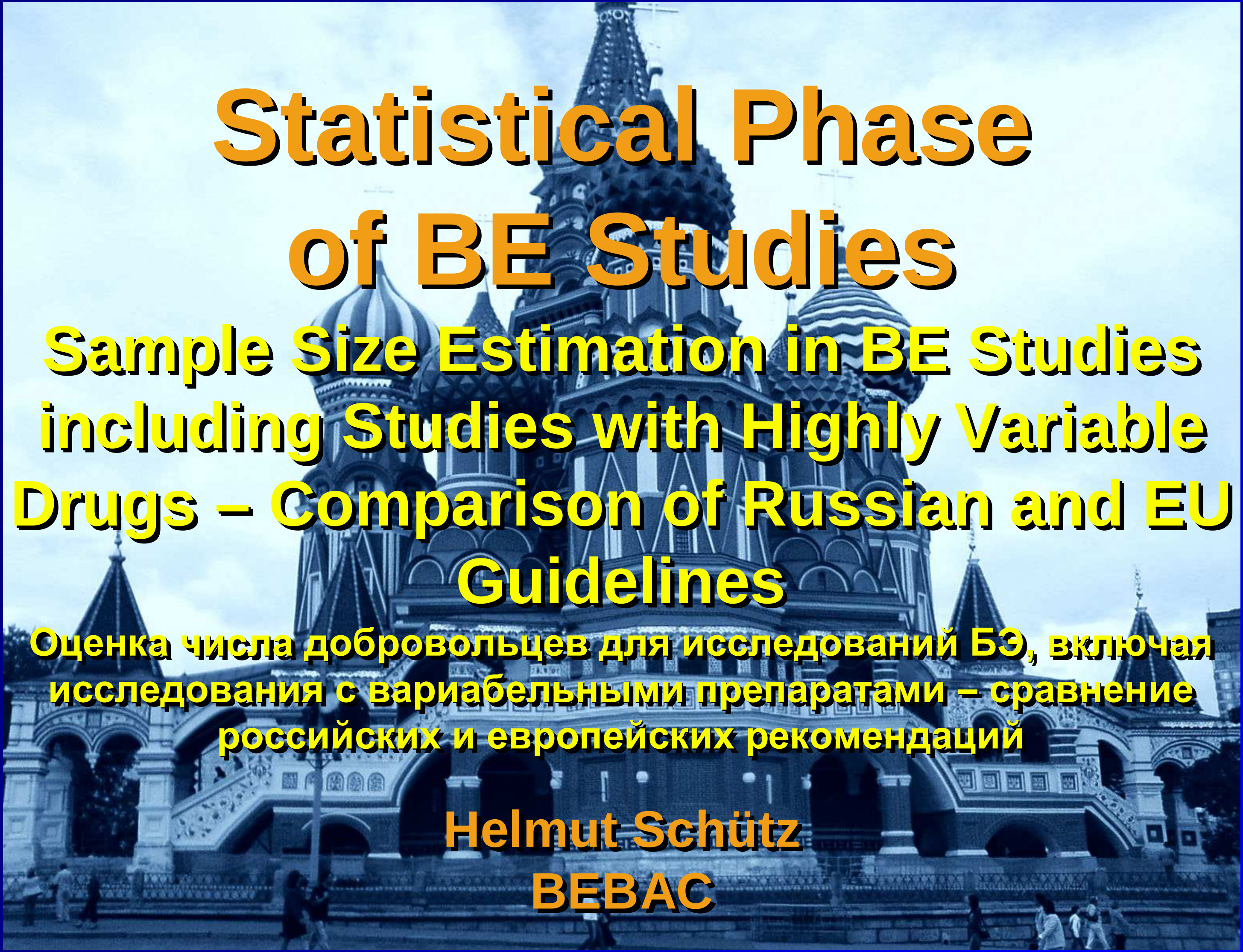


Statistical Phase of BE Studies

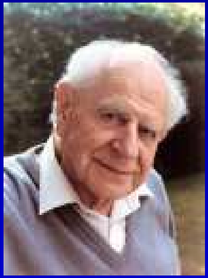
Sample Size Estimation in BE Studies including Studies with Highly Variable Drugs – Comparison of Russian and EU Guidelines

Оценка числа добровольцев для исследований БЭ, включая исследования с вариабельными препаратами – сравнение российских и европейских рекомендаций

Helmut Schütz
BEBAC

To bear in Remembrance...

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.



Когда теория кажется единственно верной, воспринимайте это как знак токо, что вы не поняли ни теорию, ни проблему, которую данная теория описывает.

Karl R. Popper

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

Даже если мы занимаемся *прикладной* наукой – это все равно *Наука*!

Leslie Z. Benet

α and β

- All formal decisions are subjected to two types of error:
  - α Probability of Error Type I (aka Risk Type I)
  - β Probability of Error Type II (aka Risk Type II)

  Example from the justice system:

Verdict	Defendant innocent	Defendant guilty
Presumption of innocence not accepted (guilty)	**Error type I**	**Correct**
Presumption of innocence accepted (not guilty)	**Correct**	**Error type II**

α and β

- Or in more statistical terms:

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

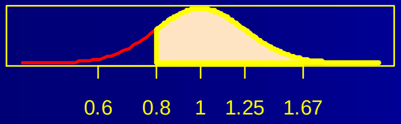
- In BE-testing the null hypothesis is bioinequivalence $(\mu_1 \neq \mu_2)$!

Decision	Null hypothesis true	Null hypothesis false
Null hypothesis rejected	**Patients' risk**	**Correct (BE)**
Failed to reject null hypothesis	**Correct (not BE)**	**Producer's risk**

α ...

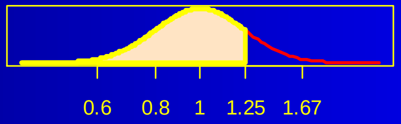
- Patient’s Risk to be treated with an inequivalent formulation (H_0 falsely rejected)
    - BA of the test compared to reference in a *particular* patient is risky *either* below 80% *or* above 125%.
    - If we keep the risk of particular patients at α 0.05 (5%), the risk of the entire population of patients (<80% *and* >125%) is 2×α (10%) – expressed as: 90% CI = 1 – 2×α = 0.90

95% one-sided CI



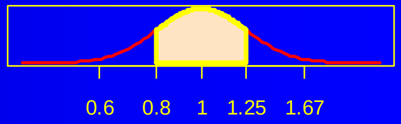
particular patient

95% one-sided CI



particular patient

90% two-sided CI = two 95% one-sided



population of patients

… and β

- Producer's Risk to get no approval for a equivalent formulation (H_0 falsely not rejected)
  - *Set* in study planning to ≤0.2 (20%), where power = 1 – β = ≥80%
  - If power is set to 80 %,
    one out of five studies will fail just by chance!

α **0.05**	**BE**
not BE	β **0.20**

0.20 = 1/5

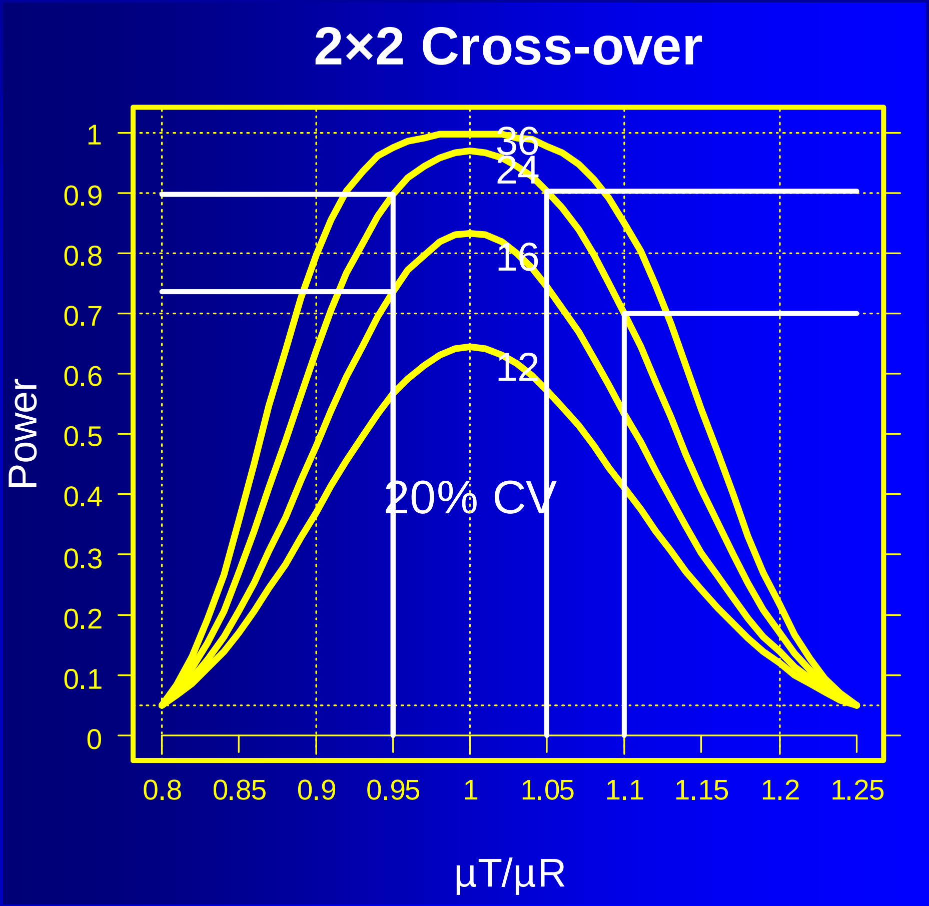
  - *A posteriori* (*post hoc*) power does not make sense!
    Either a study has demonstrated BE or not.

