

Taking a Biostatistical Approach to Designing a BE Study: Ensuring Success through Effective Planning

Part III: Models, Evaluation, Open Issues

Helmut Schütz
BEBAC

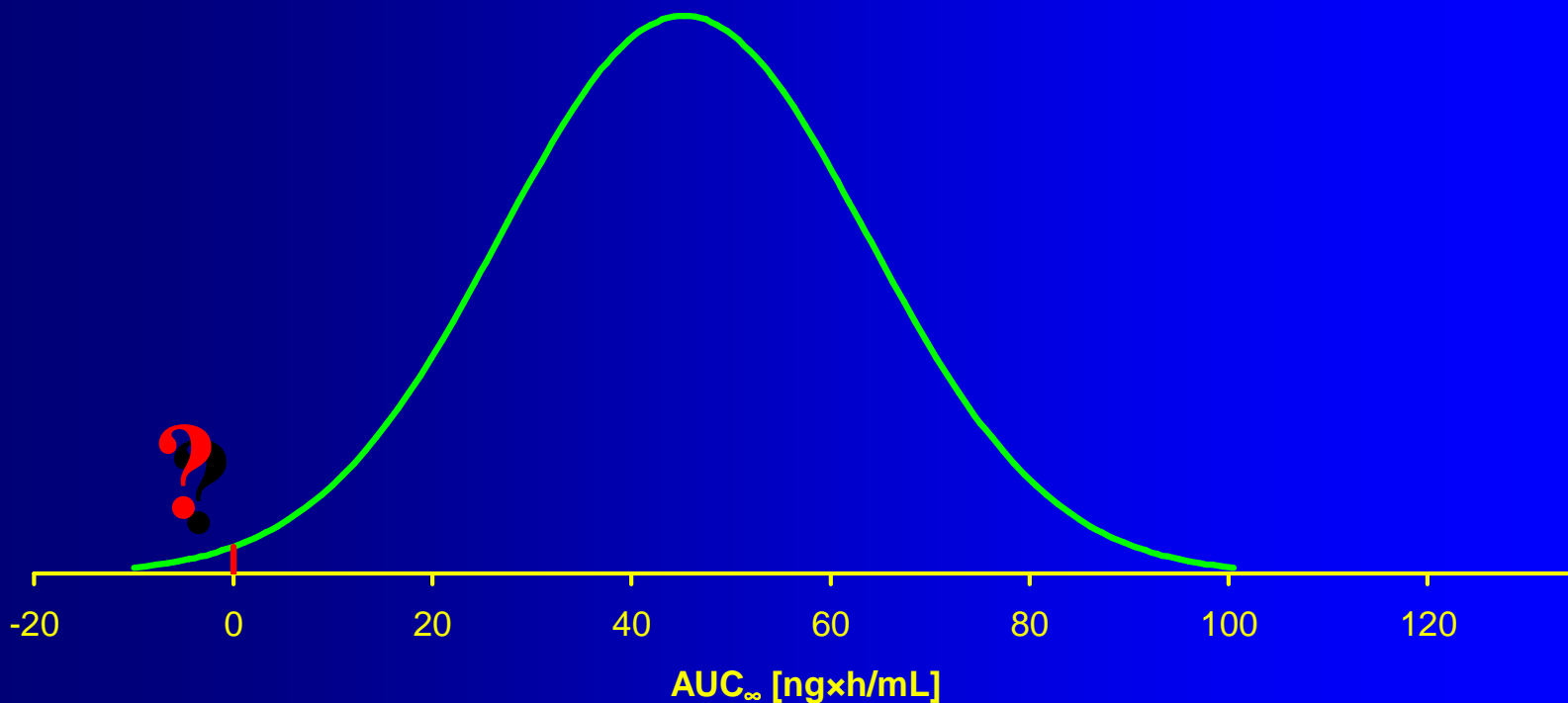
Wikimedia Commons • Stefan Kühn • Creative Commons Attribution-ShareAlike 3.0 Unported

Statistical Assumptions

- Generally accepted methods (e.g., ANOVA) rely on the Normal Distribution
- PK metrics (AUC, C_{max}) of test and reference products follow IDD (Independent Identically Distribution)
- Common sample sizes in BE studies are too small to check this assumption
- Example:
Drug XYZ, 20 mg single dose, 405 subjects;
 AUC_{∞} : mean 45.3 ± 18.4 (CV 40.7%)

Statistical Assumptions

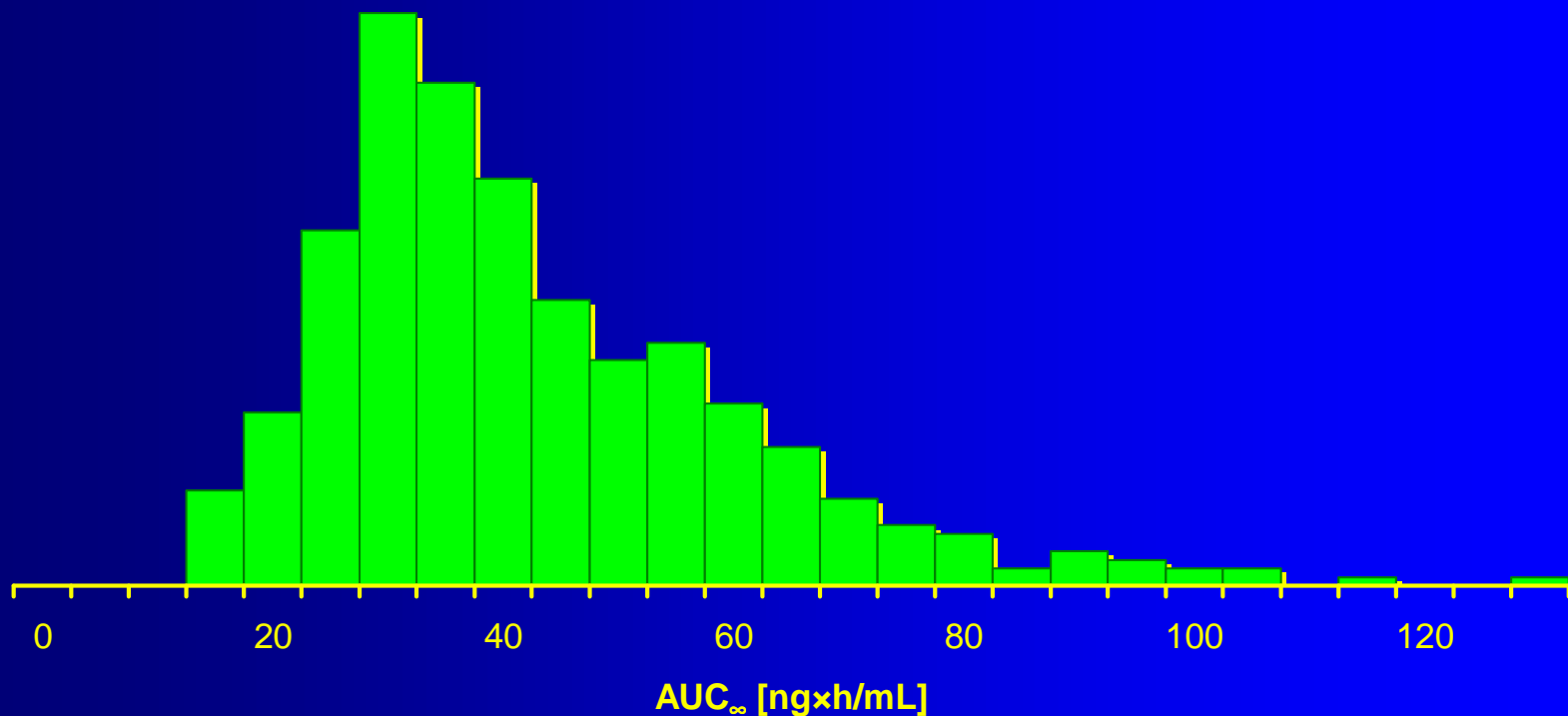
20 mg XYZ s.d. (405 subjects)
mean 45.3 ± 18.4 (CV 40.7%)



