulations.
  - Comparison of a test formulation with several references.
- Standard design for establishment of dose proportionality.

Studies of >2 formulations

- Disadvantages
  - Not mentioned in any guideline – except Brazil's ANVISA.
  - Statistical analysis more complicated – especially in the case of drop outs.
  - *May* need measures against multiplicity, increasing the sample size.

Studies of >2 formulations

- Bonferroni-correction needed if more than 1 formulation will be marketed (for 3 simultaneous comparisons without correction patient's risk increases from 5% to 14%).

k	$P_{\alpha=0.05}$	$P_{\alpha=0.10}$	$\alpha_{adj.}$	$P_{\alpha adj.}$	$\alpha_{adj.}$	$P_{\alpha adj.}$
1	5.00%	10.00%	0.0500	5.00%	0.100	10.00%
2	9.75%	19.00%	0.0250	4.94%	0.050	9.75%
3	14.26%	27.10%	0.0167	4.92%	0.033	6.67%
4	18.55%	34.39%	0.0125	4.91%	0.025	9.63%
5	22.62%	40.95%	0.0100	4.90%	0.020	9.61%
6	26.49%	46.86%	0.0083	4.90%	0.017	9.59%

Studies of >2 formulations

- Often a wrong design is applied, namely
  - a repeated latin square, instead of
  - a Williams' design.
- Example for 3 treatments (T_1, T_2, R)

3 sequence latin square

Seq.	P_1	P_2	P_3
1	T_1	T_2	R
2	T_2	R	T_1
3	R	T_1	T_2

6 sequence Williams' design

Seq.	P_1	P_2	P_3
1	T_1	T_2	R
2	T_2	R	T_1
3	R	T_1	T_2
4	T_1	R	T_2
5	T_2	T_1	R
6	R	T_2	T_1

Parallel Groups

- Sometimes it is infeasible or even impossible to demonstrate BE from a 'conventionally' designed cross-over study; a study in parallel groups should be employed:
  - Drugs with long half lives.
  - Potentially toxic drugs.
  - Studies in patients, where the condition of the disease irreversibly changes.

Parallel Groups

- Design Issues
  - EMEA NfG on BA/BE
    - 3.2.4 Genetic phenotyping
      Phenotyping and/or genotyping of subjects should be considered for […] all studies using parallel group design.
      If a drug is known to be subject to major genetic polymorphism, studies could be performed in panels of subjects of known phenotype or genotype for the polymorphism in question.
  - Since the comparison is based on inter-subject effects,
    - stratify groups for phenotype/genotype.
    - run two studies of the respective phenotype/genotype (?)
    - one study of the major phenotype/genotype (?)

Parallel Groups

- Evaluation
  - FDA/CDER, Statistical Approaches to Establishing Bioequivalence (January 2001)
    - Section VI. B.1.d. Parallel Designs
      For parallel designs, the confidence interval for the difference of means in the log scale can be computed using the total between-subject variance. As in the analysis for replicated designs (section VI. B.1.b), equal variances should not be assumed.
  - The conventional *t*-test depends on the assumption that samples come from populations that have identical variances.
    - 'Naive pooling' of variances is relatively robust against unequal variances, but rather sensitive to inbalanced data.
    - If assumptions are violated, the conventional *t*-test becomes liberal (*i.e.*, the CI is too tight; **patient's risk > 5%**).

Sample data set

- Will be used throughout the lecture
- 2×2 Cross-over Study
  - 24 subjects (balanced: TR=RT=12)
  - Single dose
  - Target parameter: AUC_{0-t}
  - CV_{intra} 20.0 %
  - CV_{inter} 32.6 %
  - http://bebac.at/downloads/24sub.txt (CSV-format)

Trt	Rand	Sub	P1	P2
1	RT	1	44.1	39.1
1	RT	2	33.6	23.8
1	RT	3	45.5	40.8
2	TR	4	19.5	21.1
2	TR	5	67.2	51.5
2	TR	6	25.7	30.1
1	RT	7	35.3	26.7
1	RT	8	26.0	36.5
1	RT	9	38.2	57.8
2	TR	10	33.6	32.5
2	TR	11	25.1	36.8
2	TR	12	44.1	42.9
1	RT	13	25.6	20.1
1	RT	14	58.0	45.3
1	RT	15	47.2	51.8
2	TR	16	16.5	21.4
2	TR	17	47.3	39.4
2	TR	18	22.6	17.3
1	RT	19	17.5	30.1
1	RT	20	51.7	36.0
1	RT	21	24.5	18.2
2	TR	22	36.3	27.2
2	TR	23	29.4	39.6
2	TR	24	18.3	20.7

Parallel Groups: Example

- Evaluation (sample data set, period 1 only)
  - Original data set
    - Balanced (T 12, R 12)
    - Equal variances (s^2_R 0.1292, s^2_T 0.1796)
      F-ratio test *p* 0.5947
      Levene test *p* 0.5867
  - Modified data set:
    - Values of subjects 4 – 6 × 3
    - Subjects 22 – 24 removed
    - Inbalanced (T 9, R 12)
    - Unequal variances (s^2_R 0.1292, s^2_T 0.5639)
      F-ratio test *p* 0.0272
      Levene test *p* 0.1070

Parallel Groups: Example

- Evaluation (original data set)
  - Is your software able to give the correct answer?

Program / Method	equal variances	unequal variances
'manual' (Excel 2000)	63.51% – 110.19%	63.48% – 110.25%
R 2.5.0 (2007)	63.51% – 110.19%	63.49% – 110.22%
NCSS 2001 (2001)	63.51% – 110.19%	63.49% – 110.22%
STATISTICA 5.1H (1997)	63.51% – 110.19%	63.49% – 110.22%
WinNonlin 5.2 (2007)	63.51% – 110.20%	not implemented!
Kinetica 4.4.1 (2007)	63.51% – 110.19%	not implemented!
EquivTest/PK (2006)	63.51% – 110.18%	not implemented!

Parallel Groups: Example

- Evaluation (modified data set)

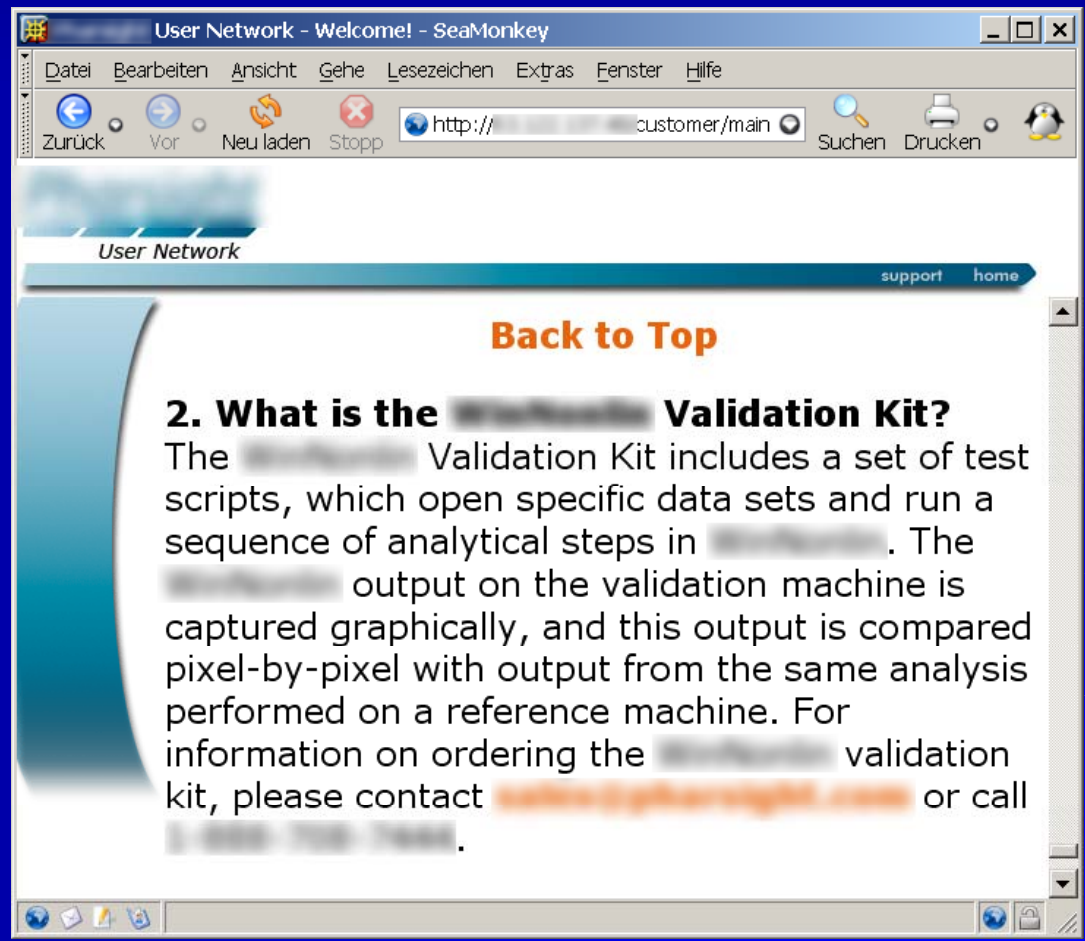
Program	equal variances	unequal variances
R 2.5.0 (2007)	81.21% – 190.41%	76.36% – 202.51%
NCSS 2001 (2001)	81.21% – 190.41%	76.36% – 202.51%