Power (1 – β)

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



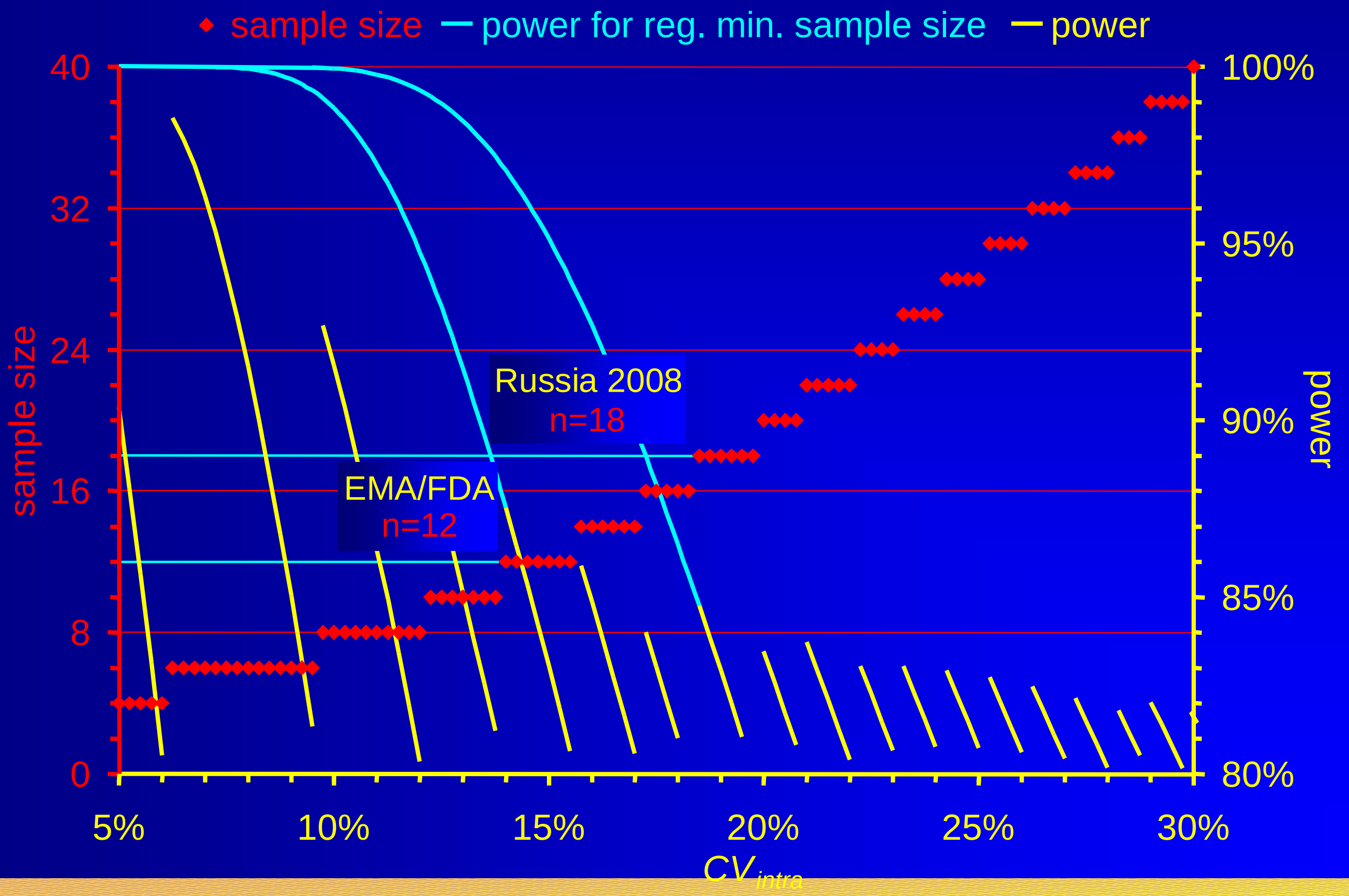
Power *vs.* Sample Size

- It is not possible to *directly* estimate the required *sample size*.
- *Power* is estimated instead; the smallest sample size which fulfills the minimum target power is used.
  - Example: α 0.05, target power 80% (β 0.2), T/R 0.95, CV_{intra} 20% → minimum sample size 19 (power 81%), rounded *up* to the next *even* number in a 2×2 study to get balanced sequences (power 83%)

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

Power *vs.* Sample Size

2×2 cross-over, T/R 0.95, AR 80–125%, target power 80%



Sample Size Estimation

- The estimated *CV* carries some uncertainty (in a pivotal study it is more likely that one will be able to reproduce the pilot’s 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!*
  - Помните: *CV* это всего лишь *оценка* – *она не выбита на скрижалях!*

Uncertainty of CV

- *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 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)

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 ≥80%) based on a n=10 pilot study

CV%	CV fixed	CV uncertain	‘penalty’ uncert./fixed
20	20	24	+20.0%
25	28	36	+28.6%
30	40	52	+30.0%
35	52	68	+30.8%
40	66	86	+30.3%

If pilot n=24:
n=72, penalty +9.1%

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

Pilot Studies: Sample Size

- Moderate sized pilot studies (sample size ~12–24) lead to more consistent results (both *CV* and T/R)
  - If stated in the protocol, BE may be even claimed in the pilot study, and no further study will be necessary (US-FDA).
  - If one has previous hints of high intra-subject variability (>30%), a pilot study size of *at least* 24 subjects is reasonable.

Justification

- Best description given 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.

Good Scientific Practice!

- Influental factors can be tested
  - Sampling schedule: matching C_{max}, lag-time (first point C_{max} problem), reliable estimate of λ_z
  - Food, posture, clinical set-up, …
  - Bioanalytical method: LLOQ, ULOQ, linear range, metabolite interferences, ICSR
  - Select formulations (candidate tests ↔ one reference or one test ↔ >one reference)
  - Variabilty of PK metrics / location of T/R
  - If design issues (formulations, clinics, bioanalytics) are already known, a Two-Stage sequential design might be a better alternative!

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 one needs is the 90% geometric confidence interval and the sample size.

Algebra…

- Calculation of CV_{intra} from Confidence Interval
  - Point estimate (PE) from the Confidence Interval

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

  - Estimate the number of subjects / sequence (example 2×2 cross-over)
    - If total sample size (N) is an even number, *assume* (!) $n_1 = n_2 = ½N$
    - If N is an odd number, *assume* (!) $n_1 = ½N + ½$, $n_2 = ½N – ½$ (*not* $n_1 = n_2 = ½N$!)
  - Difference between one CL and the PE in log-scale; use the CL which is given with more significant digits

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

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
16	8	24.74

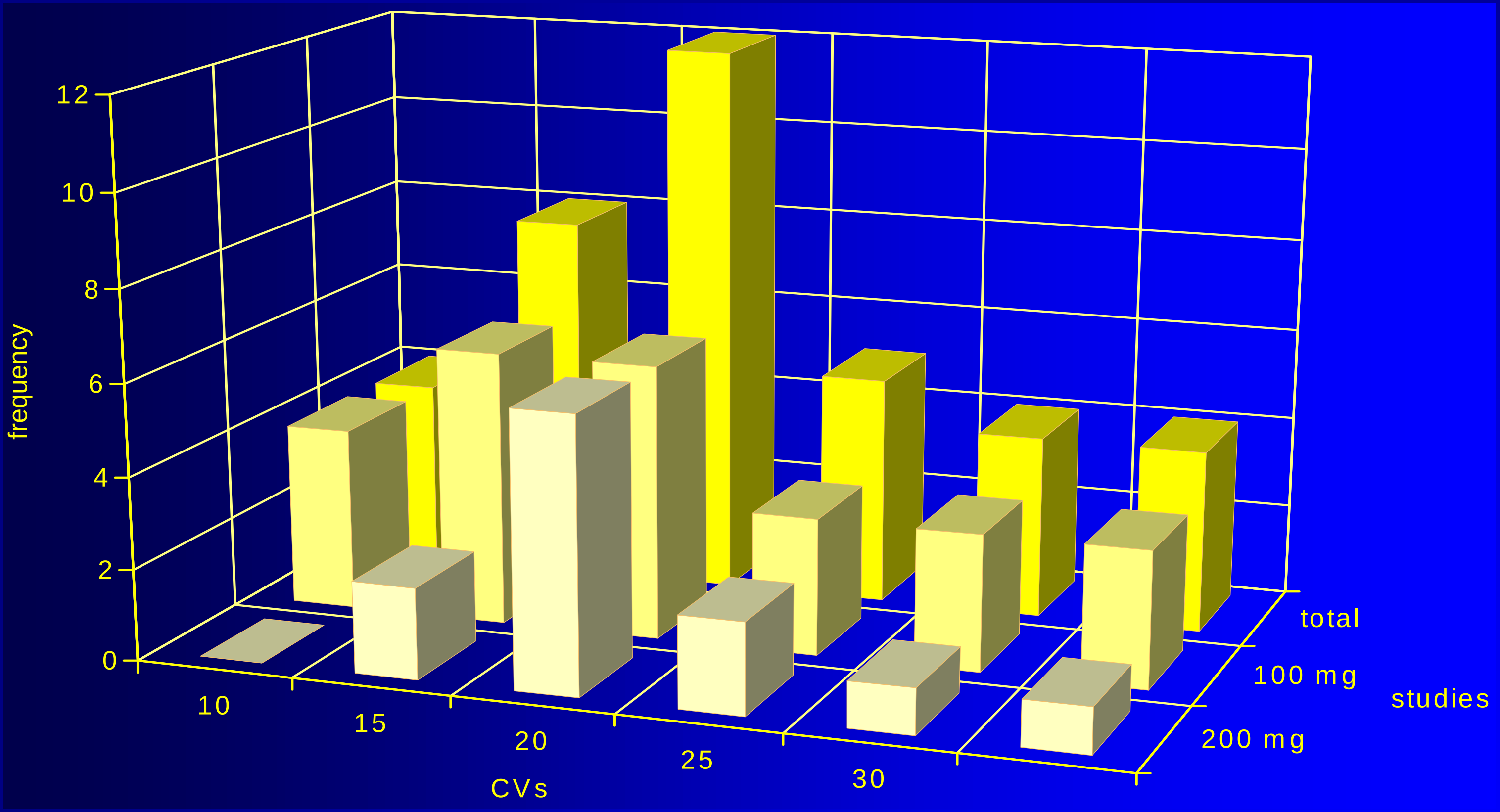
True sequences in study

No Algebra…

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

```
require(PowerTOST)
100*CVfromCI(lower=0.91, upper=1.15, n=21, design='2x2', alpha=0.05)
[1] 22.19886
```

Literature data



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

Sample Size (Guidelines)

- Recommended minimum
  - 12 WHO, EU, CAN, NZ, AUS, AR, MZ, ASEAN States, RSA, Russia?
  - 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 (2008)
  - 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.”*

EMEA

- 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.