Statistical Assumptions

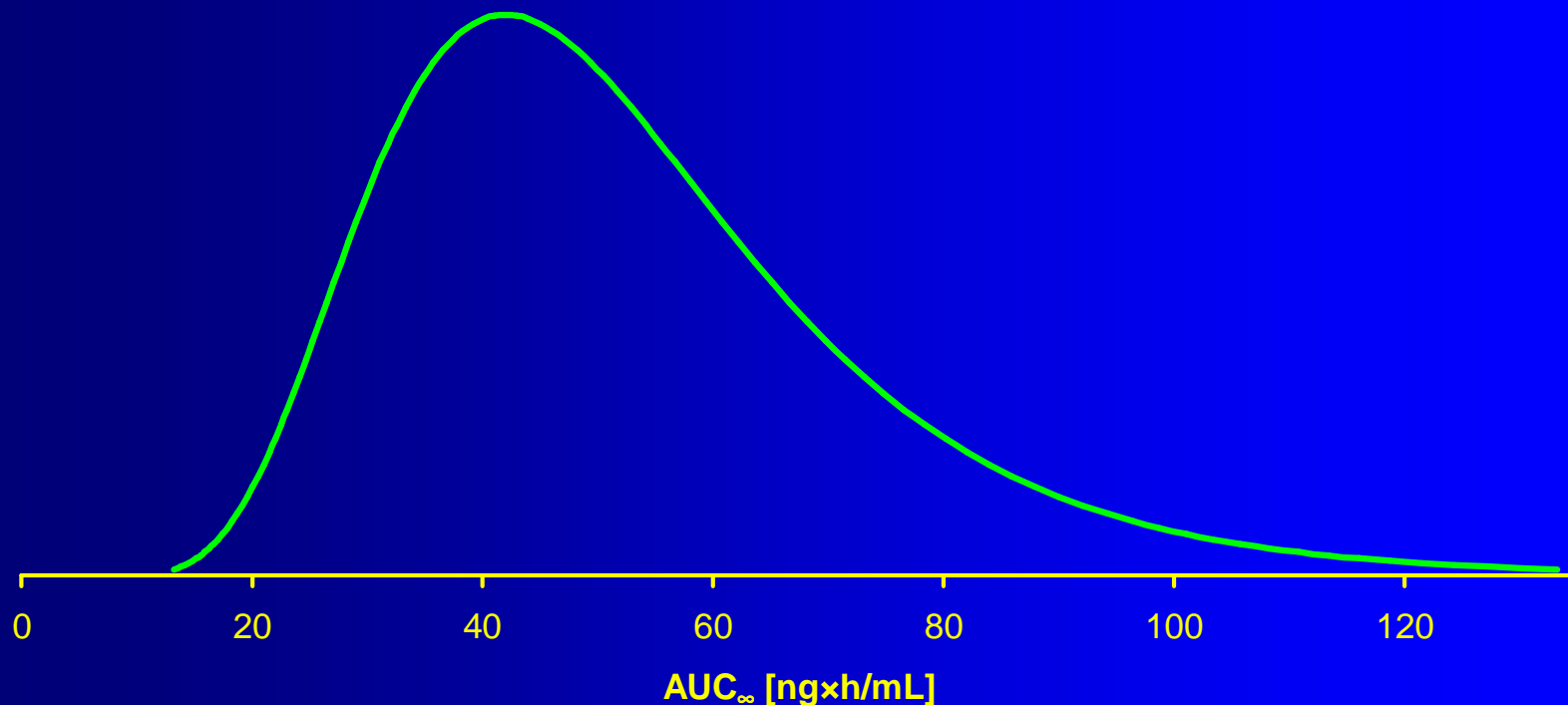
20 mg XYZ s.d. (405 subjects)

min 15.3, Q_1 32.7, median 40.7, Q_3 55.3, max 134.8



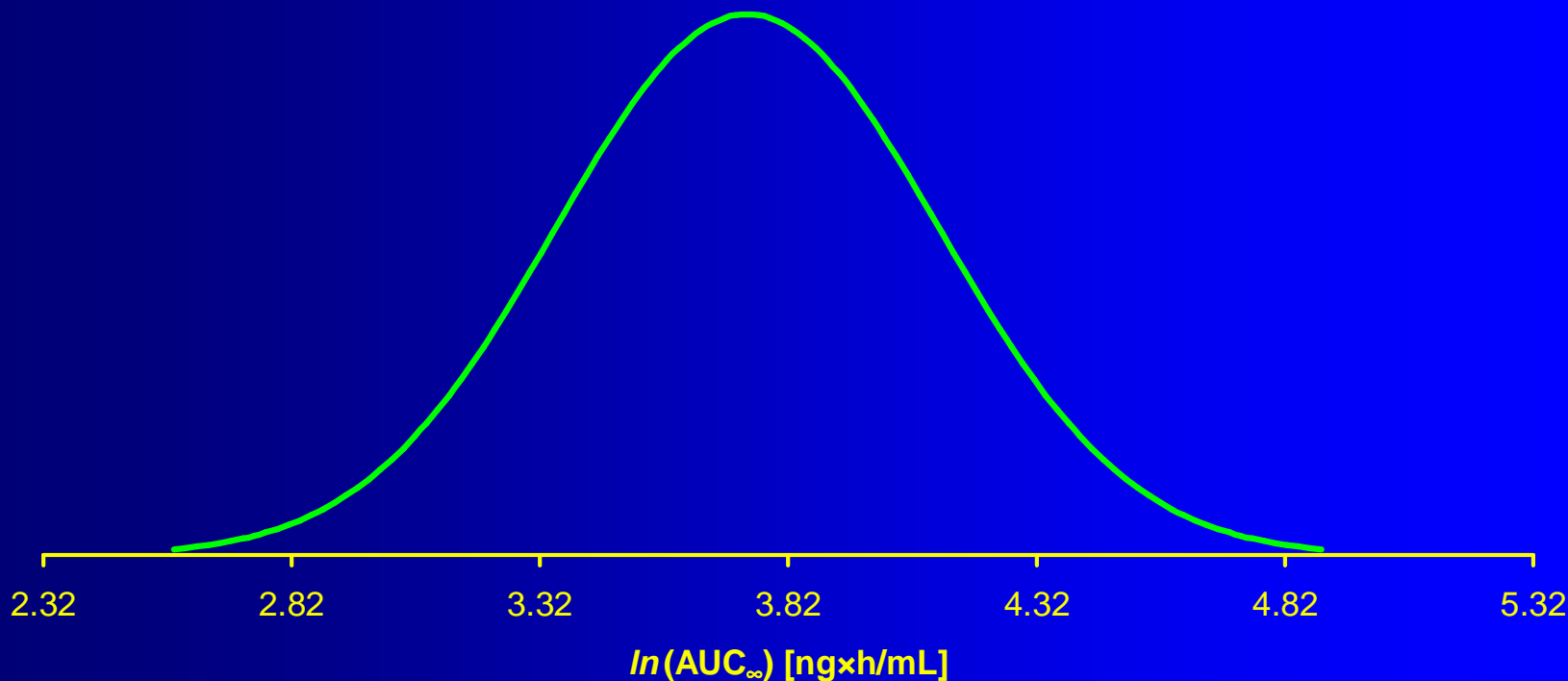
Statistical Assumptions

20 mg XYZ s.d. (405 subjects)
geometric mean 42.0 ± 16.6 (CV 39.4%)