  - Inflated α-risk in 'conventional' *t*-test (naive pooling) is reflected in a tighter confidence interval.
  - Preliminary testing for equality in variances is flawed*) and should be avoided (FDA).
  - Approximations (*e.g.*, Satterthwaite, Aspin-Welch, Howe, Milliken-Johnson) are currently *not implemented* in packages 'specialized' in BE (WinNonlin, Kinetica, EquivTest/PK)!

*) Moser, B.K. and Stevens, G.R.; Homogeneity of variance in the two-sample means test. Amer. Statist. 46, 19-21 (1992)

Side note

Validated?
Sure!
But what if
2 × 2 = 5 ?



Sample Size

- Minimum Number of Subjects
  - **12** – WHO, EU, CAN, NZ, AUS, Malaysia, Argentina, ASEAN States, South Africa (20 for MR).
  - **12** (?) – USA: The total number of subjects in the study should provide adequate power for BE demon-stration […]. For modified-release products, a pilot study can help determine the sampling schedule to assess lag time and dose dumping. A pilot study that documents BE may be appropriate, provided its design and exe-cution are suitable and a sufficient number of subjects (*e.g.*, 12) have completed the study.
  - **24** – Saudia Arabia (12 – 24 if statistically justifiable).
  - **24** – Brazil.

Sample Size

- Maximum Number of Subjects
  - New Zealand:
    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 in BE-Guidelines (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.

Sample Size: Planning

- NfG on the Investigation of BA/BE
  - 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.

Sample Size: Planning

- NfG on the Investigation of BA/BE
  - 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.*, *prazoles?).

Sample Size: Planning

- NfG on the Investigation of BA/BE
  - 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.

Sample Size: Planning

- NfG on the Investigation of BA/BE
  - 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***in***equivalent 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 restrictive legislations (*e.g.*, Brazil's ANVISA), α must be tightened to 2.5 % for NTIDs (95 % confidence interval).

Sample Size: Planning

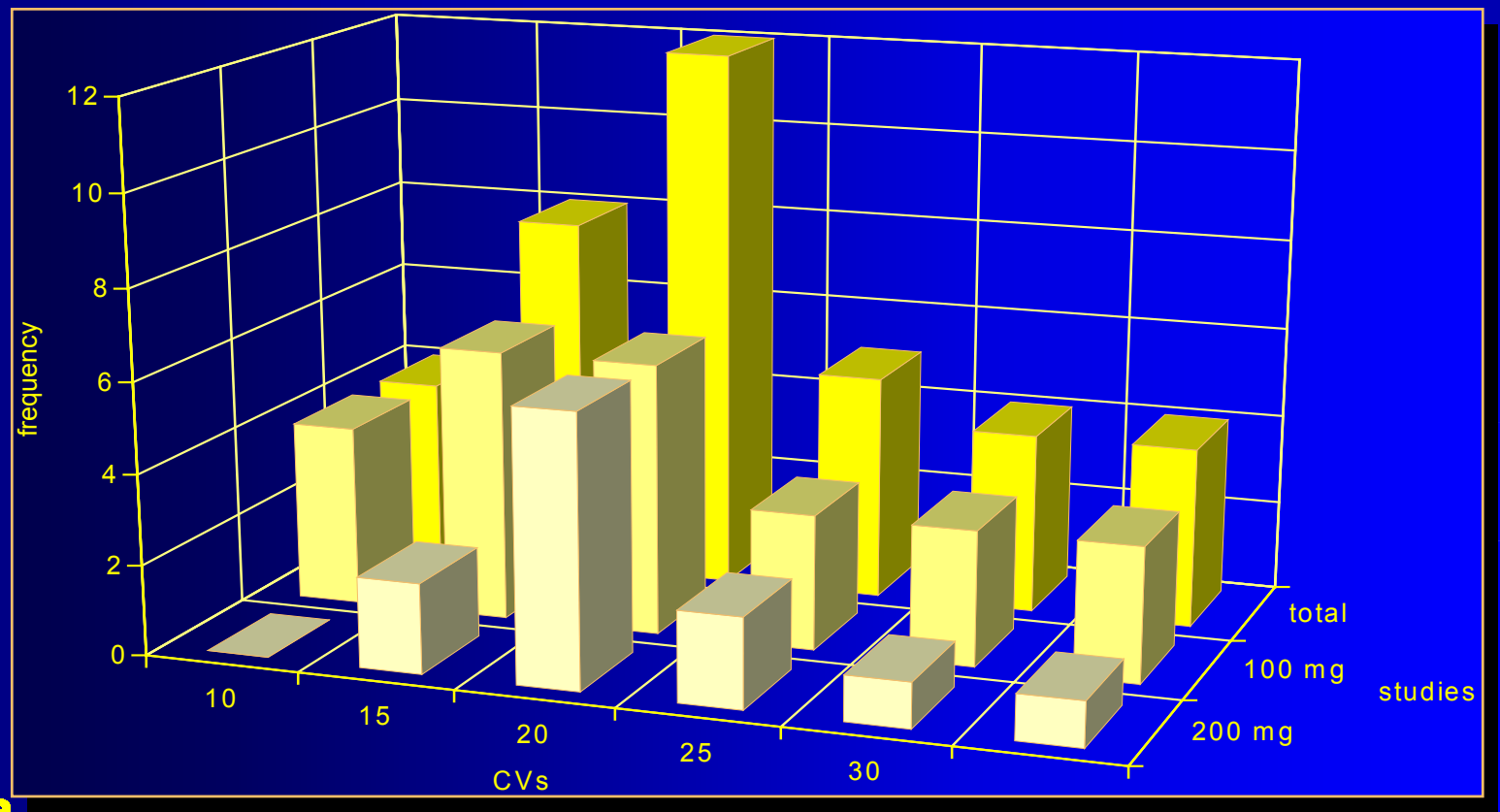
- NfG on the Investigation of BA/BE
  - Problems/solutions
    - … the *expected deviation (Δ) from the reference* …
      - Reliable estimate only from a previous full-sized study.
      - If you are using data from a pilot study, allow for a safety margin.
      - If no data are available, commonly a GMR (geometric test/reference-ratio) of 0.95 (Δ = 5 %) is used.
      - If more than Δ = 10 % is expected, questions from the ethics committee are likely.

Sample Size: Planning

- NfG on the Investigation of BA/BE
  - 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 – β).
        Remember: 1 out of 5 studies will fail just by chance!
      - If you plan for power of less than 70 %, problems with the ethics committee are likely.
      - If you plan for power of more than 90 % (especially with low variability drugs), problems with the regulator are possible ('forced bioequivalence').
      - Add subjects according to the expected drop-out rate!