EMEA

- 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?).

EMEA

- 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

EMEA

- 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***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 legislations (*e.g.*, Brazil's ANVISA), α must be tightened to 2.5% for NTIDs (95% confidence interval)

EMEA

- 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 – β).
      - 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!

EMEA

- 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
      - EMA GL (2010): content of batches must not differ more than 5%

EMA

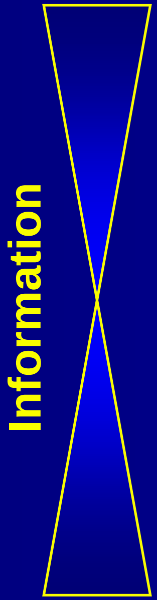
- BE Guideline (2010)
  - The number of subjects to be included in the study should be based on an *appropriate* sample size calculation.

Cookbook?

Hierarchy of Designs

- The more ‘sophisticated’ a design is, the more information can be extracted.
  - Hierarchy of designs:
    Full replicate (TRTR | RTRT or TRT | RTR), ⇘
    Partial replicate (TRR | RTR | RRT) ⇘
    Standard 2×2 cross-over (RT | RT) ⇘
    Parallel (R | T)
  - Variances which can be estimated:

Parallel:	total variance (between + within)
2×2 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%* (Within) →
    - Inter-subject *CV%* (Between)
    - Total *CV%* (Pooled)

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

$$CV_{inter}\% = 100\cdot\sqrt{e^{\frac{MSE_B-MSE_W}{2}}-1}$$

$$CV_{total}\% = 100\cdot\sqrt{e^{\frac{MSE_B+MSE_W}{2}}-1}$$

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, 376–8 (1994)

Coefficient(s) of Variation

- *CVs* of *higher* design levels not accessible.
  - If only mean±SD of reference available…
    - Avoid (often quoted) ‘rule of thumb’ CV_{intra} ~50% of CV_{total}
    - Do not 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

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 can not directly be 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 estimations…
    - ‘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	12.5	14	15	16	17.5	18	20	22
PowerTOST 1.0-00 (2012)	exact	Owen's Q	**4**	**6**	**8**	**8**	**10**	**12**	**12**	**14**	**16**	**16**	**20**	**22**
Patterson & Jones (2006)	noncentr. *t*	AS 243	**4**	**5**	**7**	**8**	**9**	**11**	**12**	**13**	**15**	**16**	**19**	**22**
Diletti *et al.* (1991)	noncentr. *t*	Owen's Q	**4**	**5**	**7**	NA	**9**	NA	**12**	NA	**15**	NA	**19**	NA
nQuery Advisor 7 (2007)	noncentr. *t*	AS 184	**4**	**6**	**8**	**8**	**10**	**12**	**12**	**14**	**16**	**16**	**20**	**22**
FARTSSIE 1.6 (2008)	noncentr. *t*	AS 243	**4**	**5**	**7**	**8**	**9**	**11**	**12**	**13**	**15**	**16**	**19**	**22**
EFG 2.01 (2009)	noncentr. *t*	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 *t*	?	NA	**5**	**7**	**8**	**9**	**11**	**12**	**13**	**15**	**16**	**19**	**22**
Hauschke *et al.* (1992)	approx. *t*		NA	NA	**8**	**8**	**10**	**12**	**12**	**14**	**16**	**16**	**20**	**22**
Chow & Wang (2001)	approx. *t*		NA	**6**	**6**	**8**	**8**	**10**	**12**	**12**	**14**	**16**	**18**	**22**
Kieser & Hauschke (1999)	approx. *t*		**2**	NA	**6**	**8**	NA	**10**	**12**	**14**	NA	**16**	**20**	**24**

			CV%											
original values	Method	Algorithm	22.5	24	25	26	27.5	28	30	32	34	36	38	40
PowerTOST 1.0-00 (2012)	exact	Owen's Q	**24**	**26**	**28**	**30**	**34**	**34**	**40**	**44**	**50**	**54**	**60**	**66**
Patterson & Jones (2006)	noncentr. *t*	AS 243	**23**	**26**	**28**	**30**	**33**	**34**	**39**	**44**	**49**	**54**	**60**	**66**
Diletti *et al.* (1991)	noncentr. *t*	Owen's Q	**23**	NA	**28**	NA	**33**	NA	**39**	NA	NA	NA	NA	NA
nQuery Advisor 7 (2007)	noncentr. *t*	AS 184	**24**	**26**	**28**	**30**	**34**	**34**	**40**	**44**	**50**	**54**	**60**	**66**
FARTSSIE 1.6 (2008)	noncentr. *t*	AS 243	**23**	**26**	**28**	**30**	**33**	**34**	**39**	**44**	**49**	**54**	**60**	**66**
EFG 2.01 (2009)	noncentr. *t*	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 *t*	?	**23**	**26**	**28**	**30**	**33**	**34**	**39**	**44**	**49**	**54**	**60**	**66**
Hauschke *et al.* (1992)	approx. *t*		**24**	**26**	**28**	**30**	**34**	**36**	**40**	**46**	**50**	**56**	**64**	**70**
Chow & Wang (2001)	approx. *t*		**24**	**26**	**28**	**30**	**34**	**34**	**38**	**44**	**50**	**56**	**62**	**68**
Kieser & Hauschke (1999)	approx. *t*		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. Δ = ln(0.8)-ln(T/R) = -0.1719
2. Start with e.g. n=8/sequence
   1. df = n · 2 – 1 = 8 × 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/sequence (N=18.716 → 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/sequence (N=18.3422 → 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 → 19):
   Use 10 subjects/sequence (20 total)

S-C Chow and H Wang (2001)

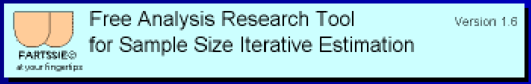
Patient's risk α 0.05, Power 80% (Producer's risk β 0.2), AR [0.80 – 1.25], CV 0.2 (20%), T/R 0.95

1. Δ = ln(T/R) – ln(1.25) = 0.1719
2. Start with e.g. n=8/sequence
   1. df_α = roundup(2·n-2)·2-2 = (2×8-2)×2-2 = 26
   2. df_β = roundup(4·n-2) = 4×8-2 = 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/sequence (N=17.7446 → 18)
   1. df_α = roundup(2·n-2)·2-2=(2×8.8723-2)×2-2 = 30
   2. df_β = roundup(4·n-2) = 4×8.8723-2 = 34
   3. $t_{\alpha,df}$ = 1.6973
   4. $t_{\beta/2,df}$ = 0.8523
   5. new n = $\beta^2 \cdot [(t_{\alpha,df} + t_{\beta/2,df})^2/\Delta^2$ = $0.2^2 \times (1.6973+0.8538)^2 / 0.1719^2$ = 8.8045
4. Convergence reached (N=17.6090 → 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 plausibility of software’s results
- Approximations based on noncentral *t* (FARTSSIE17)