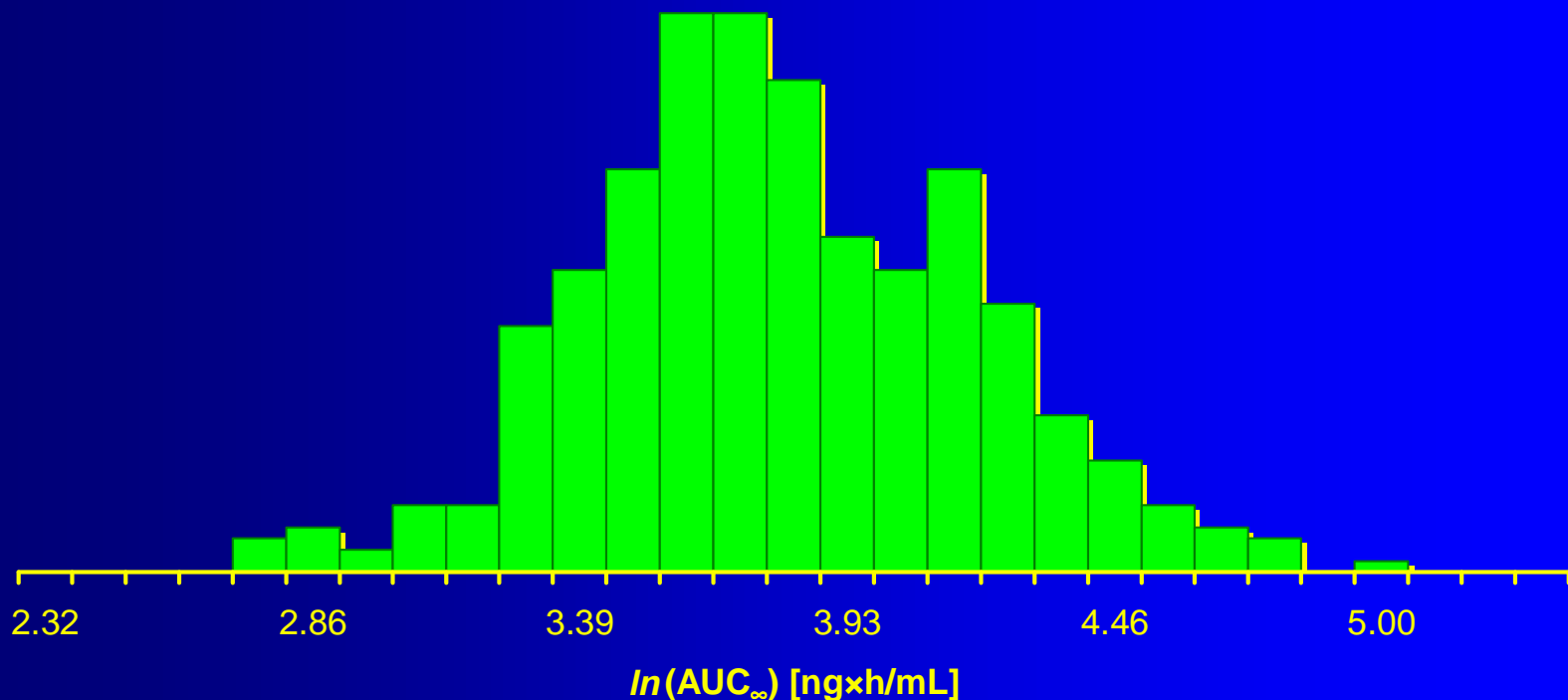
Statistical Assumptions

20 mg XYZ s.d. (405 subjects)
geometric mean 42.0 ± 16.6 (CV 39.4%)



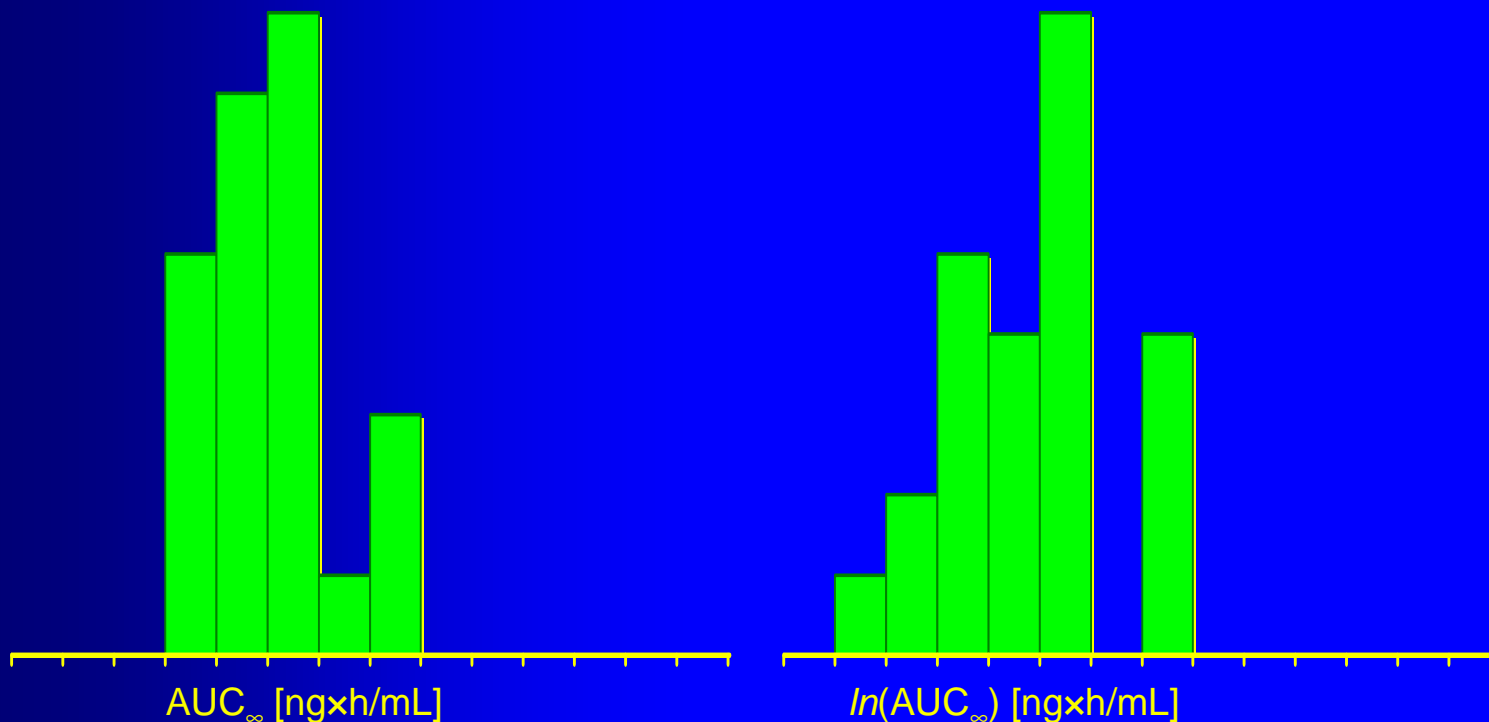
Statistical Assumptions

20 mg XYZ s.d. (405 subjects)
min 2.73, Q_1 3.49, median 3.71, Q_3 4.01, max 4.90



Statistical Assumptions

20 mg XYZ s.d. (24 subjects)



Statistical Assumptions

- BE testing started in the early 1980s with an acceptance range of 80% – 120% of the reference based on the normal distribution.
- Was questioned in mid 1980s
 - Like many biological variables AUC and C_{\max} *do not* follow a normal distribution
 - Negative values are impossible
 - The distribution is skewed to the right
 - Might follow a **lognormal** distribution
 - Serial dilutions in bioanalytics lead to multiplicative errors

Statistical Assumptions

- 'Problems' with logtransformation
 - If we transform the 'old' acceptance limits of 80% – 120%, we get -0.2231, +0.1823.
 - These limits are *not symmetrical* around 100% any more, the maximum power is obtained at $e^{0.1823-0.2231} = 96\% \dots$
 - Solution:
lower limit = $1 - 0.20$, upper limit = $1/\text{lower limit}$
 $\ln(0.80) = -0.2231$ and $\ln(1.25) = +0.2231$.
Symmetrical around 0 in the log-domain and around 100% in the backtransformed domain ($e^0=100\%$).

Statistical Assumptions

- ‘Problems’ with logtransformation
 - Discussion, whether more bioinequivalent formula-
tions will pass due to ‘5% wider’ limits
lower limit = $1 - 0.20$, upper limit = $1/\text{lower}$
80.00% – 125.00% (width **45.00%**)
instead of keeping the ‘old’ width
lower limit = $1 - 0.1802$, upper limit = $1/\text{lower}$
81.98% – 121.98% (width **40.00%**)
or even become more strict by setting
upper limit = $1 + 0.20$, lower limit = $1/\text{upper}$
83.33% – 120.00% (width **36.67%**)
80% – 125% was chosen for convenience (!)

BE-Statistics

- Based on a given design (pilot study \leftrightarrow pivotal, healthy subjects \leftrightarrow patients, single dose \leftrightarrow multiple dose, parallel groups \leftrightarrow cross-over \leftrightarrow replicate)
 - estimate the lowest feasible sample size to meet the aimed target:
 - In a pilot study the CV and test/reference-ratio for further product development or planing a pivotal study;
 - In a pivotal study to meet regulatory requirements (maintaing patient's risk) in demonstrating BE.
 - Write an SAP and evaluate the study.