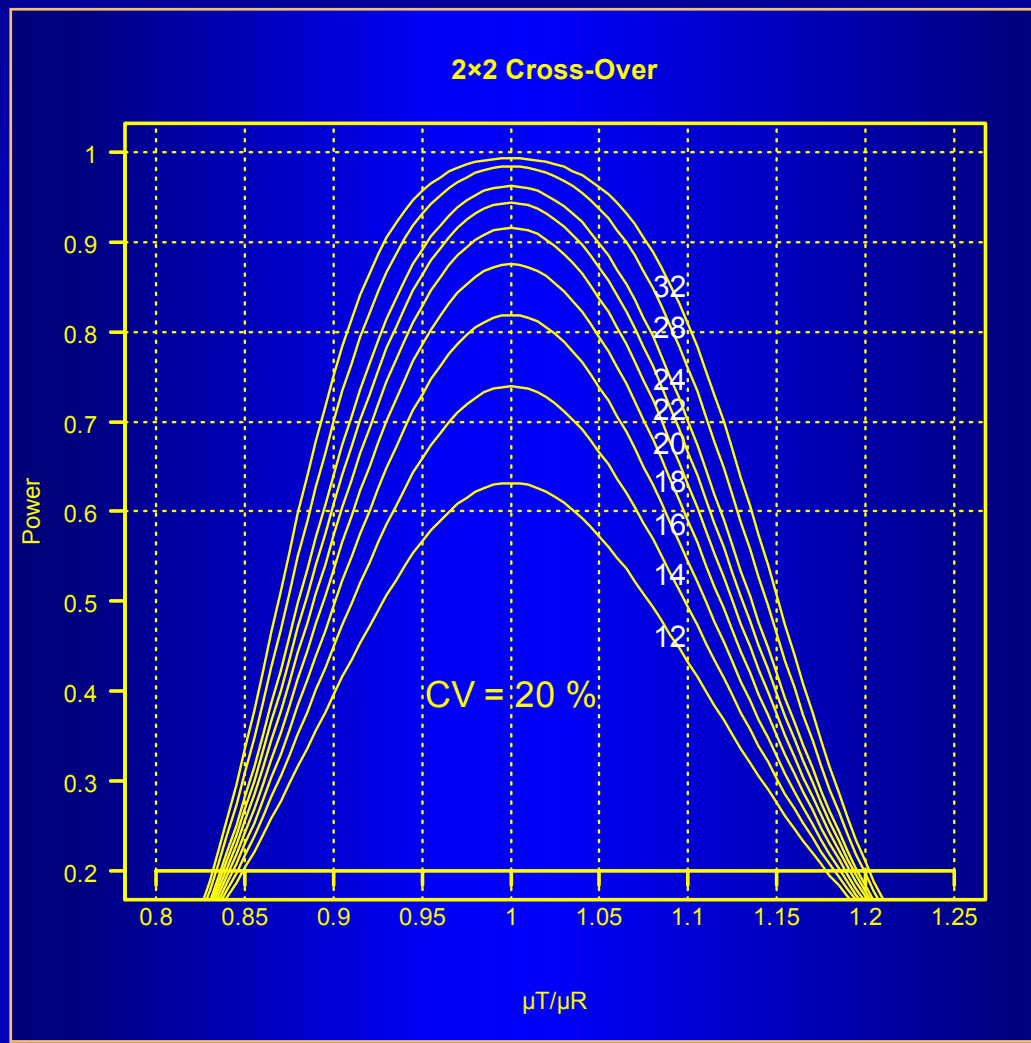
Sample Size: Planning

Doxicycline (37 studies ref. by Blume/Mutschler 1996)