Free Analysis Research Tool for Sample Size Iterative Estimation

Version 1.6

FARTSSIE© at your fingertips

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

or R / 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, theta0=0.95,
  CV=0.30, design='2x2')
```

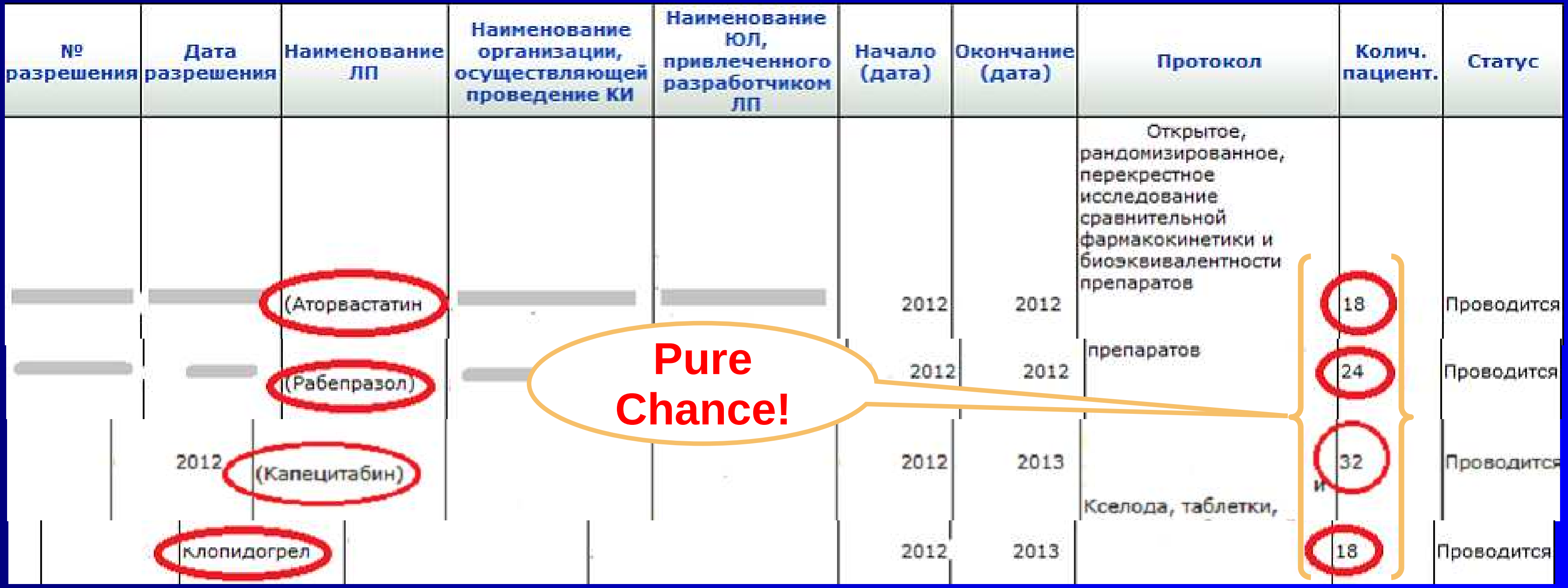
```
alpha   <- 0.05     # alpha
CV      <- 0.30     # intra-subject CV
theta1  <- 0.80     # lower acceptance limit
theta2  <- 1/theta1 # upper acceptance limit
theta0  <- 0.95     # expected ratio T/R
PwrNeed <- 0.80     # minimum power
Limit   <- 1000     # Upper Limit for Search
n       <- 4        # 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(theta0)-log(theta1))/s
  nc2   <- sqrt(n)*(log(theta0)-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')
```

Which Power?

- Generally Producer's Risk 10–20%
  - Plan for 90% – allowing for contingency *e.g.*,
    - drop-outs,
    - CV_{intra} higher than assumed,
    - deviation of test from reference larger than expected.
  - Power >90% might lead to ethical problems ('forced bioequivalence').
  - FDA (2001): 80–90%
  - EMA (2010): 'appropriate'…
  - Russia (2008): ≥80%

Sufficient Sample Size?!

- Atorvastatin, Rapeprazol, Capecitabine, Clopidogrel: Highly Variable Drugs!



№ разрешения	Дата разрешения	Наименование ЛП	Наименование организации, осуществляющей проведение КИ	Наименование ЮЛ, привлеченного разработчиком ЛП	Начало (дата)	Окончание (дата)	Протокол	Колич. пациент.	Статус
[redacted]	[redacted]	(Аторвастатин	[redacted]	[redacted]	2012	2012	Открытое, рандомизированное, перекрестное исследование сравнительной фармакокинетики и биоэквивалентности препаратов	18	Проводится
[redacted]	[redacted]	(Рабепразол)	[redacted]		2012	2012	препаратов	24	Проводится
	2012	(Капецитабин)			2012	2013	Кселода, таблетки,	32 и	Проводится
		клопидогрел			2012	2013		18	Проводится

Pure Chance!

Friendly Reminder

Дружественное напоминание

4.2. Число испытуемых

В исследование должно быть включены испытуемые в количестве достаточном для обеспечения статистической значимости исследования. При этом мощность статистического теста для проверки биоэквивалентности должна поддерживаться на уровне не меньше 80% для выявления 20%-ных различий между основными показателями сравнения.

End of the Story?

- *'Doing the maths'* is just *part* of the job!
  - Does it make sense to rely on 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 variability.
    - Generally insufficient information about the clinical setup (*e.g.*, posture control).
    - Review studies critically; don't try to mix oil with water.

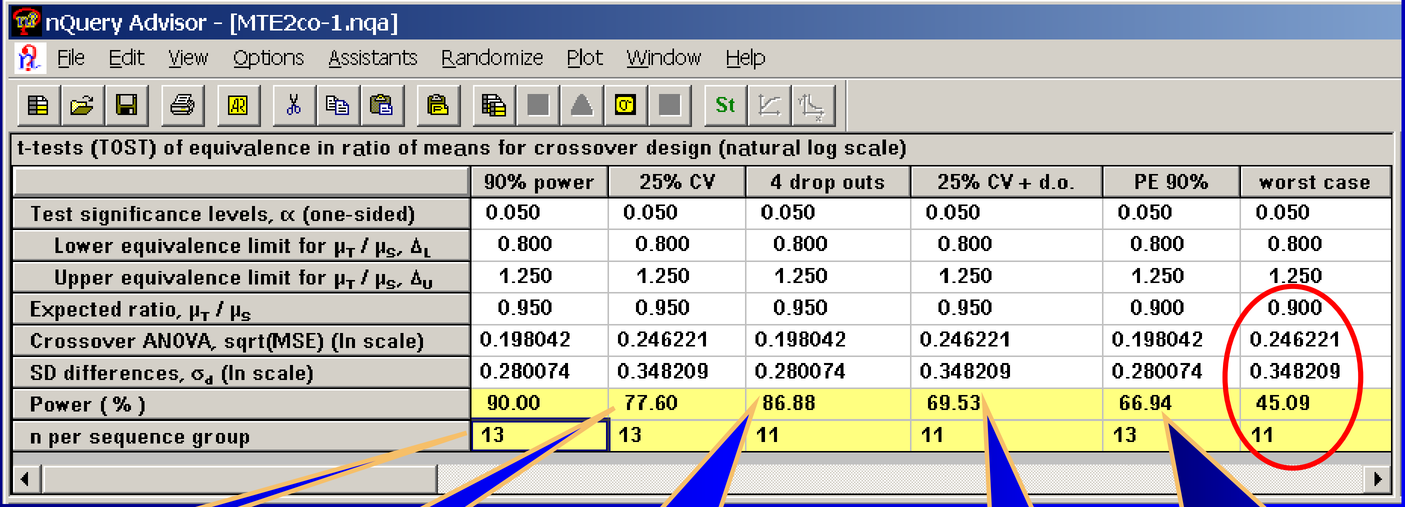
Sensitivity Analysis

- 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
St

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 μ_T / μ_S, Δ_L	0.800	0.800	0.800	0.800	0.800	0.800
Upper equivalence limit for μ_T / μ_S, Δ_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(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, theta0=0.95,
             theta1=0.8, theta2=1.25, CV=0.2, design='2x2')

+++++++++++ 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 estimate Power for a given sample size, use function *power.TOST*