Power vs. Sample Size

- It is not possible to *directly* calculate the needed sample size.
- Power is calculated instead, and the lowest sample size which fulfills the minimum target power is used.
 - Example: α 0.05, target power 80% ($\beta=0.2$), T/R 0.95, CV_{intra} 20% \rightarrow minimum sample size 19 (power 81%), rounded *up* to the next even number in a 2x2 study (power 83%).

n	power
16	73.54%
17	76.51%
18	79.12%
19	81.43%
20	83.47%

Power Curves

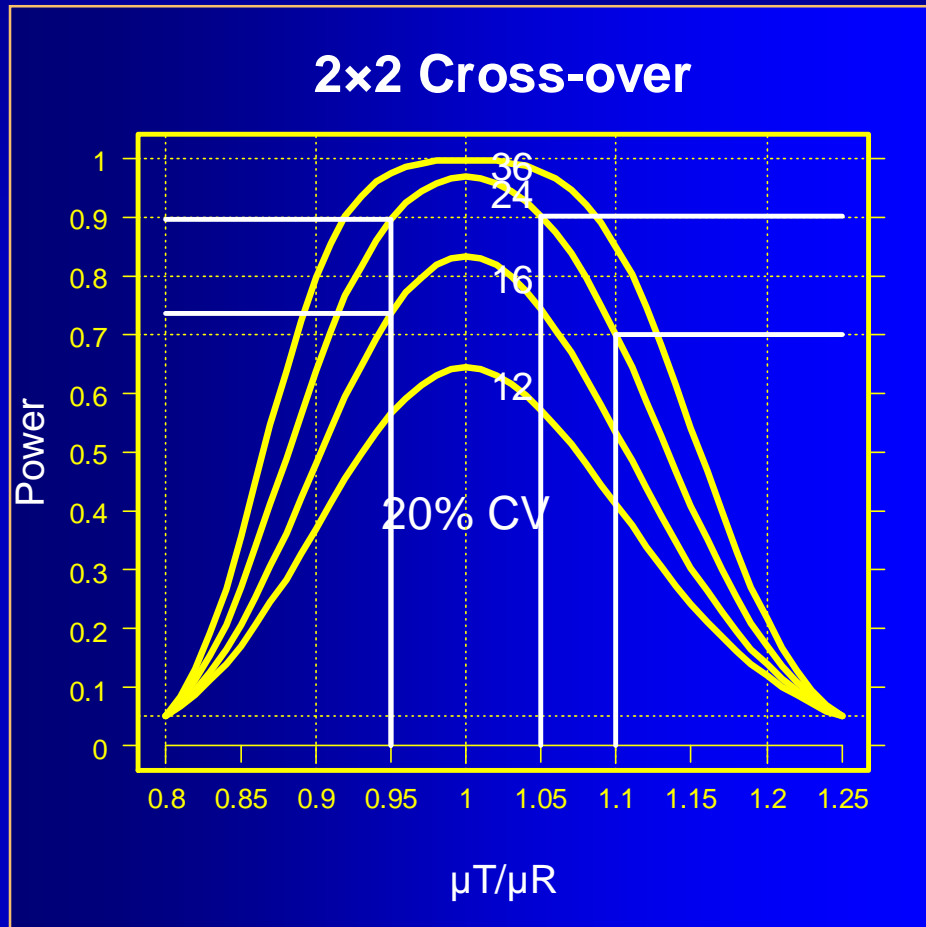
Power to show

BE with 12 – 36 subjects for

CV_{intra} 20%

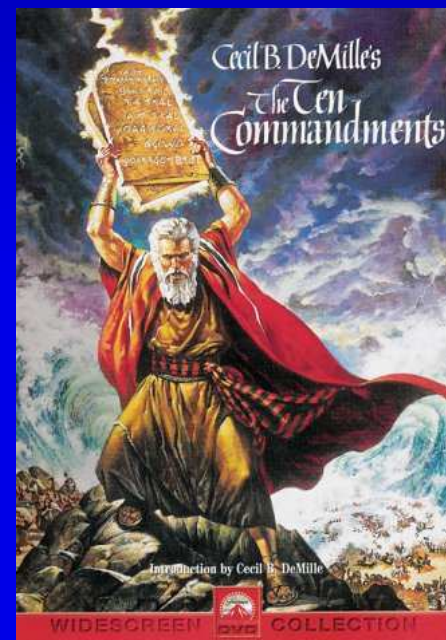
n 24 → 16:
power 0.896 → 0.735

μ_T/μ_R 1.05 → 1.10:
power 0.903 → 0.700



Pilot Studies

- Estimated CV has a high degree of uncertainty (in the pivotal study it is more likely that you 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 you have tested, lesser degrees of freedom will result in worse estimates.
 - Remember: CV is an *estimate* – ***not carved 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)
expn = 10
n.TOST(alpha = 0.05,
targetpower = 0.80, theta1 = 0.80,
theta2 = 1.25, diff = 0.95,
cv = 0.40, dfcv = 22, 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 you 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 you 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.

Pilot Studies: Sample Size

- *Do not* use the pilot study's CV, but calculate an upper confidence interval!
 - Gould (1995) recommends a 75% CI (*i.e.*, a producer's risk of 25%).
 - Apply Bayesian Methods (Julious and Owen 2006, Julious 2010).
 - Unless you are under time pressure, a Two-Stage Sequential Design will help in dealing with the uncertain estimate from the pilot study.

Sequential Designs

- ... have a long and accepted tradition in later phases of clinical research (mainly Phase III).
 - Based on work by Armitage *et al.* (1969), McPherson (1974), Pocock (1977), O'Brien and Fleming (1979) and others.
 - First proposal by LA Gould (1995) in the area of BE did not get regulatory acceptance in Europe, but
 - stated in the current Canadian Draft Guidance (November 2009).
 - Two-Stage Design acceptable in the EU (BE GL 2010, Section 4.1.8)

Sequential Designs

- Penalty for the interim analysis (94.12% vs. 90% CI)
 - Moderate increase in sample sizes
 - Example: T/R 95%, power 80%
 - ~10% increase (sim's by Gould 1995)
 - Comparison to a fixed sample design is based on a delusion – assuming a ‘known’ CV!
 - On the long run (many studies) sequential designs will need *less* subjects.

CV%	90% CI	94.12% CI	ratio
10	8	8	1.000
15	12	14	1.167
20	20	24	1.200
25	28	34	1.214
30	40	48	1.200

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.

'Internal Pilot Study Design'

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

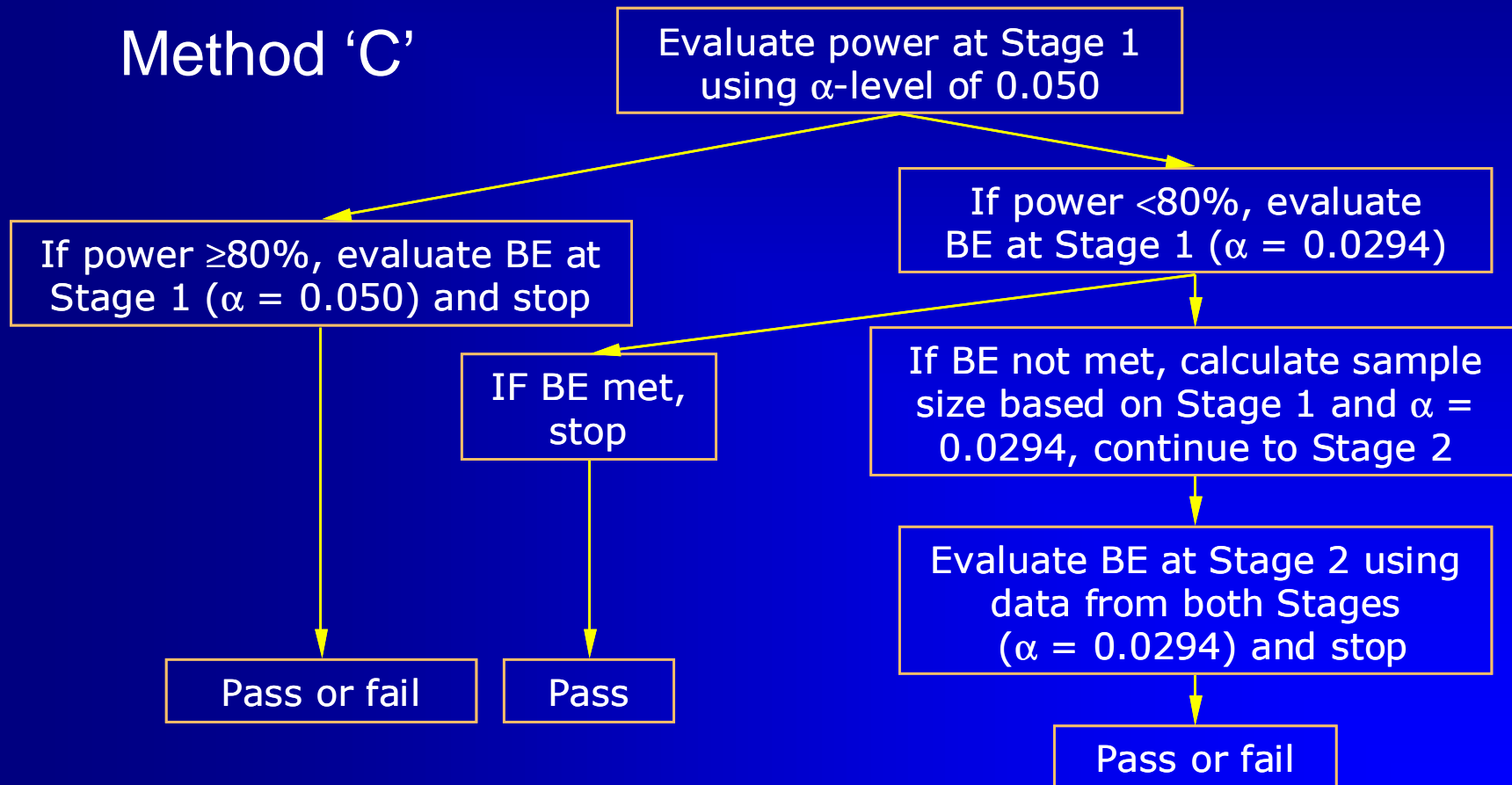
- EMA GL on BE (2010)
 - Section 4.1.8 (cont'd)
 - Plan to use a two-stage approach must be pre-specified in the protocol along with the adjusted significance levels to be used for each of the analyses.
 - When analysing the combined data from the two stages, a term for stage should be included in the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