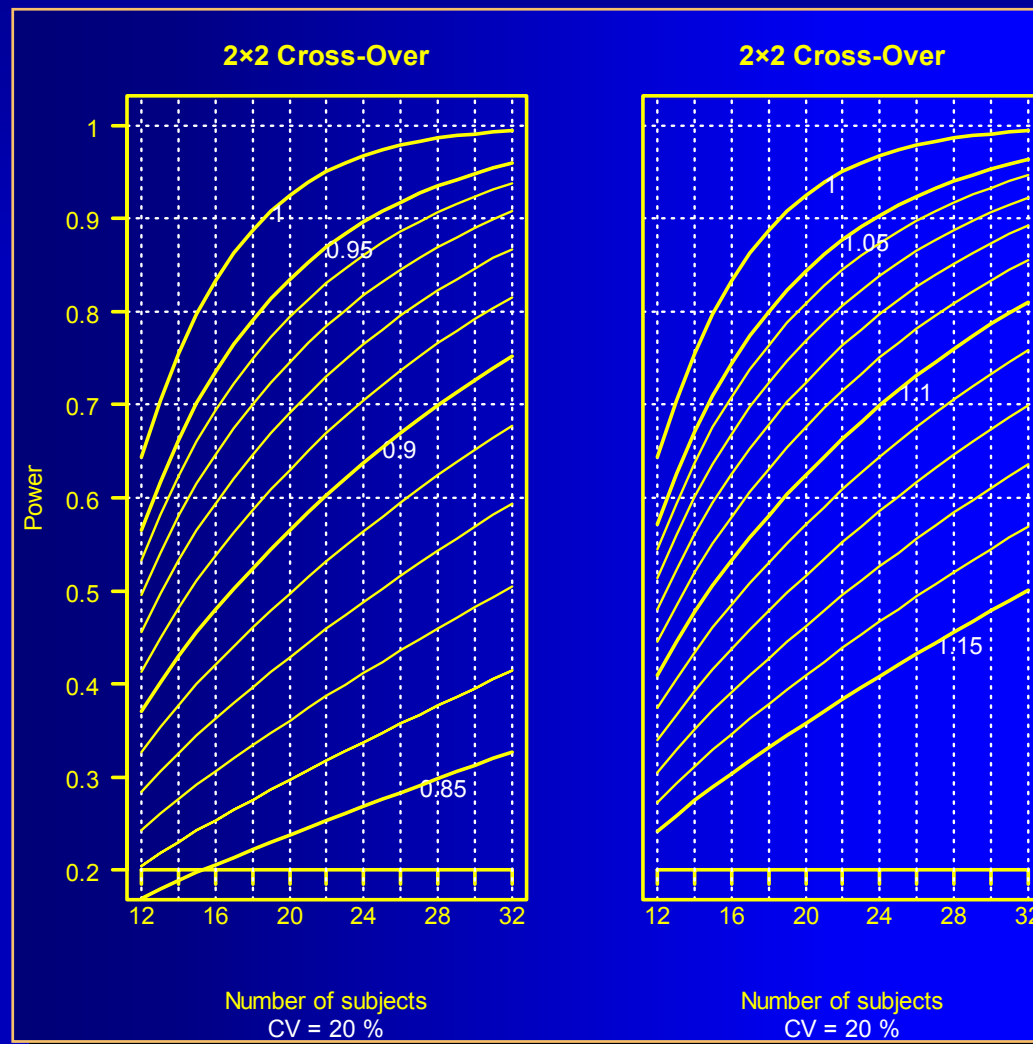
Sample Size: Power

Power to show BE with 12 – 32 subjects for CV_{intra} = 20%



Sample Size: Power

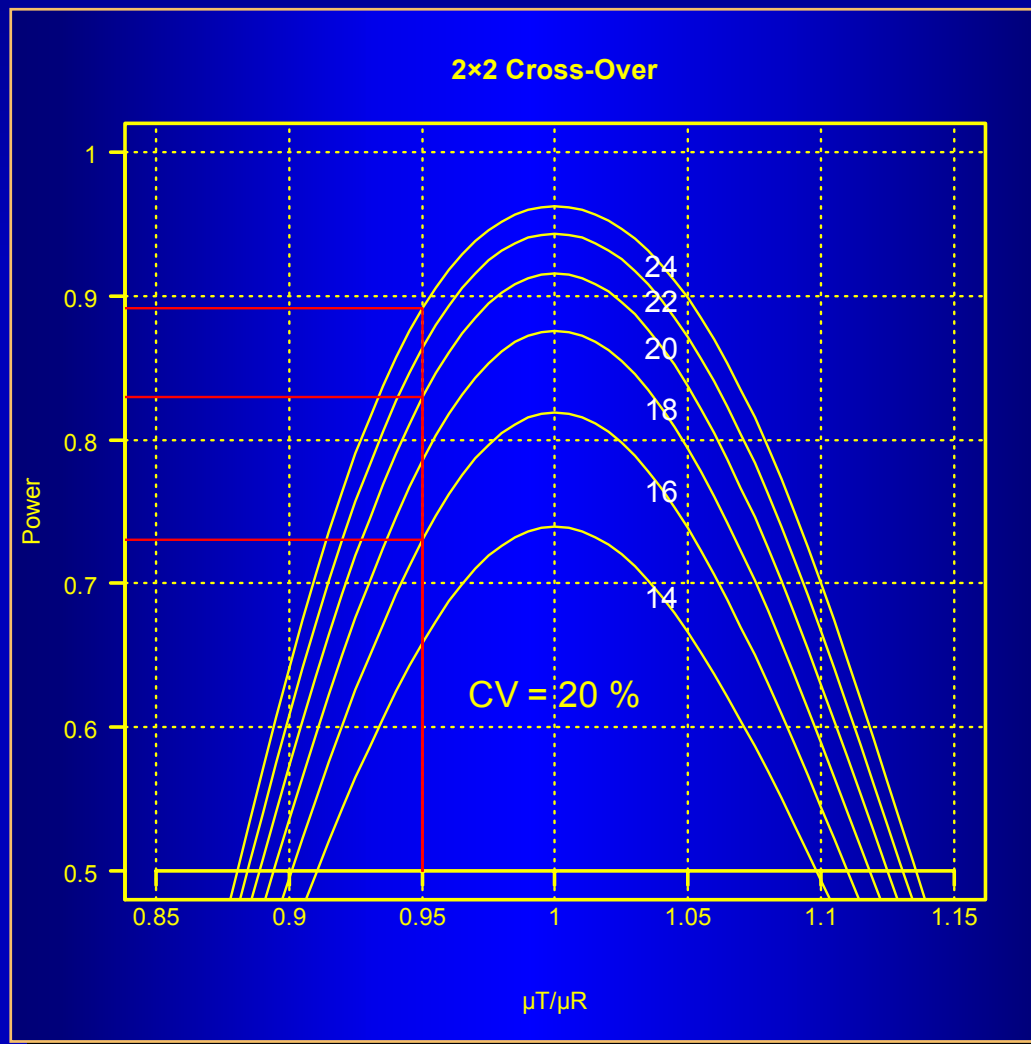
Power to show BE with 12 – 32 subjects for CV_{intra} = 20%



Sample Size: Power

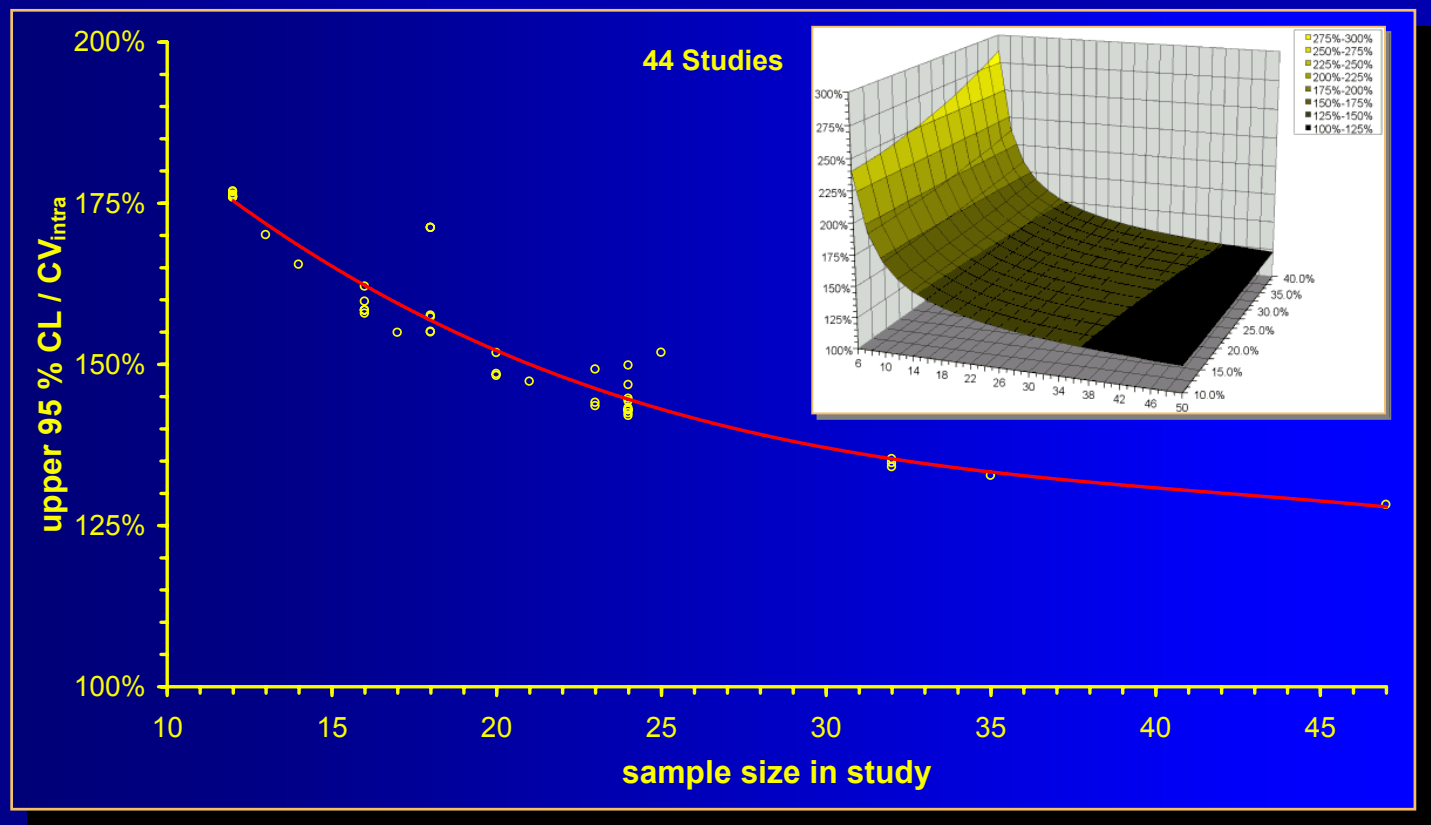
Effect of drop-outs on power to show BE (CV_{intra} 20%, GMR 0.95):

- n=24: 0.891
- n=20: 0.829 (-7 %)
- n=16: 0.730 (-12 %)



Sample Size: Planning

Estimated CV and upper 95 % CL



Sample Size: Sensitivity Analysis

- ICH E9
  - 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.

Sample Size: Sensitivity Analysis

- Sample data set
- n_{eq}: sample size to demonstrate BE for an expected deviation of -5% and 80% power.
  - Main study n=24: 96.4% (90% CI: 87.5%-106.5%)
    - CV_{intra} 20.00% ⇒ n_{eq} 18 CL_{upper} of CV 26.91% ⇒ n_{eq} 32
  - 4 subsets (I-IV) of sample size 6 ('pilot studies')
    - **I** 91.1% (77.7%-107.3%)
      CV_{intra} 13.15% ⇒ n_{eq} 10 CL_{upper} of CV 31.82% ⇒ n_{eq} 44
    - **II** 101.7% (77.8%-135.2%)
      CV_{intra} 22.74% ⇒ n_{eq} 24 CL_{upper} of CV 57.28% ⇒ n_{eq} 140
    - **III** 96.1% (78.2%-119.4%)
      CV_{intra} 17.32% ⇒ n_{eq} 14 CL_{upper} of CV 42.53% ⇒ n_{eq} 78
    - **IV** 94.6% (66.8%-137.7%)
      CV_{intra} 30.02% ⇒ n_{eq} 40 CL_{upper} of CV 79.07% ⇒ n_{eq} 264

Sample Size: Sensitivity Analysis

- Sample data set
  - 2 subsets (V-VI) of sample size 12 ('pilot studies')
    - **V** 96.5% (83.9%-111.6%)
      CV_{intra} 19.47% $\Rightarrow$ n_{eq} 18 CL_{upper} of CV 31.47% $\Rightarrow$ n_{eq} 44
    - **VI** 95.6% (83.9%-111.6%)
      CV_{intra} 22.14% $\Rightarrow$ n_{eq} 22 CL_{upper} of CV 35.93% $\Rightarrow$ n_{eq} 56

Sample Size: Sensitivity Analysis

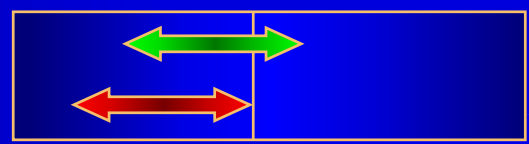
- Observations
  - Subset III: Point estimate (PE) 96.1%, CV 17.32%
    - Calculating the sample size for -5% and performing the main study in 14 subjects would have a fairly high probability of failure.
    - Ignoring the uncertainty in PE (and to a much greater extent) in CV is not a good idea.
  - Subset IV: PE 94.6%, CV 30.02%
    - Planing for 40 subjects, the study will very likely be over-powered.
    - Being cautious (upper CL of 79.07% $\Rightarrow$ n_{eq} 264!) would even lead to a wrong decision, that we have to deal with a highly variable drug, and subsequently unnecessary complicated design issues (*e.g.*, a replicate design with ScABE).