```
require(PowerTOST)
power.TOST(theta0=0.95, CV=0.25, n=26)
[1] 0.7760553

power.TOST(theta0=0.95, CV=0.20, n=22)
[1] 0.8688866

power.TOST(theta0=0.95, CV=0.25, n=22)
[1] 0.6953401

power.TOST(theta0=0.90, CV=0.20, n=26)
[1] 0.6694514

power.TOST(theta0=0.90, CV=0.25, n=22)
[1] 0.4509864
```

Sensitivity Analysis

- Must be done *before* the study *(a priori)*
- The Myth of retrospective (*a posteriori* or *post hoc*) Power…
  - High power does not further support the claim of already demonstrated bioequivalence.
  - Low power does not invalidate a bioequivalent formulation.
  - Further reader:

    RV Lenth (2000)
    JM Hoenig and DM Heisey (2001)
    P Bacchetti (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.

RV Lenth
Two Sample-Size Practices that I don't recommend
http://www.math.uiowa.edu/~rlenth/Power/2badHabits.pdf

Add-on / Two-Stage Designs

- Sometimes properly designed and executed studies fail due to
  - pure chance (producer's risk hit),
  - false (over-optimistic) 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.

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
    - new methods stated in recent guidelines.

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, …)
    - Two Methods (B/C) for T/R 0.95 and 80% power.
    - Simulations for n_1 12–60 and CV 10–100%.
    - Three of BEBAC's protocols accepted by German BfArM, one product approved in 06/2011.

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

Review of Guidelines

- Canada (May 2012)

  Potvin *et al.* Method C recommended.

- FDA (Jun 2012)

  Potvin *et al.* Method C recommended.

  API specific guidances: Loteprednol, Dexamethasone / Tobramycin.

- EMA (Jan 2010)

  Acceptable; Potvin *et al.* Method B preferred.

- Russia?

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.

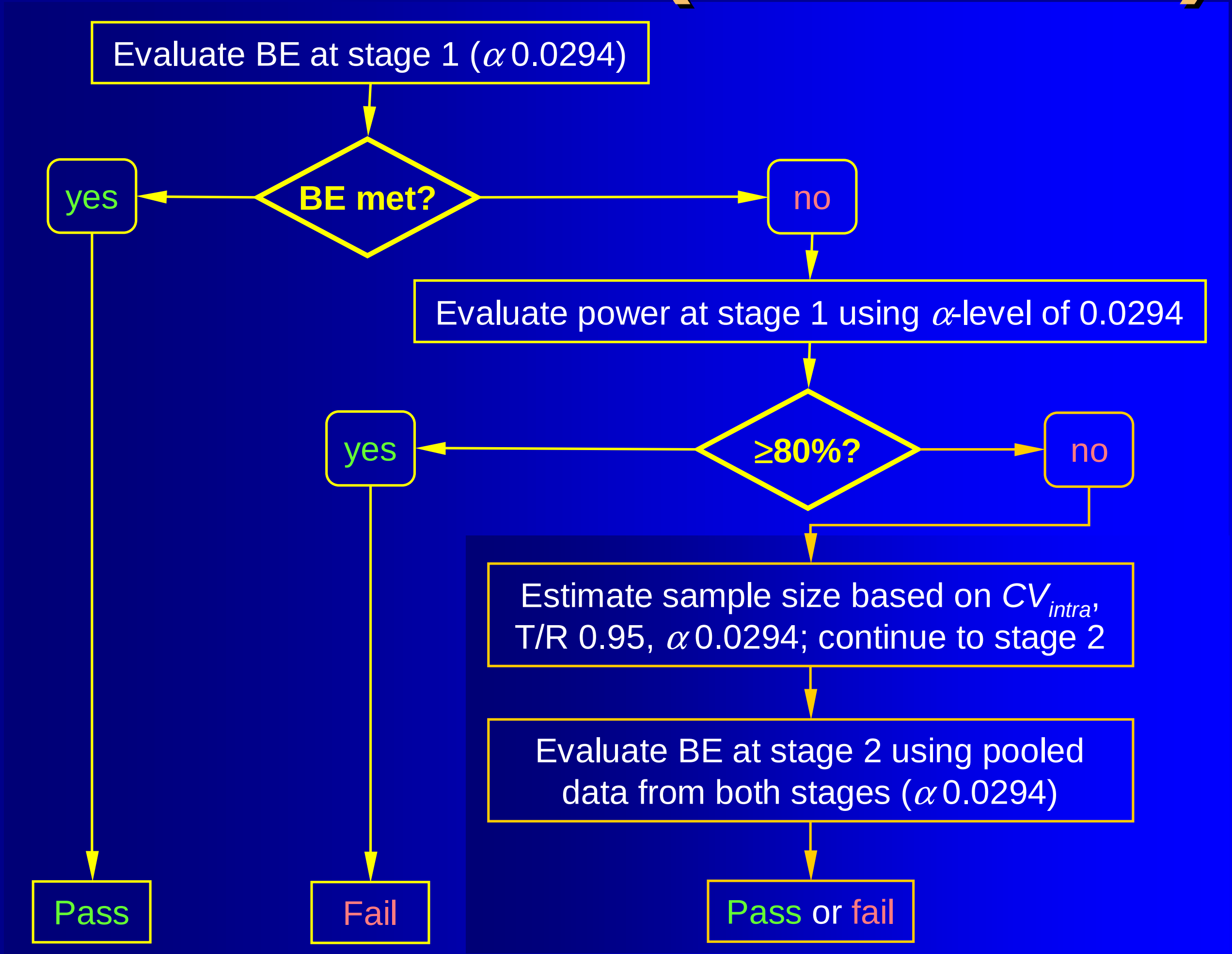
Two-Stage Design

- 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.

Potvin *et al.* (Method B)



Evaluate BE at stage 1 (α 0.0294)