Two-Stage Design

- Method by Potvin *et al.* (2007) promising
 - Supported by 'The Product Quality Research Institute' (members: FDA-CDER, Health Canada, USP, AAPS, PhRMA,...)
 - Likely to be implemented by US-FDA
 - Should be acceptable as a Two-Stage Design in the EU
 - Two of BEBAC's protocols approved by BfArM and competent EC in May and December 2009

Potvin et al. (2007)

Method 'C'



Potvin *et al.* (2007)

● Technical Aspects

- Only *one* Interim Analysis (after Stage 1)
- If possible, use software (too wide step sizes in Diletti's tables)
- Should be called 'Interim 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
- No stop criterion for Stage 2! Must be clearly stated in the protocol (may be unfamiliar to the IEC, because standard in Phase III).

Potvin et al. (2007)

- Technical Aspects (cont'd)
 - Adjusted α of 0.0294 (Pocock 1977)
 - If power is $<80\%$ in Stage 1 and in the pooled analysis (data from Stages 1 + 2), α 0.0294 is used (*i.e.*, the $1-2\times\alpha=94.12\%$ CI is calculated)

Model	Fixed Effects	Variance Structure	Options	General Options
Confidence Level		<input type="text" value="94.12"/>	%	
Percent of Reference to Detect		<input type="text" value="20"/>	%	
Anderson-Hauck Lower Limit		<input type="text" value="0.8"/>		
Anderson-Hauck Upper Limit		<input type="text" value="1.25"/>		

	Dependent	Ratio_Ref_	CI_User_Lower	CI_User_Upper
1	Ln(Cmax)	105.26625	98.075483	112.98422
2	Ln(AUClast)	99.163003	96.124777	102.29726
3	Ln(AUCINF_pred)	98.82479	95.726146	102.02374

- Overall patient's risk is ≤ 0.0500

Potvin *et al.* (2007)

- Technical Aspects (cont'd)
 - If the study is stopped after Stage 1, the (conventional) statistical model is:
fixed: treatment+period+sequence
random: subject(sequence)
 - If the study continues to Stage 2, the model for the combined analysis is:
fixed: treatment+period+sequence+stage×treatment
random: subject(sequence×stage)
 - No poolability criterion; combining is *always allowed* – even for significant differences between Stages.

Potvin *et al.* (2007)

- Advantage
 - Currently the only *validated* procedure for BE!
- Drawbacks
 - *Not validated* for a correction of effect size (PE) observed in Stage 1 (must continue with the one used in sample size planning).
 - No stop criterion (EMA GL on BE?)
 - Not validated for any other design than the conventional 2x2 crossover (no higher order cross-overs, no replicate designs).

Designs

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

- Hierarchy of designs:

Full replicate (TRTR | RTRT) ↗

Partial replicate (TRR | RTR | RRT) ↗

Standard 2x2 cross-over (RT | TR) ↗

Parallel (R | T)

Power

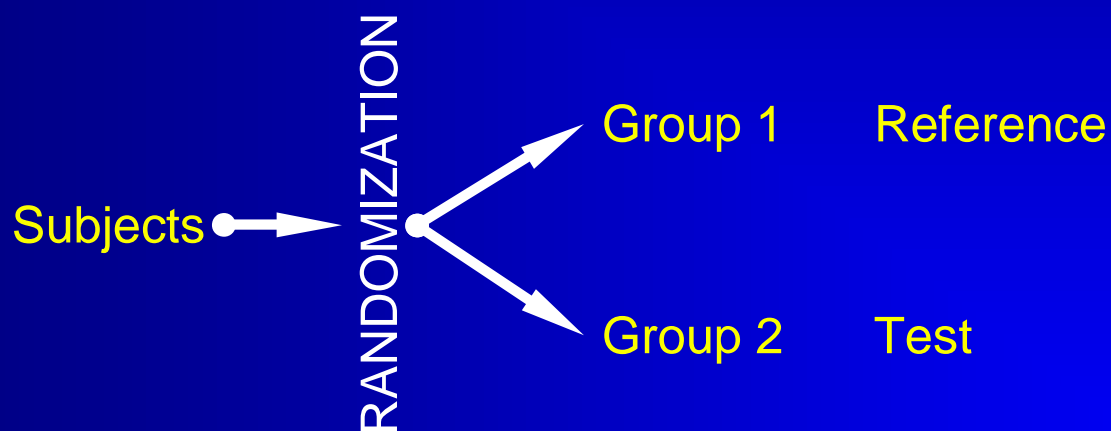
Designs

Power

- Parallel Groups (patients, long half-life drugs)
- Cross-over (generally healthy subjects)
 - Standard 2x2x2
 - Higher Order Designs (more than two formulations)
 - Latin Squares
 - Variance Balanced Designs (Williams' Designs)
 - Incomplete Block Designs
 - Replicate designs