Sample Size: Sensitivity Analysis

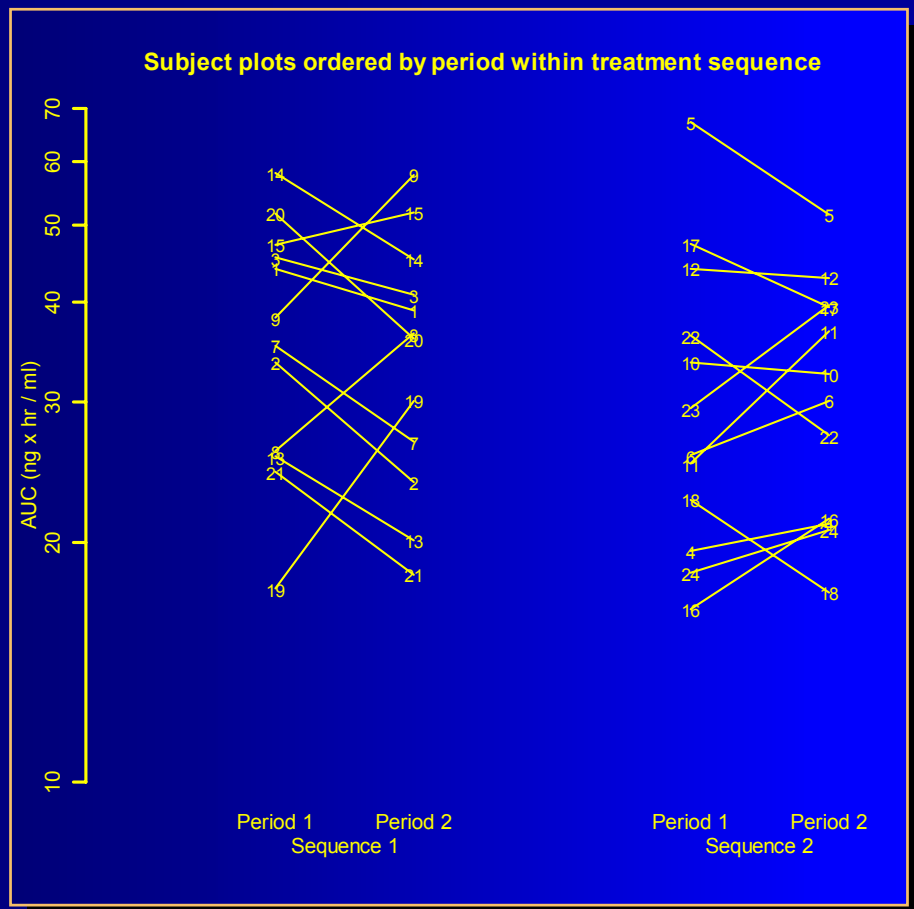
- Observations
  - Subsets of size 12 lead to more consistent results.
    - If you have stated such a procedure in your protocol, even BE may be claimed in both subsets, and no further study will be necessary.
    - If you want to use the upper CL in sample size estimation, you also get more consistant values.
    - If you have some previous hints of high intra-subject variability (>30%), a pilot study size of at least 16 subjects is reasonable.
- Conclusions
  - 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.

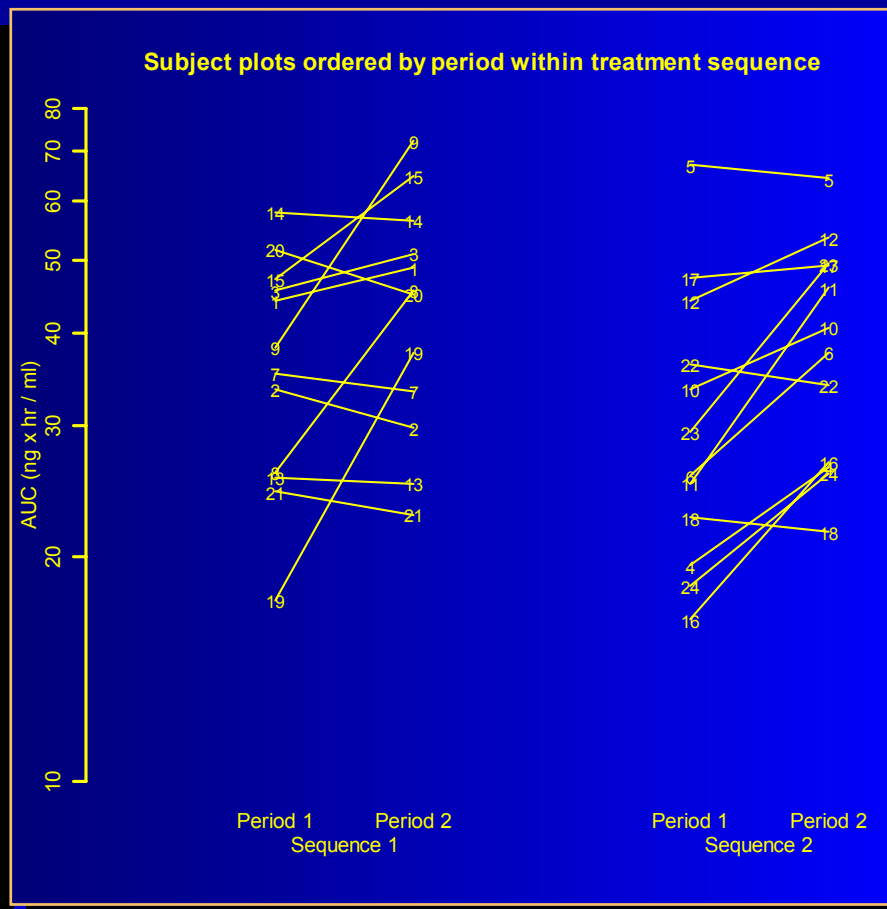
Low Variability

- Drugs / Drug Products with CV_{intra} <10%
  - No specific regulations 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).
  - Denmark
    - DKMA considers that the 90% CI for the ratio test versus reference should include 100% […].
    - Deviations may be accepted if they can be adequately justified not to have impact on either the overall therapeutic effect or safety profile of the product.