BE met?

yes

no

Evaluate power at stage 1 using α-level of 0.0294

≥80%?

yes

no

Estimate sample size based on CV_{intra}, T/R 0.95, α 0.0294; continue to stage 2

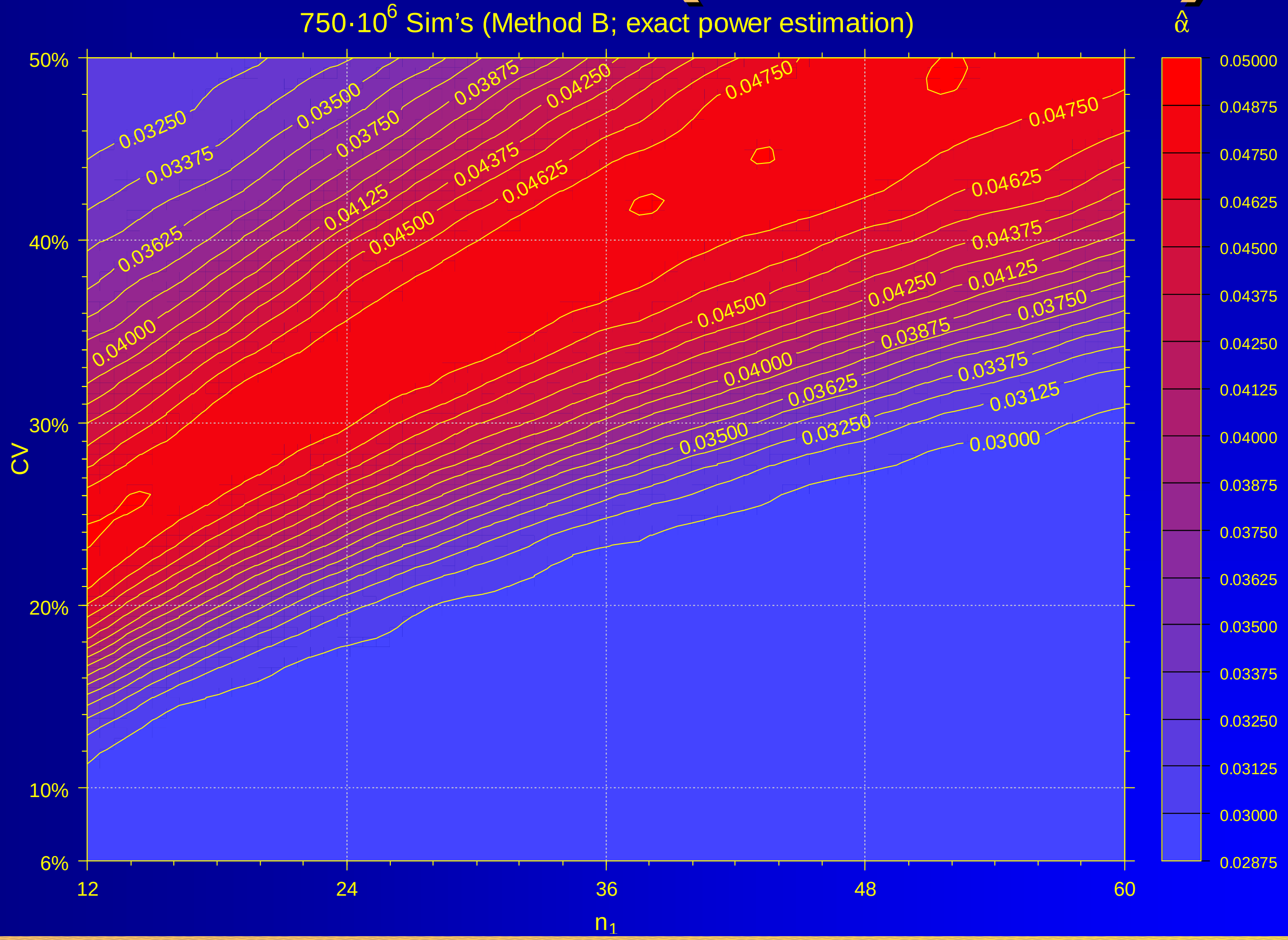
Evaluate BE at stage 2 using pooled data from both stages (α 0.0294)

Pass

Fail

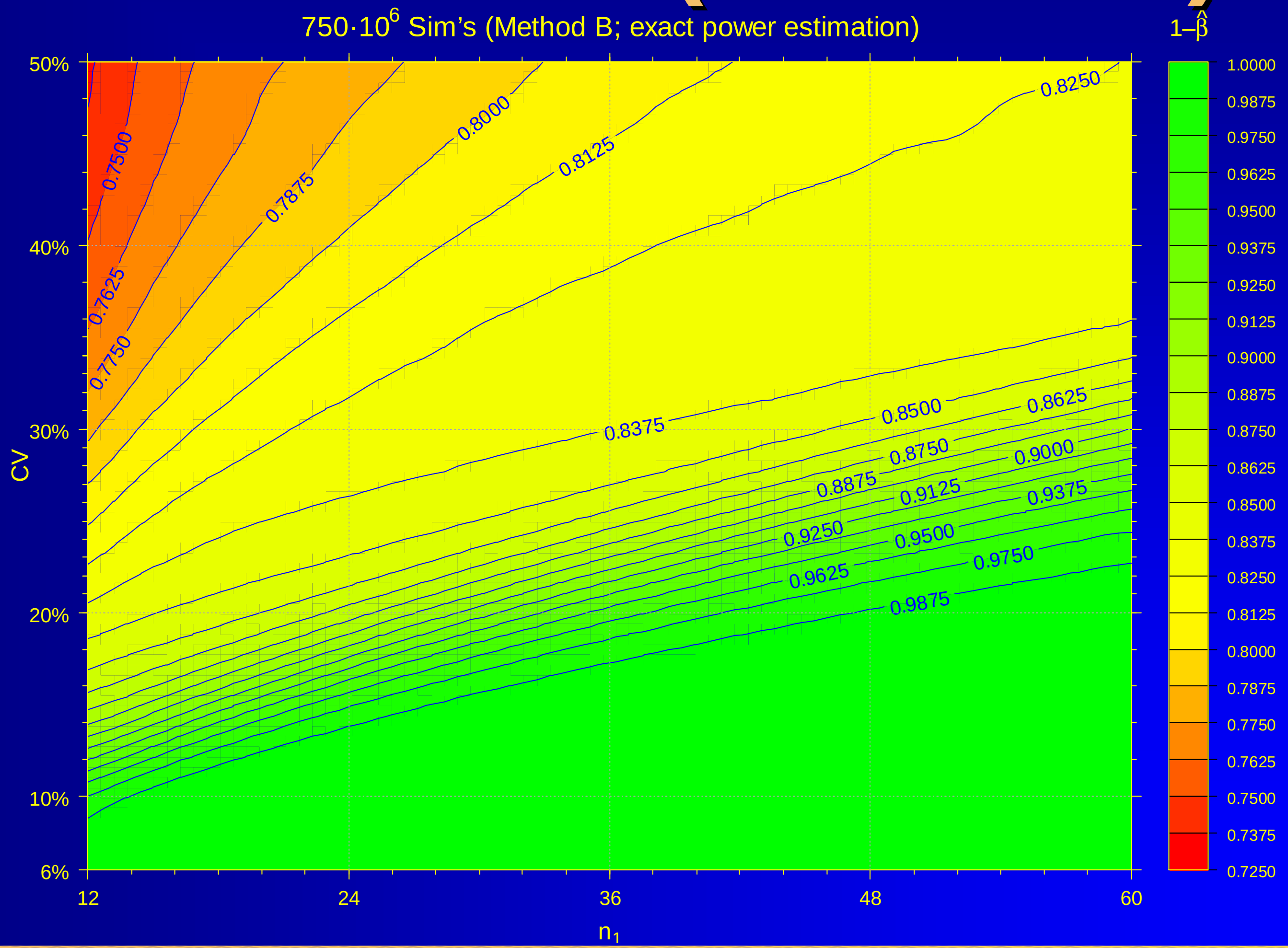
Pass or fail

Potvin *et al.* (Method B)

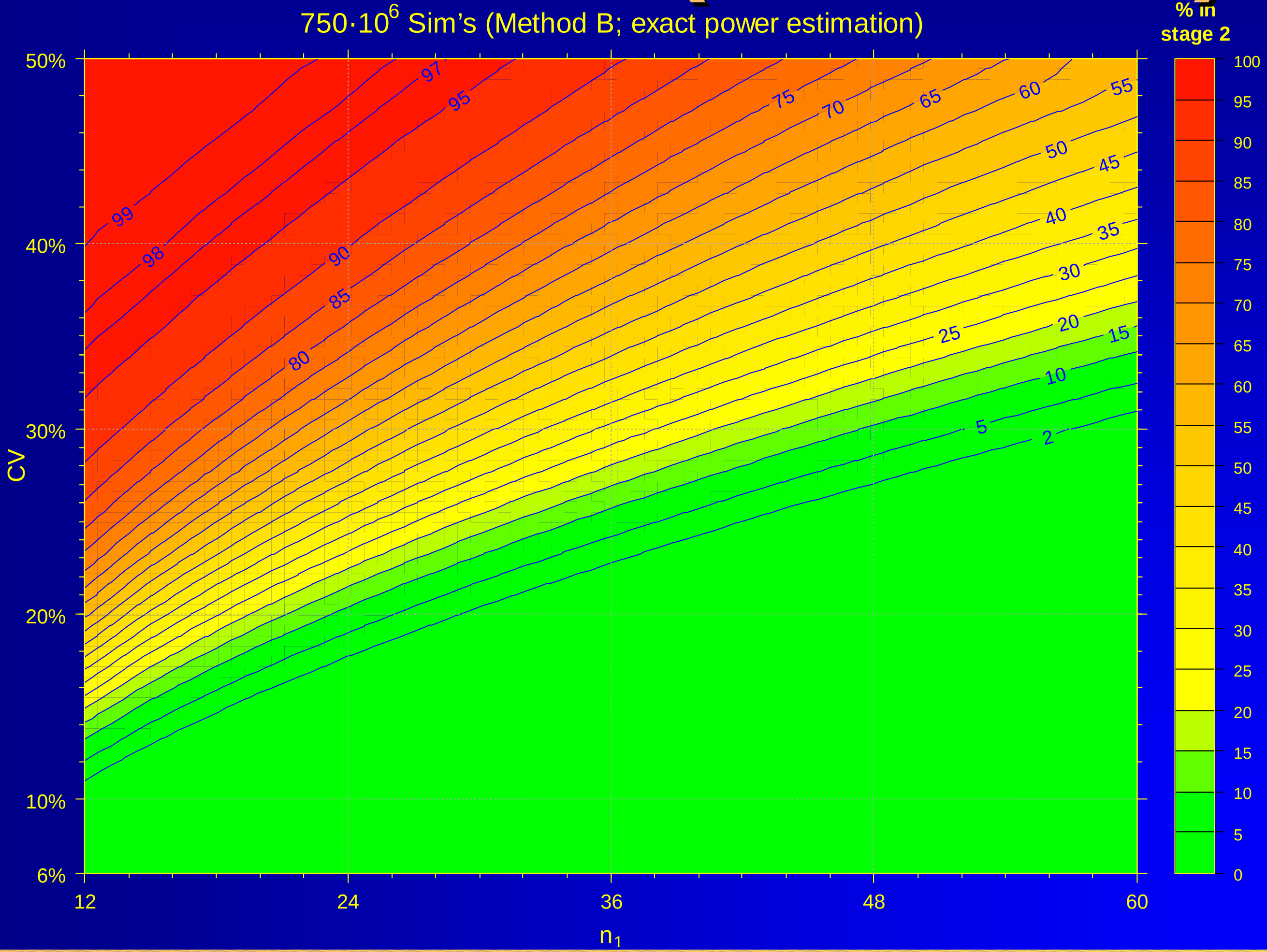


750·10^6 Sim's (Method B; exact power estimation)

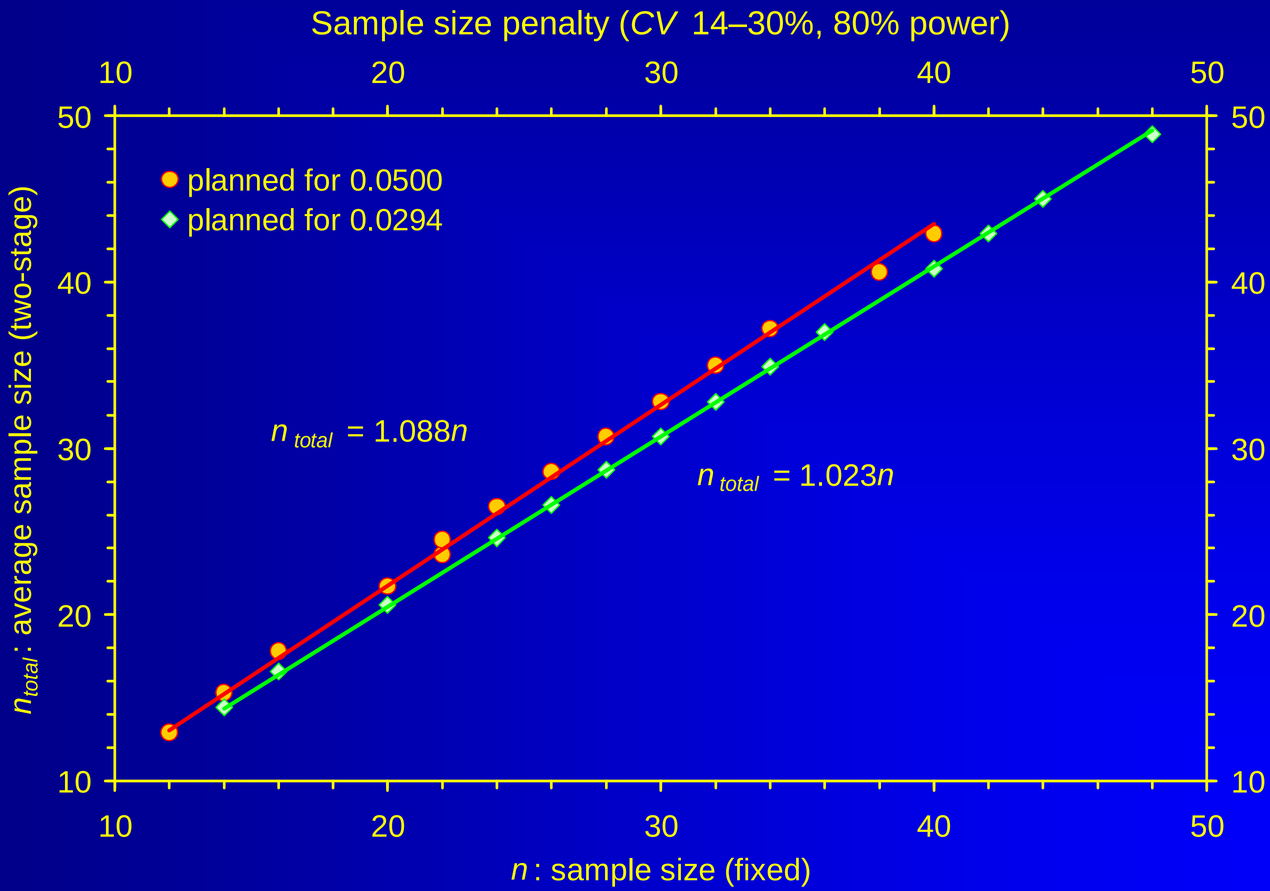
Potvin *et al.* (Method B)



Potvin *et al.* (Method B)



Potvin *et al.* (Method B)



Potvin *et al.* (Method B)

- Technical Aspects
  - Only *one* Interim Analysis (after stage 1).
  - Use software (wide step sizes in Diletti’s tables); preferrable the exact method (avoid approximations).
  - Should be termed ‘Interim Power Analysis’ *not* ‘Bioequivalence Assessment’ in the protocol.
  - No *a posteriori* Power – only a validated method in the decision tree.
  - No adjustment for T/R observed in stage 1 (not fully adaptive).

Potvin *et al.* (Method B)

- Technical Aspects (cont'd)
  - No 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).
  - Pocock's α 0.0294 is used in stage 1 and in the pooled analysis (data from stages 1 + 2), *i.e.*, the $1 - 2\times\alpha = 94.12\%$ CI is calculated.
  - Overall patient's risk preserved at ≤0.05.

Potvin *et al.* (Method B)

- 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. No need to test this effect.

Potvin *et al.* (Method B)

- Technical Aspects (cont’d)
  - Potvin *et al.* used a simple approximative power estimation based on the shifted *t*-distribution.
  - If possible use the exact method (Owen; *R* package *PowerTOST* `method = 'exact'`) or at least one based on the noncentral *t*-distribution (*PowerTOST* `method = 'noncentral'`).
  - Power obtained in stage 1 (example 2 from Potvin):

method	power
approx. (shifted *t*)	50.49%
approx. (noncentral *t*)	52.16%
exact	52.51%

Potvin *et al.* (Method B)