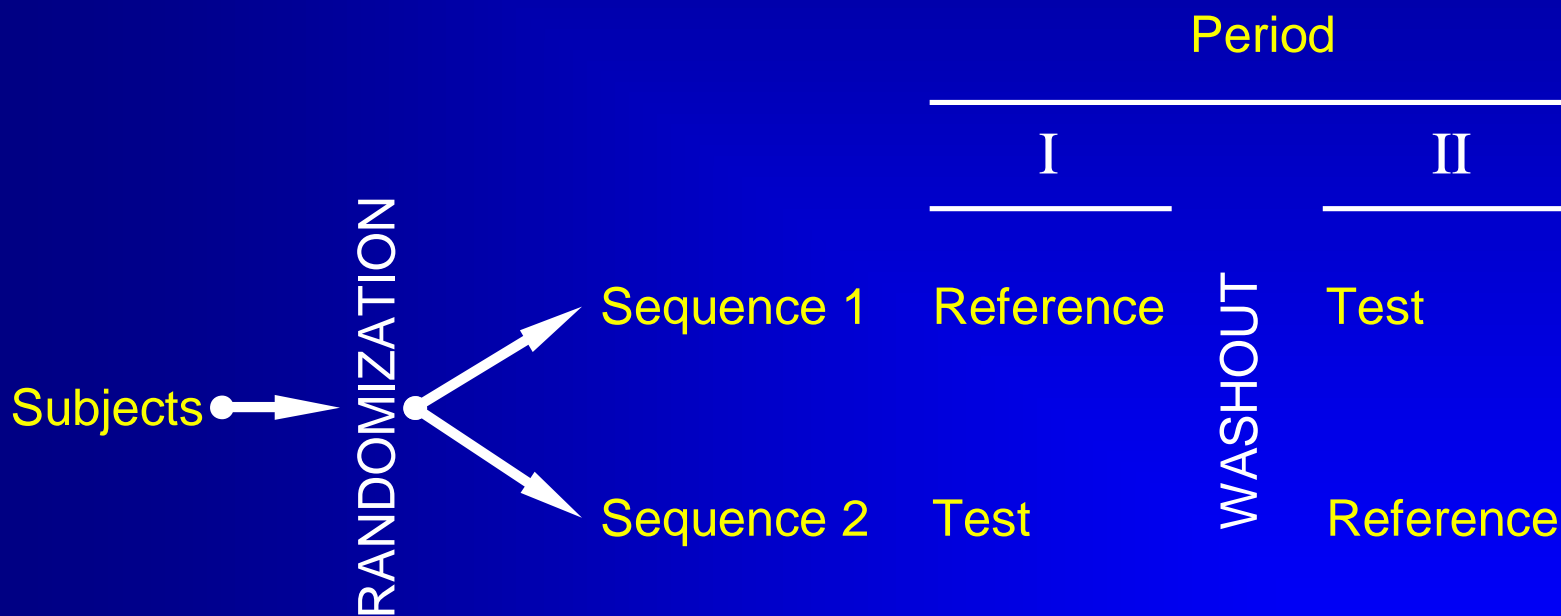
Parallel Groups

- Two-Group Parallel Design



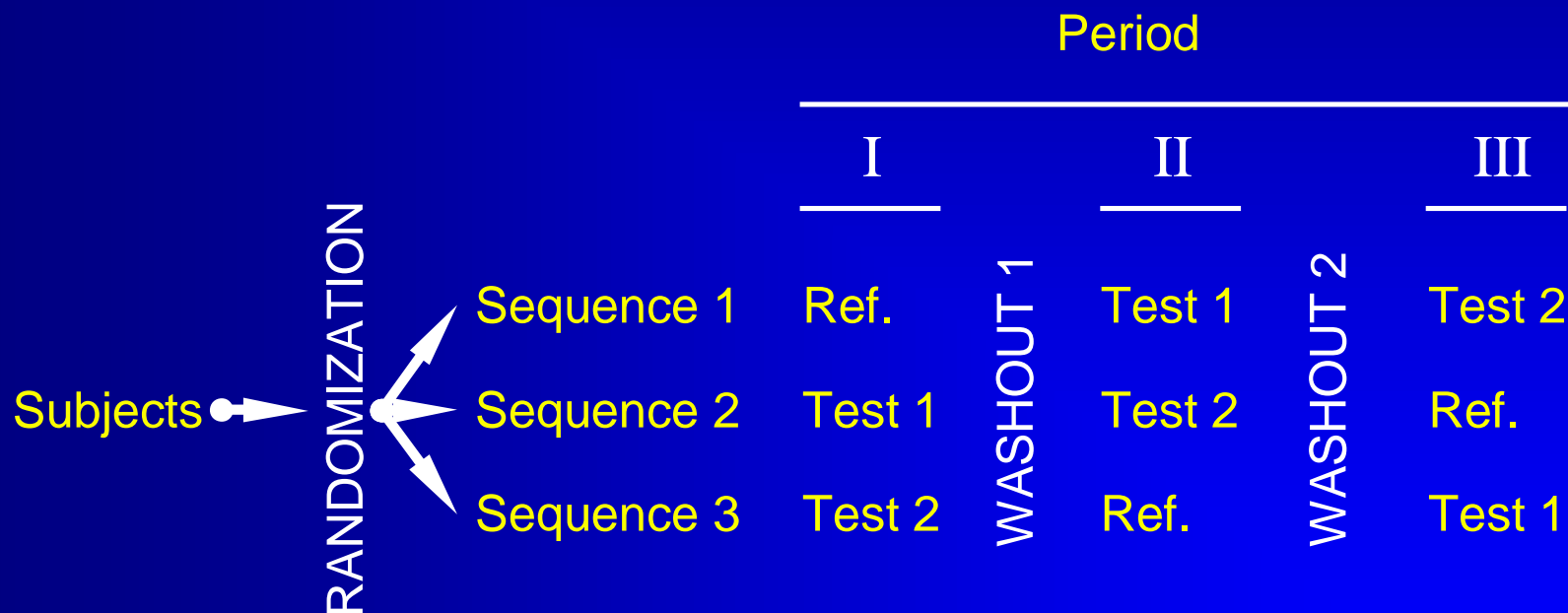
Cross-over Designs

- Standard 2x2x2 Design



Cross-over Designs

- 3x3x3 Latin Square Design



Cross-over Designs

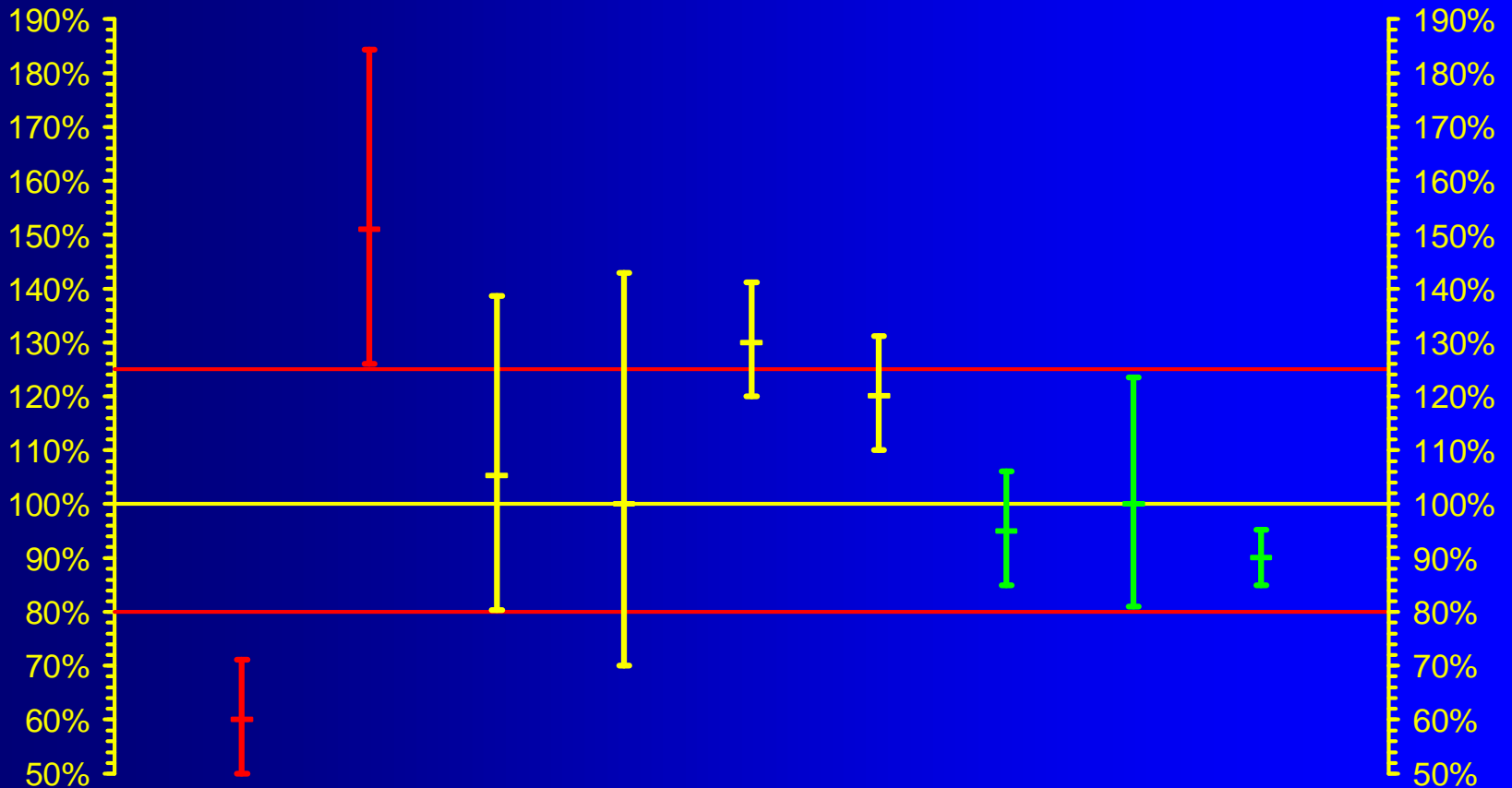
- Williams' Design for three treatments

Sequence	Period		
	I	II	III
1	R	T ₂	T ₁
2	T ₁	R	T ₂
3	T ₂	T ₁	R
4	T ₁	T ₂	R
5	T ₂	R	T ₁
6	R	T ₁	T ₂

BE Assessment

- The *width* of the confidence interval depends on the variability observed in the study.
- The *location* of the confidence interval depends on the observed test/reference-ratio.
- Decision rules:
 - Confidence Interval (CI) entirely outside the Acceptance Range (AR): **Bioinequivalence proven.**
 - CI overlaps the AR, but is not entirely within the AR: **Bioequivalence not proven.**
 - CI entirely within the AR: **Bioequivalence proven.**

BE Assessment



Algebra...