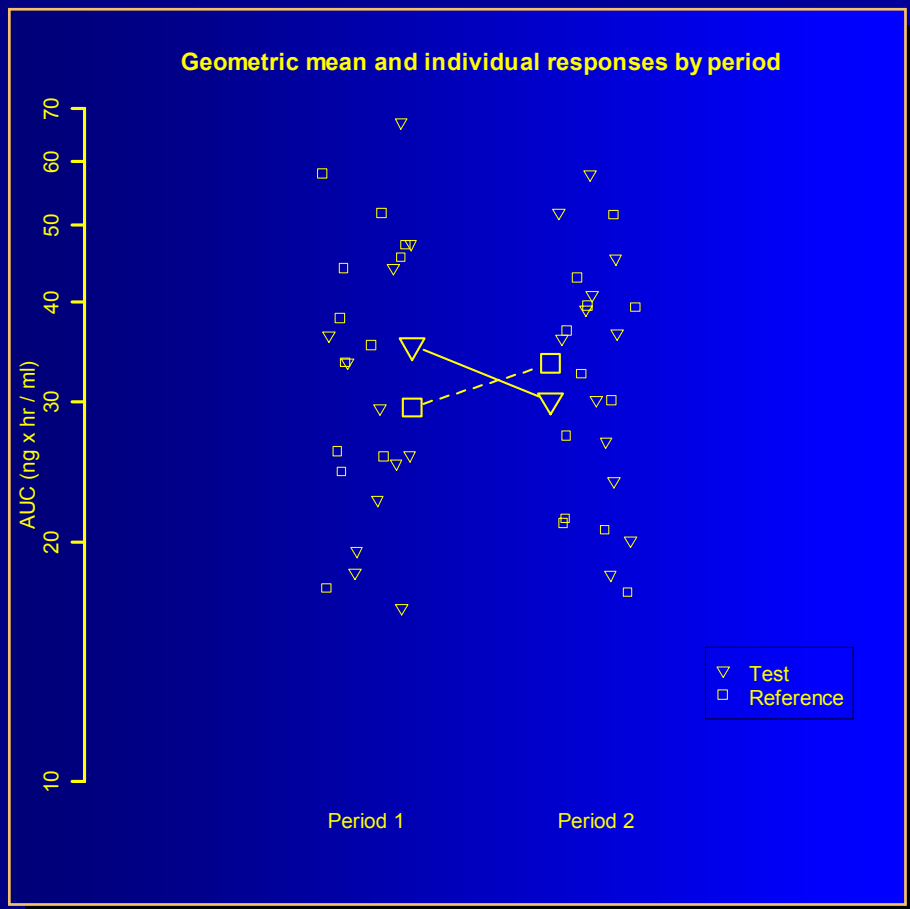


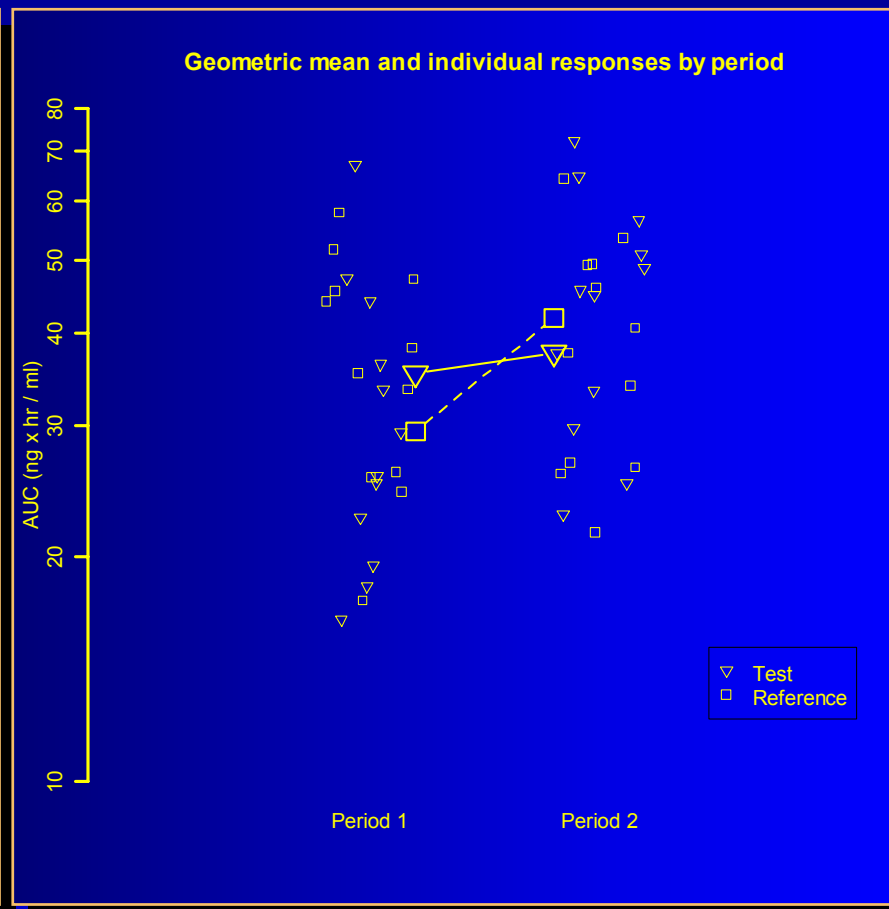
Nuisance: period effect





Nuisance: period effect

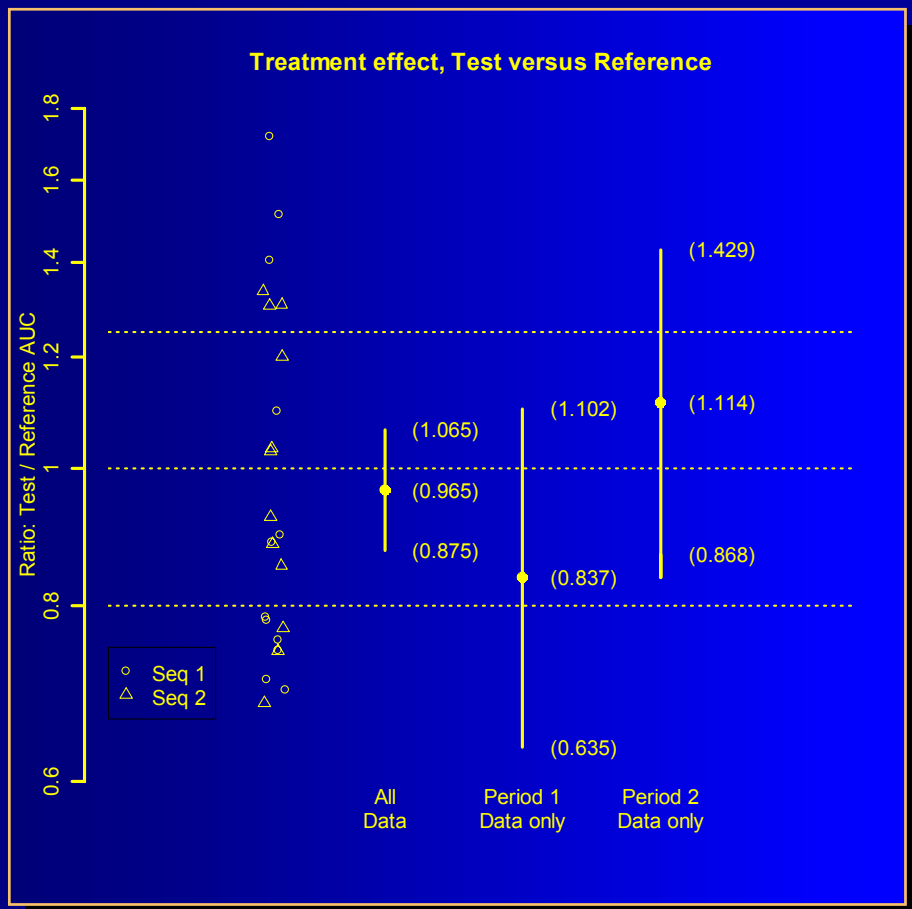


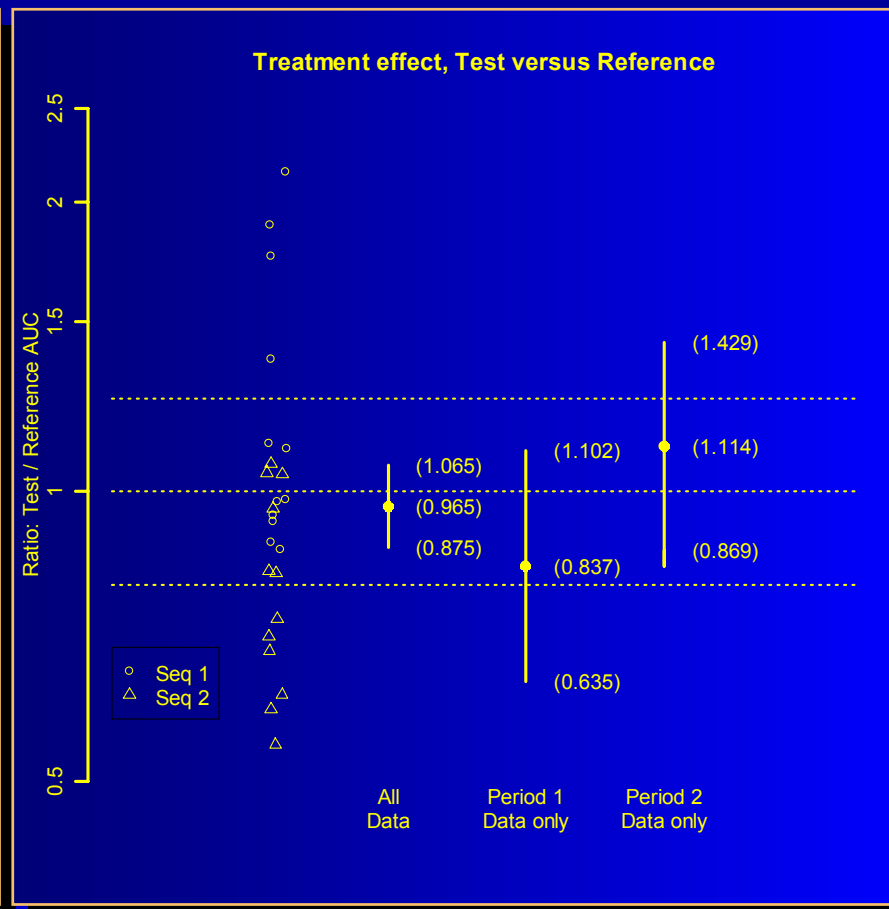


Nuisance: period effect

- Original data
  - AUC(p_2/p_1): 98.4%
  - Period: *p* 0.7856 (95% CI: 87.4% –110.8%)
  - Sequence: *p* 0.3239 (95% CI: 86.0% –154.8%)
  - GMR: 96.5% (90% CI: 87.5% –106.5%)
- Modified data (p_2 +25% of original values)
  - AUC(p_2/p_1): 123.0%
  - Period: *p* 0.0015 (95% CI: 109.3% –138.5%)
  - Sequence: *p* 0.3239 (95% CI: 86.0% –154.8%)
  - GMR: 96.5% (90% CI: 87.5% –106.5%)

Nuisance: period effect





Nuisance: sequence effect

- In a ‘standard’ 2×2 cross-over design
  - the sequence effect is confounded with
    - the carryover effect, and
    - the formulation-by-period interaction.
  - Therefore, a statistically significant sequence effect could indicate that there is
    - a true sequence effect,
    - a true carryover effect,
    - a true formulation by period interaction, or
    - a failure of randomization.

Nuisance: sequence effect

- 'Two-stage analysis'[1)] was – and still is – often applied.
  - Test for a significant sequence effect at α 0.10