```
Model Specification and User Settings
       Dependent variable : Response
                Transform : LN
              Fixed terms : int+Sequence+Period+Treatment
   Random/repeated terms : Sequence*Subject

Final variance parameter estimates:
    Var(Sequence*Subject)     0.408682
            Var(Residual)     0.0326336
          Intrasubject CV     0.182132

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

                        Classical
    CI User = (   92.9330,  127.2838)
    Failed to show average bioequivalence for confidence=94.12 and percent=20.0.
```

12 subjects in stage 1, conventional BE model

CV_{intra} 18.2%

α 0.0294

Failed with 94.12% Confidence Interval

Potvin *et al.* (Method B)

```
require(PowerTOST)
power.TOST(alpha=0.0294, theta0=0.95,
           CV=0.182132, n=12, design='2x2',
           method='exact')

[1] 0.5251476
```

α 0.0294, T/R 95% – *not* 108.76% observed in stage 1!
CV$_{intra}$ 18.2%, 12 subjects in stage 1

Power 52.5% – 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', method='exact',
            print=TRUE)

+++++++++++ 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
```

Estimate total sample size:
α 0.0294, T/R 95%, *CV*$_{intra}$ 18.2%, 80% power

Total sample size 20: include another 8 in stage 2

Potvin *et al.* (Method B)