- Calculation of 90% CI (2-way cross-over)
 - Sample size (N) 24, Point Estimate (PE) 102.30%, Residual Mean Squares Error (MSE) from ANOVA (\ln -transformed values) 0.04798, t -value (2α , $N-2$ degrees of freedom) 1.717
 - Standard Error (SE_{Δ}) of the mean difference

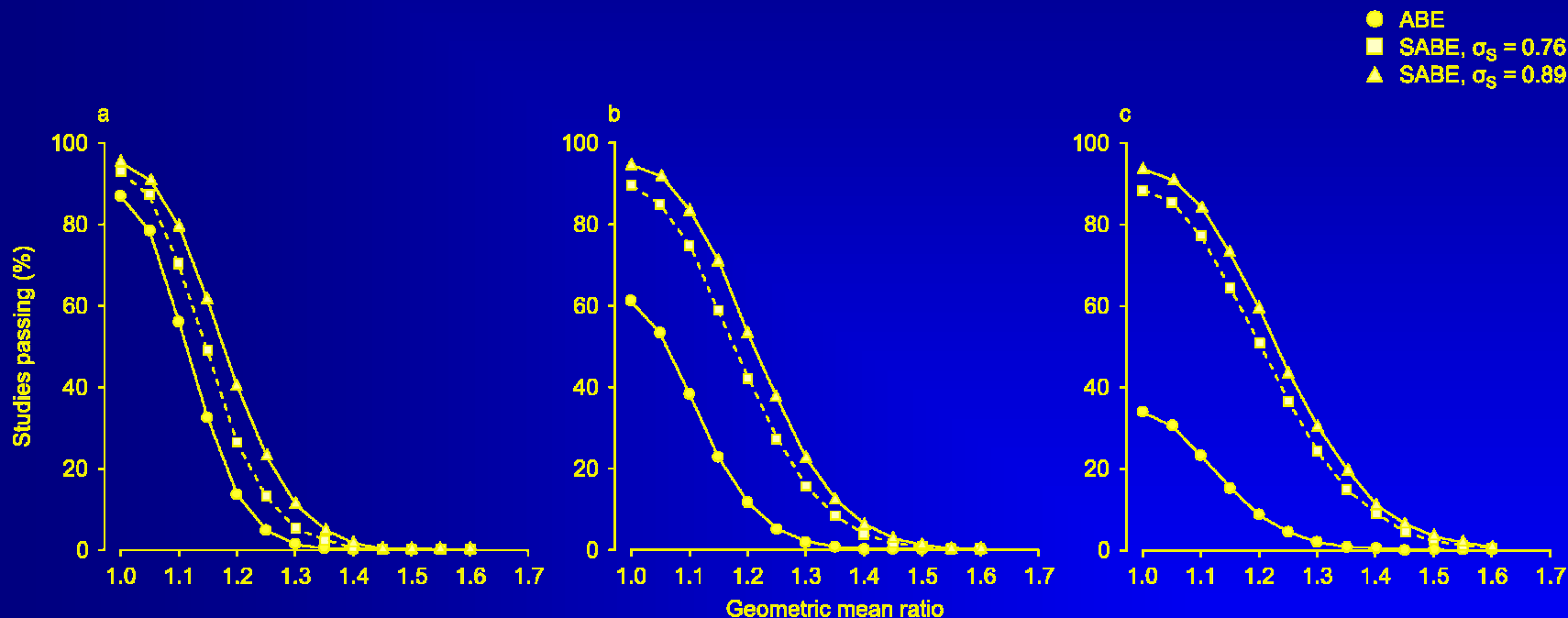
$$SE_{\Delta} = \sqrt{\frac{2 \cdot MSE}{N}} = \sqrt{\frac{2 \times 0.04798}{24}} = 0.063232$$

- Confidence Interval

$$CL_L = e^{\ln PE - t_{2\alpha, df} \cdot SE_{\Delta}} = e^{0.02274 - 1.717 \times 0.063232} = e^{0.02274 - 1.717 \times 0.063232} = 0.9178$$

$$CL_H = e^{\ln PE + t_{2\alpha, df} \cdot SE_{\Delta}} = e^{0.02274 + 1.717 \times 0.063232} = e^{0.02274 + 1.717 \times 0.063232} = 1.1403$$

HVDs/HVDPs



Totfalushi *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.

HVDs/HVDPs


- EU GL on BE (2010)
 - Average Bioequivalence (ABE) with Expanding Limits (ABEL)
 - 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 ($\theta_s=0.760$)

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

HVDs/HVDPs

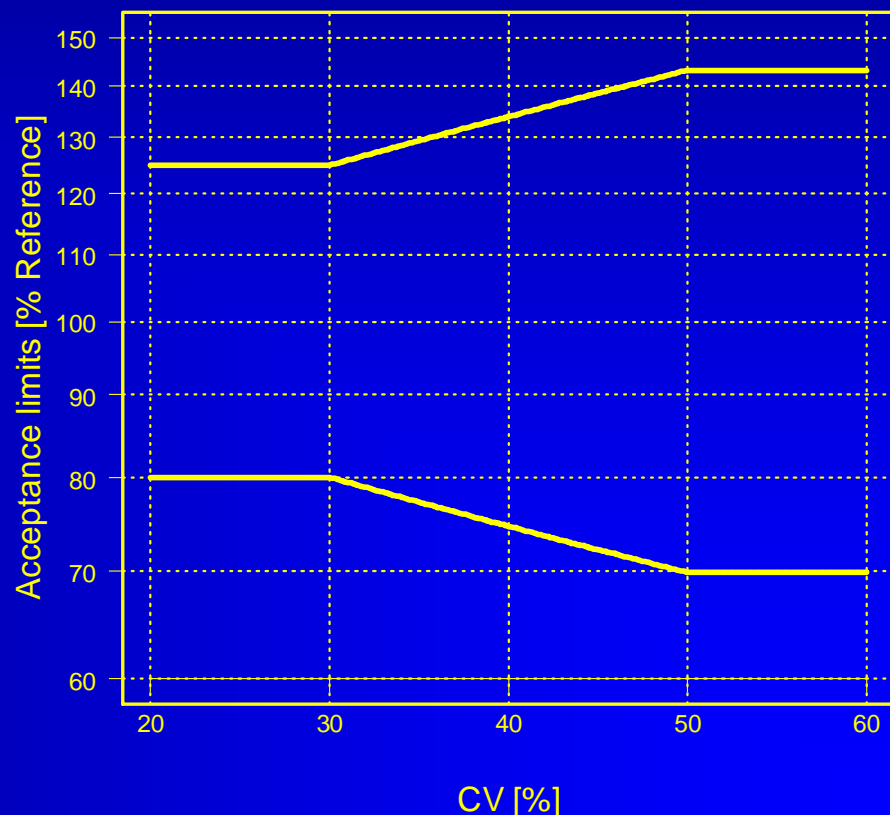
- EU GL on BE (2010)
 - Scaling allowed for C_{\max} only (*not* AUC!) – based on $CV_{WR} > 30\%$ in the actual study (no reference to previous studies).
 - Limited to a maximum of CV_{WR} 50% (*i.e.*, higher CVs are treated *as if* $CV = 50\%$).
 - GMR restricted within 80.00% – 125.00% in any case.
 - At higher CVs only the GMR is of importance!
 - No commercial software for sample size estimation can handle the GMR restriction.
 - Expect a solution from the  community soon...