  - If a significant sequence effect is found, evaluation of the first period as a parallel design
- This procedure was shown to be statistically flawed.[2)]

1) J.E. Grizzle;
The two-period change over design and ist use in clinical trials.
Biometrics 21, 467-480 (1965)

2) P. Freeman;
The performance of the two-stage analysis of two-treatment, two-period cross-over trials.
Statistics in Medicine 8, 1421-1432 (1989)

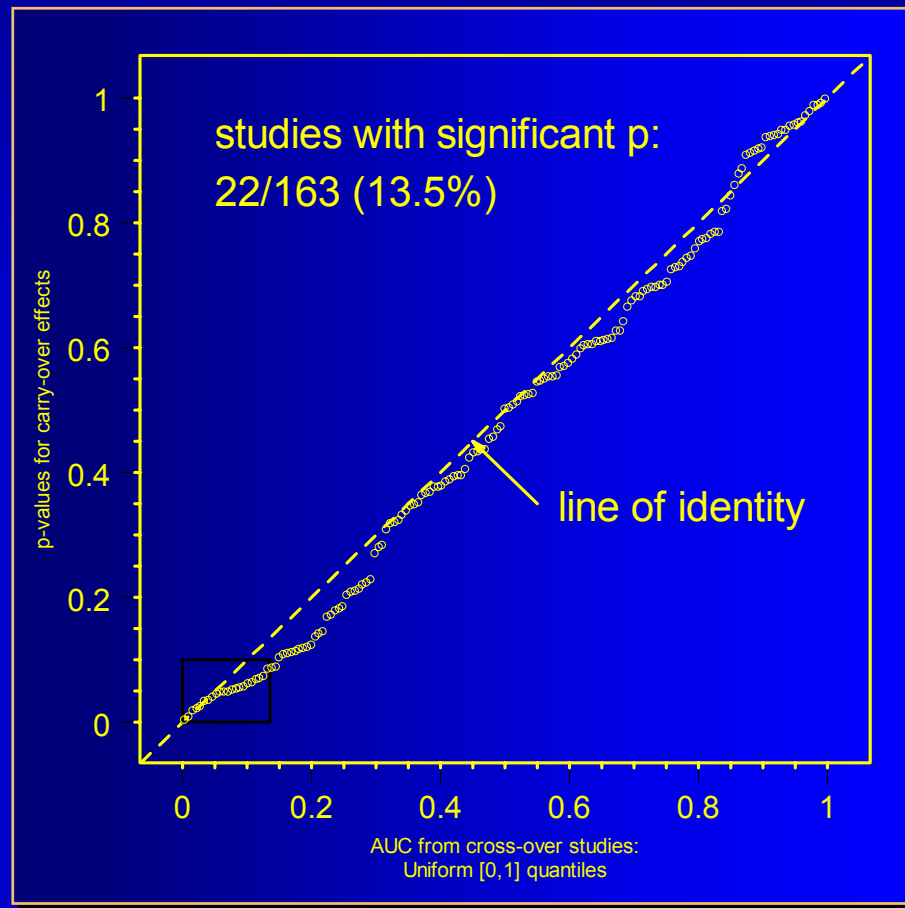
Nuisance: sequence effect

- In a large metastudy significant sequence effects were found at $\approx \alpha$, both for AUC and C_{max}.*)
  - 2×2 studies (n=324)
    - AUC: 34/324 (10.5%) C_{max}: 37/324 (11.4%)
  - 6×3 studies (n=96)
    - AUC: 4/96 (4.2%) C_{max}: 4/96 (4.2%)
  - For both metrics the distribution of *p* values followed closely Uniform [0,1]

*) D'Angelo, G., Potvin, D., and J. Turgeon; Carry-over effects in bioequivalence studies. J. Biopharm. Stat. 11, 35-43 (2001)

Nuisance: sequence effect

- These results could be confirmed (20 published studies, 143 studies from BEBAC's database; AUC):
  - Significant sequence effects in 22/163 studies (13.5%)
- Significant sequence effects in properly planned studies should be considered a statistical artefact (significant results are obtained in α of studies)

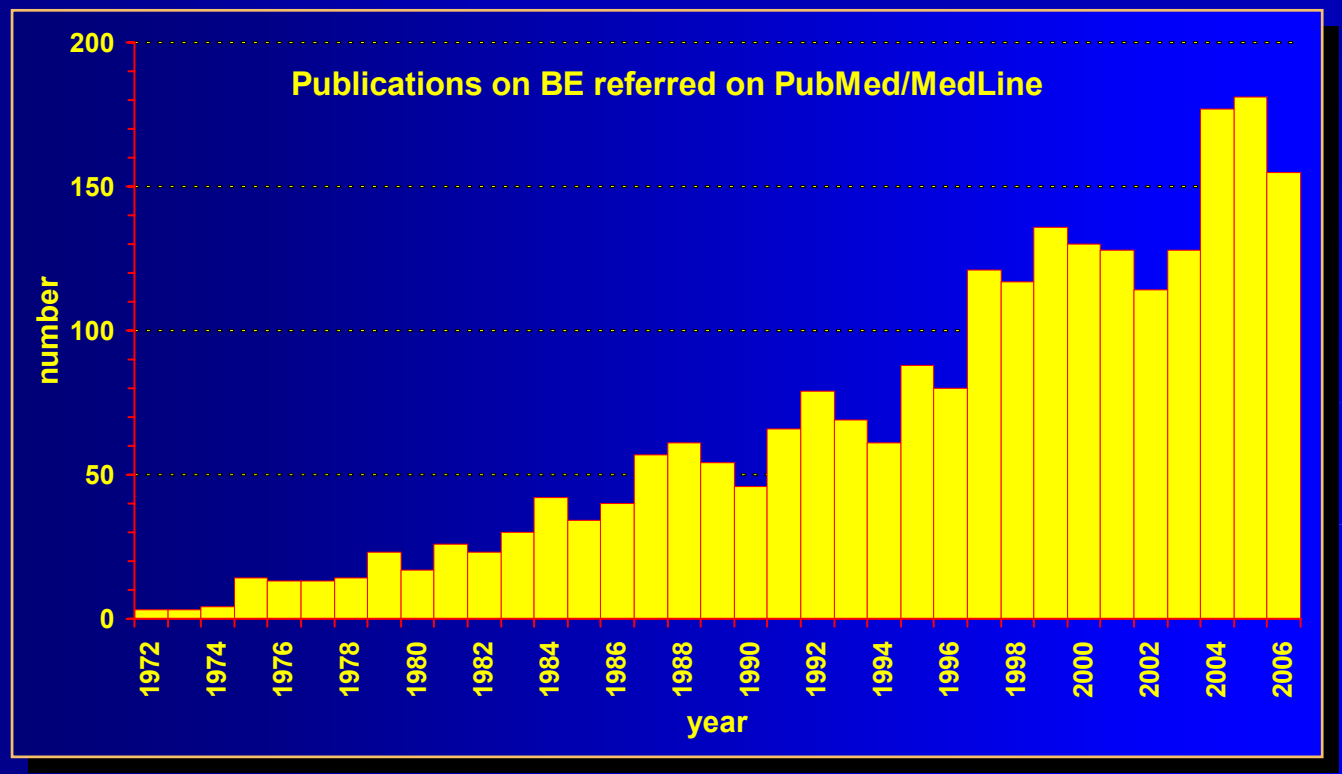


Nuisance: sequence effect

- Conclusions
  - No valid procedure exists to correct for a true sequence/carry-over effect
  - A true sequence/carry-over is highly unlikely in a BE study if
    - the study is performed in healthy subjects,
    - the drug is not an endogenous entity, and
    - an adequate washout period (no predose concentrations) was maintained.
  - Testing for a sequence effect is futile…

Are we making progress?

PubMed/MedLine: (bioequivalence) OR (comparative AND bioavailability), Field: Title/Abstract, Limits: Humans, Publication Date



Are we making progress?

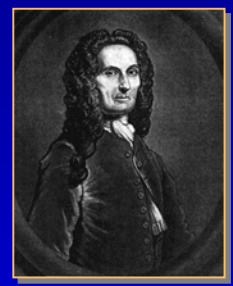
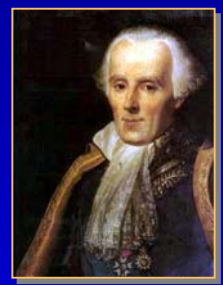
- About 3 000 – 10 000 BE studies / year are conducted worldwide; only ~ 1 – 5% of them are published.
- Although a standard for publishing data of BE studies was already suggested in 1992,[1)]
  - a review in 2002 found only 17 complete data sets on AUC and 12 on C_{max}.[2)]
  - Since no 'real world' data are available, proposed methods (*e.g.*, reference-scaled ABE) rely entirely on simulations!
  - Studies seen by regulators are 'selection biased'.

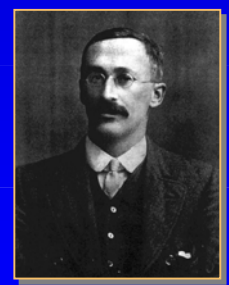
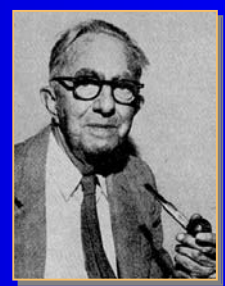
1) Sauter, R., Steinijans, V.W., Diletti, E., Böhm, E. and H.-U. Schulz; Int. J. Clin. Pharm. Ther. Toxicol. 30/Suppl.1, S7-30 (1992)

2) Nakai, K., Fujita, M. and M. Tomita; Int. J. Clin. Pharmacol. Ther. 40, 431-438 (2002)

Bell curve (and beyond?)

- Abraham de Moivre (1667-1754), Pierre-Simon Laplace (1749-1827)
  Central limit theorem 1733, 1812
- Carl F. Gauß (1777-1855)
  Normal distribution 1795
- William S. Gosset, aka Student (1876-1937)
  t-distribution 1908
- Frank Wilcoxon (1892-1965)
  Nonparametric tests 1945

...to be remembered

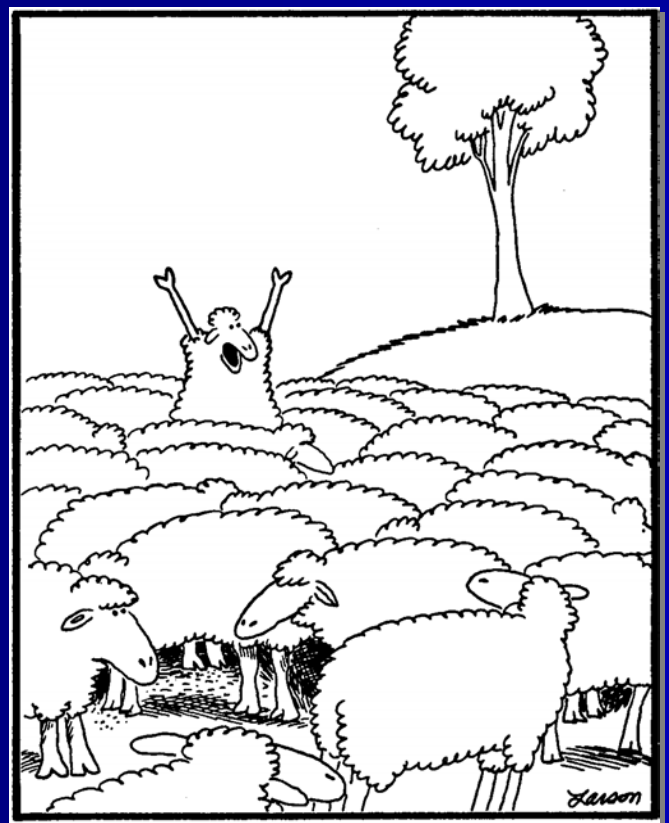
Whenever a theory appears to you as the only possible one, take this as a sign that you have neither understood the theory nor the problem which it was intended to solve.

Karl R. Popper

Even though it's applied science we're dealin' with, it still is – *science!*

Leslie Z. Benet

Conclusions, Outlook



- David Bourne's (Uni. Oklahoma) e-mail list
  - A rather active list (3200 members, about 50 postings/week) covering almost any aspect of PK / PD / BA…
    - Subscription http://www.boomer.org/pkin/
    - Search page http://www.boomer.org/pkin/simple.html
- BA and BE Forum (BEBAC Vienna)
  - Specialized in dissolution / BA / BE / bioanalytics.
    - No registration necessary to *read* postings. http://forum.bebac.at/
    - Registration (to post own questions) http://forum.bebac.at/register.php

Statistical Evaluation of Bioequivalence Studies

Thank You!

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

Important Documents

- EMEA
  - Biostatistical Methodology in Clinical Trials (1993)
  - NfG on the Investigation of BA/BE (2001)
  - Points to Consider on Multiplicity Issues in Clinical Trials (2002)
  - BA/BE for HVDs/HVDPs: Concept Paper (2006)
  - Questions & Answers on the BA and BE Guideline (2006)
- ICH
  - E3: Structure and Content of Clinical Study Reports (1995)
  - E6: Good Clinical Practice (1996)
  - E8: General Considerations for Clinical Trials (1997)
  - E9: Statistical Principles for Clinical Trials (1998)
- WHO
  - Handbook for GCP (2005)
  - Fortieth Report - TRS No. 937 (2006)
    - Annex 7: Multisource (generic) pharmaceutical products: guidelines on registration requirements to establish interchangeability
    - Annex 8: Proposal to waive in vivo bioequivalence requirements for *WHO Model List of Essential Medicines* immediate-release, solid oral dosage forms
    - Annex 9: Additional guidance for organizations performing *in vivo* bioequivalence studies
- US-FDA
  - Statistical Approaches Establishing Bioequivalence (2001)
  - Bioavailability / Bioequivalence – General Considerations (Revision 1, 2003)
- Collection of links to global documents
  http://bebac.at/Guidelines.htm