HVDs/HVDPs

- EU GL on BE (2010)

CV%	L%	U%
30	80.00	125.00
32	78.87	126.79
34	77.77	128.58
36	76.69	130.39
38	75.64	132.20
40	74.61	134.02
42	73.61	135.85
44	72.63	137.68
46	71.68	139.52
48	70.74	141.36
50	69.83	143.20

EU SABE



HVDs/HVDPs

- Replicate designs

- 4-period replicate designs:
sample size = $\frac{1}{2}$ of 2×2 study's sample size
- 3-period replicate designs:
sample size = $\frac{3}{4}$ of 2×2 study's sample size
- Reminder: number of treatments (and biosamples) identical to the 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 bioanalytics has to be improved.

Example ABEL

- RTR–TRT Replicate Design, n=18

Subj	Seq	Per	Trt	Cmax
1	1	1	R	209.91
1	1	2	T	111.05
1	1	3	R	116.36
2	1	1	R	101.16
2	1	2	T	100.31
2	1	3	R	31.71
3	1	1	R	14.83
3	1	2	T	57.10
3	1	3	R	21.47
4	1	1	R	118.71
4	1	2	T	37.34
4	1	3	R	52.29
5	1	1	R	36.11
5	1	2	T	83.95
5	1	3	R	17.76
6	1	1	R	146.44
6	1	2	T	40.45
6	1	3	R	38.34

Subj	Seq	Per	Trt	Cmax
7	1	1	R	58.49
7	1	2	T	62.80
7	1	3	R	123.23
8	1	1	R	105.34
8	1	2	T	103.32
8	1	3	R	43.67
9	1	1	R	59.73
9	1	2	T	169.03
9	1	3	R	48.26
10	1	1	R	38.34
10	1	2	T	31.19
10	1	3	R	19.43
11	2	1	T	51.95
11	2	2	R	195.71
11	2	3	T	65.87
12	2	1	T	18.72
12	2	2	R	20.63
12	2	3	T	7.45

Subj	Seq	Per	Trt	Cmax
13	2	1	T	92.76
13	2	2	R	59.54
13	2	3	T	56.84
14	2	1	T	159.20
14	2	2	R	155.50
14	2	3	T	165.31
15	2	1	T	162.41
15	2	2	R	47.31
15	2	3	T	88.23
16	2	1	T	19.44
16	2	2	R	42.80
16	2	3	T	18.93
17	2	1	T	90.58
17	2	2	R	42.39
17	2	3	T	54.57
18	2	1	T	42.96
18	2	2	R	171.86
18	2	3	T	59.15

Example ABEL

- σ_{WR} (WinNonlin)

	Dependent	Units	Statistic	Value
1	Ln(Cmax)		Difference(Delta)	-0.0011
2	Ln(Cmax)		Ratio(%Ref)	99.8939
3	Ln(Cmax)		SigmaR	0.7319
4	Ln(Cmax)		SigmaWR	0.4628

Calculate the scaled acceptance range based on the regulatory constant k (0.760) and the limiting CV_{WR} :

$$[U, L] = e^{\pm k \cdot \sigma_{WR}} \quad CV_{WR} = \sqrt{e^{\sigma_{WR}^2} - 1}$$

σ_{WR}	0.4628
CV_{WR}	0.4887
L	0.7035
U	1.4215

↻ 30% < CV_{WR} < 50%: Use calculated limits.

Example ABEL

● ABE

PE: 99.89

90% CI:

72.04, 138.52

fails ABE

fails 75 – 133

$30 < CV_{WR} < 50$

[L,U]

70.35, 142.15

passes ABEL

(90% CI within [L,U], PE within 80.00 – 125.00)

```

Bioequivalence Text - [Untitled4] (Read-only) (Derived)

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: Trt
Reference: R    LSMean=    4.069159    SE=    0.173739    GeoLSM=    58.507730
-----
Test:          T    LSMean=    4.068098    SE=    0.174718    GeoLSM=    58.445673

Difference =    -0.0011,    Diff_SE=    0.1876,    df= 16.5
Ratio(%Ref) =    99.8939

          Classical          Westlake
CI 80% = (    77.7639,    128.3217)    (    75.1692,    124.8308)
CI 90% = (    72.0378,    138.5217)    (    67.3124,    132.6876)
CI 95% = (    67.1817,    148.5344)    (    59.4138,    140.5862)
Failed to show average bioequivalence for confidence=90.00 and percent=20.0.

          Two One-Sided T-tests

Prob(< 80%)=0.1266    Prob(> 125%)=0.1244    Max=0.1266    Total=0.2510
    
```

Open Issues

- Studies in both fed and fasted states
 - Acceptable to conduct either two separate two-way cross-over studies or a four-way cross-over study.
 - Recommendation: Separate studies, because variability in fed and fasted state may be different and the treatment effect is statistically confounded with the food effect.
- Limited sampling (truncated AUC_{72})
 - May lead to *'apple-and-orange'* statistics if in a particular subject the last sample is missing or $<LLOQ$ for one of the treatments.

Open Issues

- Limited sampling (truncated AUC_{72}) cont'd
 - Recommendations
 - Truncate the AUC at the last time point where a value $>LLOQ$ is measured for both treatments, or
 - estimate C_{72} from log/linear regression of previous samples.
 - Regulatory acceptance unclear!
- Higher order cross-over studies (e.g., one test vs. two references or $T_{fed}-R_{fed}-T_{fasted}-R_{fasted}$)
 - The analysis for each comparison should be conducted excluding the data from the treatments that are not relevant for the comparison in question.

Open Issues

- Higher order cross-over studies cont'd
 - Minutes of the 3rd EGA Symposium on BE: “Training on the new Revised EMA GL on the Investigation of BE”, 1 June 2010, London
 - However, the treatment, groups, sequences and periods should have their original values maintained in the analysis, and not have the values modified. For example an observation made in period 3 should still be coded as period 3, not have the period changed to “2” because the results for that subject in one of the earlier periods has now be removed.
 - So what?

Open Issues

- Fixed and random effects, ANOVA...
 - Standard cross-over model
 - fixed: `treatment+period+sequence`
 - random: `subject(sequence)`
 - BE GL (Section 4.1.8, Statistical analysis)
 - The terms to be used in the ANOVA model are usually sequence, subject within sequence, period and formulation. Fixed effects, rather than random effects, should be used for all terms.
 - fixed: `treatment+period+sequence+subject(sequence)`
 - Contrary to all (!) textbooks on cross-over designs in bioequivalence...

Open Issues

- Fixed and random effects, ANOVA... cont'd
 - One objective of the new guidance was to completely standardise the method of analysis. While mixed models are generally useful, for bioequivalence ANOVA is considered adequate. [...] A mixed linear models approach would not be acceptable, and subjects with valid data for only one of the two treatments should be excluded. No change. The phrase “or equivalent parametric method” removed to make clear that we are insisting on ANOVA.

EMA, Overview of Comments received on Draft Guideline on the Investigation of Bioequivalence
Doc. Ref. EMA/CHMP/EWP/26817/2010, London, 20 January 2010
http://www.ema.europa.eu/docs/en_GB/document_library/Other/2010/02/WC500073572.pdf

Open Issues

- Fixed and random effects, ANOVA... cont'd
 - Questions at EGA Meeting:
 - According to statisticians of EGA member companies “subject” and “subject within sequence” should be considered as random effects – Which procedure should be used?
 - For replicate design studies mixed effect modelling seems to be necessary in order to get unbiased and separate results for intra-subject variability of test and reference.

Open Issues

- Fixed and random effects, ANOVA... cont'd
 - Answer:
 - Fixed or random models can be used as long as they are pre-specified but fixed is the preferred approach mentioned in the revised guideline. Both approaches should be acceptable and it is unlikely the agency would refuse an application based on the choice of fixed or random. There will be further discussions on the statistical guidance on these models.

My recommendation: Use fixed effects only for balanced datasets (no drop-outs).

Open Issues

- Fixed and random effects, ANOVA... cont'd

- Answer:

- There is an inherent risk if the applicant uses PROC mix and does not remove the missing data prior to evaluation as there will be a fitting of data and this will lead to a difference between PROC Mix and other SAS models. Medicines agencies will accept the use of PROC Mix or other PROC as long as the handling of missing data is pre-defined. Its use will not result in arbitration.

My recommendation: Use a mixed model for replicate design studies.

Thank You!

**Part III: Models, Evaluation,
Open Issues
*Open Questions?***

Helmut Schütz

BEBAC

Consultancy Services for
Bioequivalence and Bioavailability Studies

1070 Vienna, Austria

helmut.schuetz@bebac.at