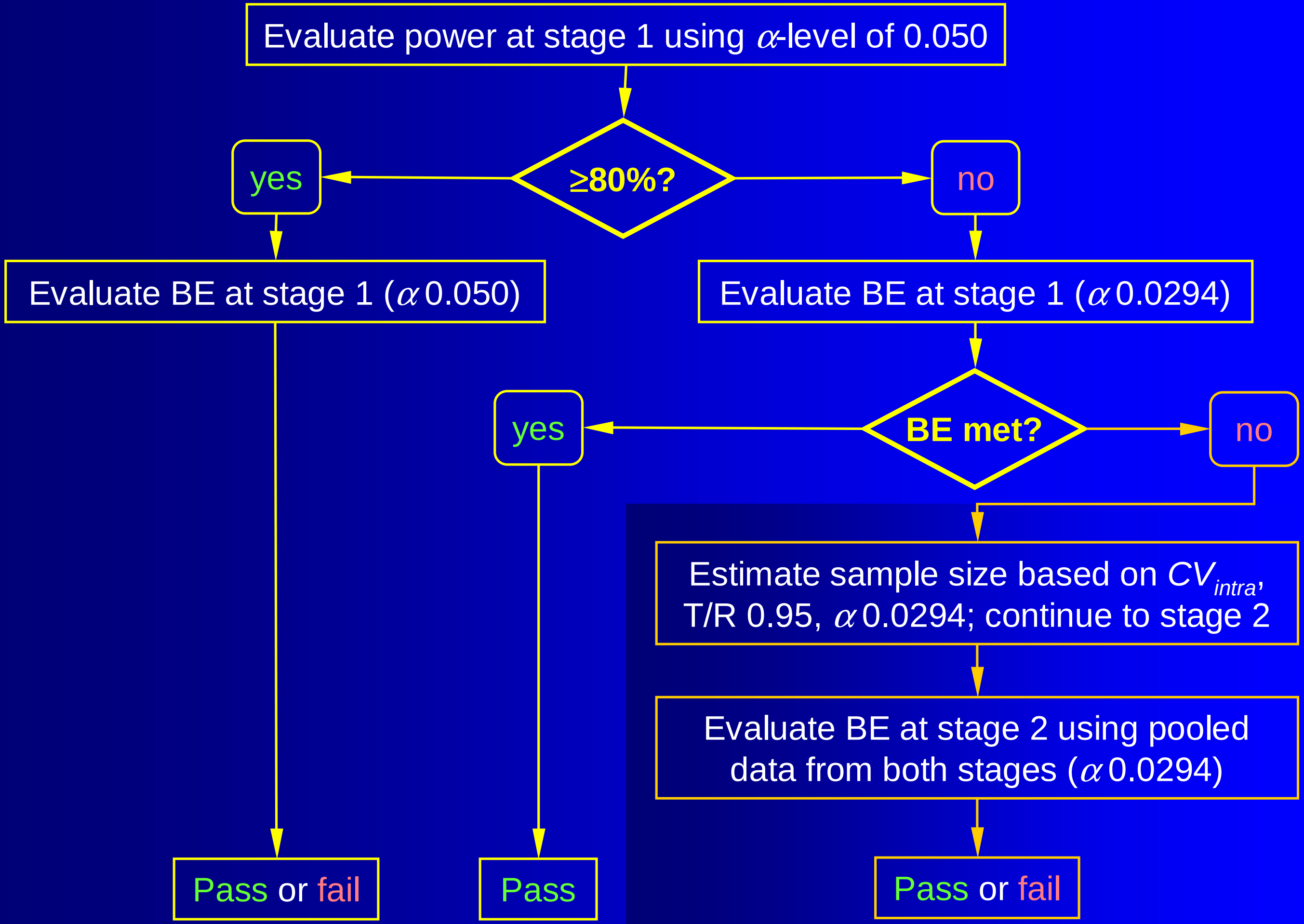
8 subjects in stage 2 (20 total), modified model in pooled analysis

```
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

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.* (Method C)



Evaluate power at stage 1 using α-level of 0.050

≥80%?

yes

Evaluate BE at stage 1 (α 0.050)

Pass or fail

no

Evaluate BE at stage 1 (α 0.0294)

BE met?

yes

Pass

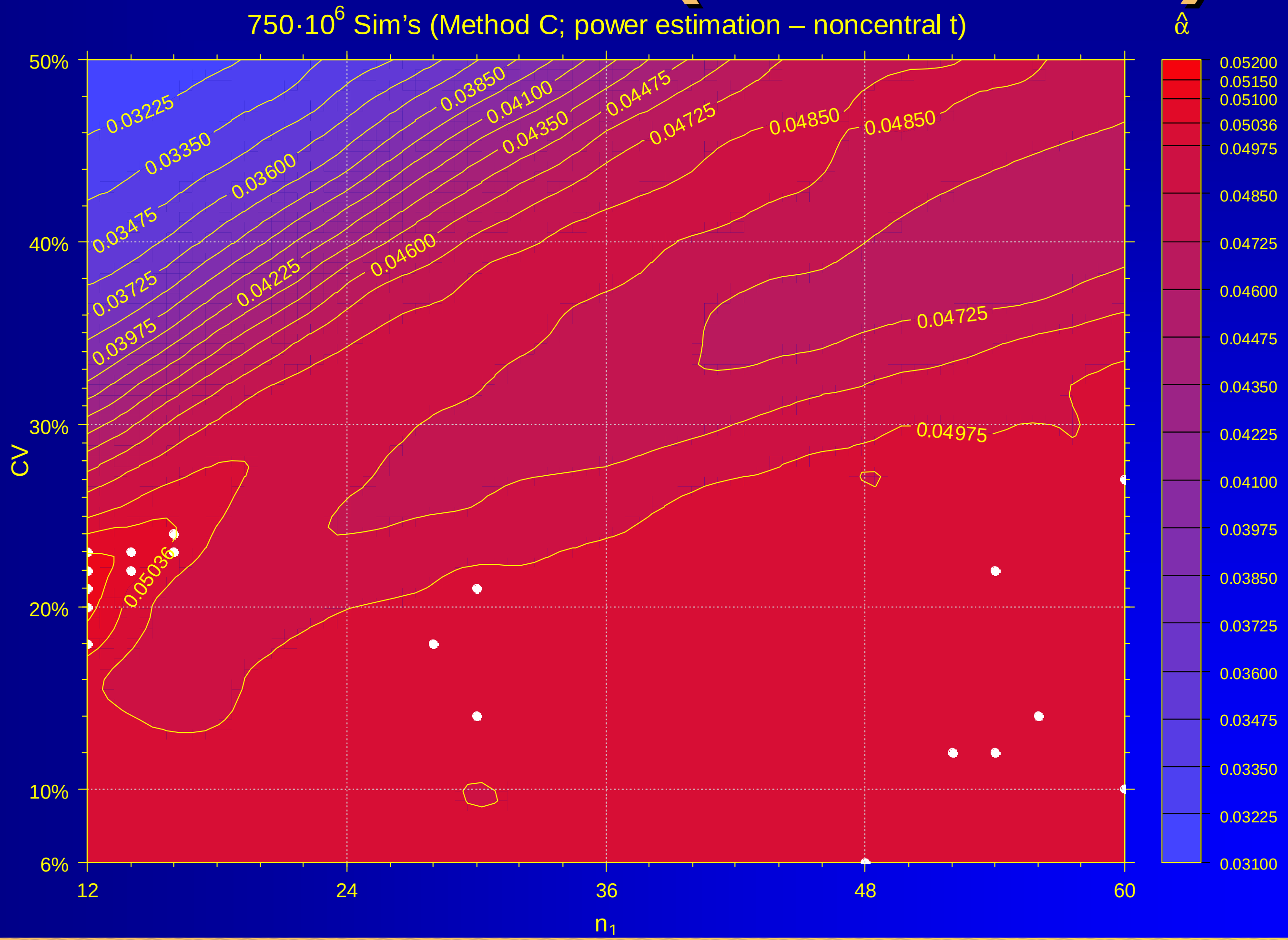
no

Estimate sample size based on CV_{intra}, T/R 0.95, α 0.0294; continue to stage 2

Evaluate BE at stage 2 using pooled data from both stages (α 0.0294)

Pass or fail

Potvin *et al.* (Method C)



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

- Pros & cons
  - Method C (*if power ≥80%!*) is a conventional BE study; no penality in terms of α needs to be applied.
  - Method C proceeds 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 ~10–20% higher.

Sequential Designs

- Methods by Potvin *et al.* (2008) limited to T/R of 0.95 and 80% power
  - Follow-up paper 2011
    - T/R 0.90 instead of 0.95.
    - Method D (like C, but α 0.0280 instead of α 0.0294).
    - Might be useful if T/R 0.95 and power 90% as well; *not validated yet!* Simulations required.

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

Sequential Designs

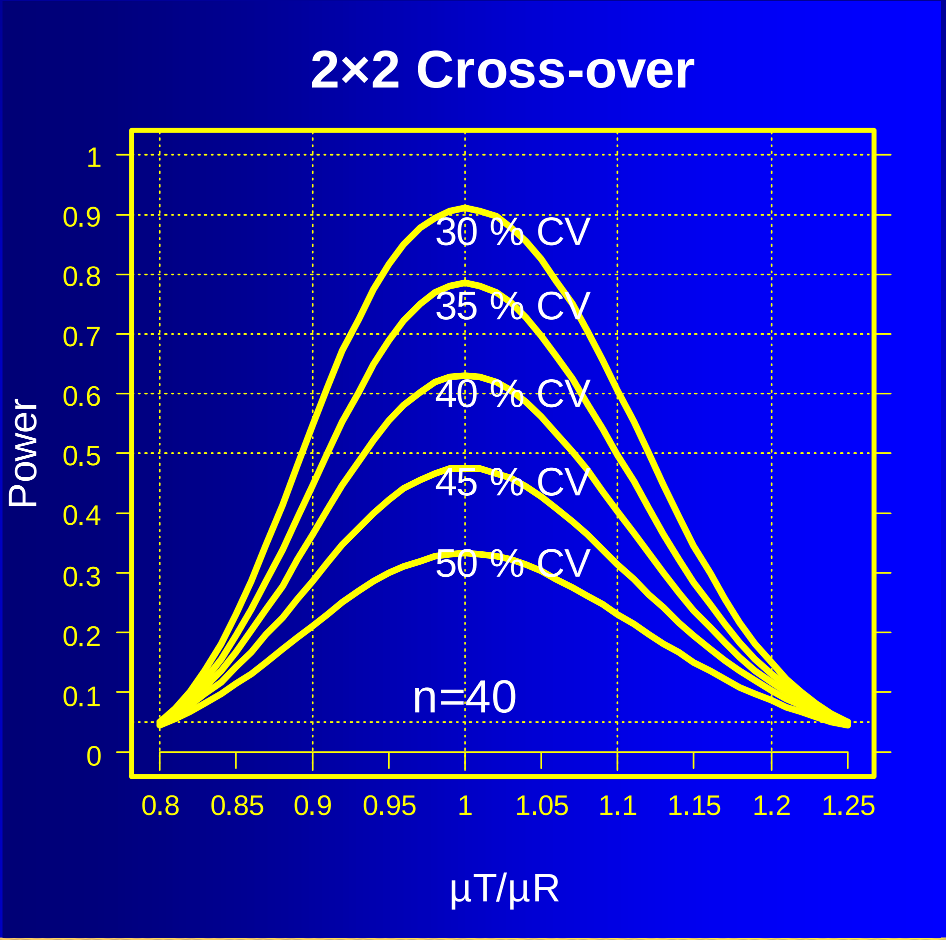
- Open issues
  - Feasibility / futility rules.
  - Arbitrary expected T/R and/or power.
  - Adaption for T/R observed in stage 1 (full adaptive design).
  - Dropping a candidate formulation from a higher-order cross-over.
  - Application to parallel designs (patients, long half-life drugs).
  - Application to replicated designs (for HVDs/HVDPs).

High variability

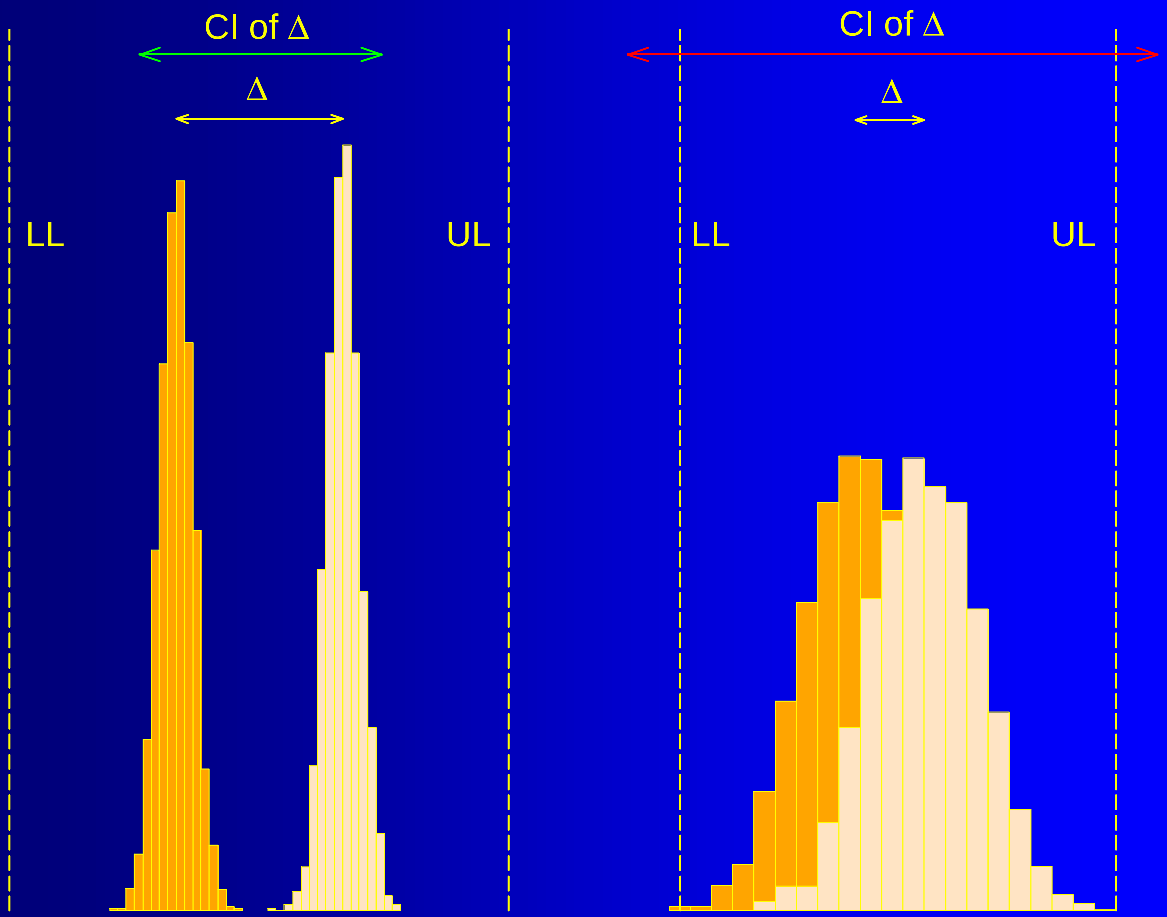
Power to show BE with 40 subjects for CV_{intra} 30–50%

μ_T/μ_R 0.95, CV_{intra} 30%
→ power 0.816
μ_T/μ_R 1.00, CV_{intra} 45%
→ power 0.476 <
Roulette 0.486 (!)

μ_T/μ_R 0.95, CV_{intra} 50%
→ n=98 (power 0.803)



High variability



Modified from Fig. 1
Tóthfalusi *et al.* (2009)

Counterintuitive concept of BE:

Two formulations with a large difference in means are declared bioequivalent if variances are low, but not bioequivalent – even if the difference is quite small – due to high variability.

Reminder

- The more ‘sophisticated’ a design is, the more information (in terms of σ^2) we obtain.
  - Hierarchy of designs:

Full replicate (TRTR | RTRT or TRT | RTR) ⇘
Partial replicate (TRR | RTR | RRT) ⇘
Standard 2×2 cross-over (RT | RT) ⇘
Parallel (R | T)

  - Assessable variances:

Parallel:	total variance (between + within)
2×2 Xover:	+ between, within subjects ⇗
Partial replicate:	+ within subjects (reference) ⇗
Full replicate:	+ within subjects (reference, test) ⇗

HVDPs (FDA)

- All (!) ANDAs submitted to FDA/OGD 2003 – 2005 (1010 studies, 180 drugs)
  - 31% (57/180) highly variable (*CV* ≥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 (FDA)

- 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

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%); different models depend on design (*e.g.*, SAS **`Proc MIXED`** for full replicate and SAS **`Proc GLM`** for partial replicate).
  - EMA: CV_{WR} ≤30%; all fixed effects model according to 2011’s Q&A-document preferred (*e.g.*, SAS **`Proc GLM`**).
  - Even if scaling is not intended, replicate design give more informations about formulation(s).

Application: HVDs/HVDPs

- Highly Variable Drugs / Drug Products (CV_{WR} >30 %)
  - ✓ USA Recommended in API specific guidances. Scaling for AUC and/or C_{max} acceptable, GMR 0.80 – 1.25; ≥24 subjects.
  - ± EU Widening of acceptance range (only C_{max}) to maximum of 69.84% – 143.19%), GMR 0.80 – 1.25. Demonstration that CV_{WR} >30% is not caused by outliers.
  - ± Russia?

Replicate designs

- Designs
  - Two-sequence three-period

    T R T

    R T R

    Sample size to obtain the same power as a 2×2×2 study: ~75%
  - Two-sequence four-period

    T R T R

    R T R T

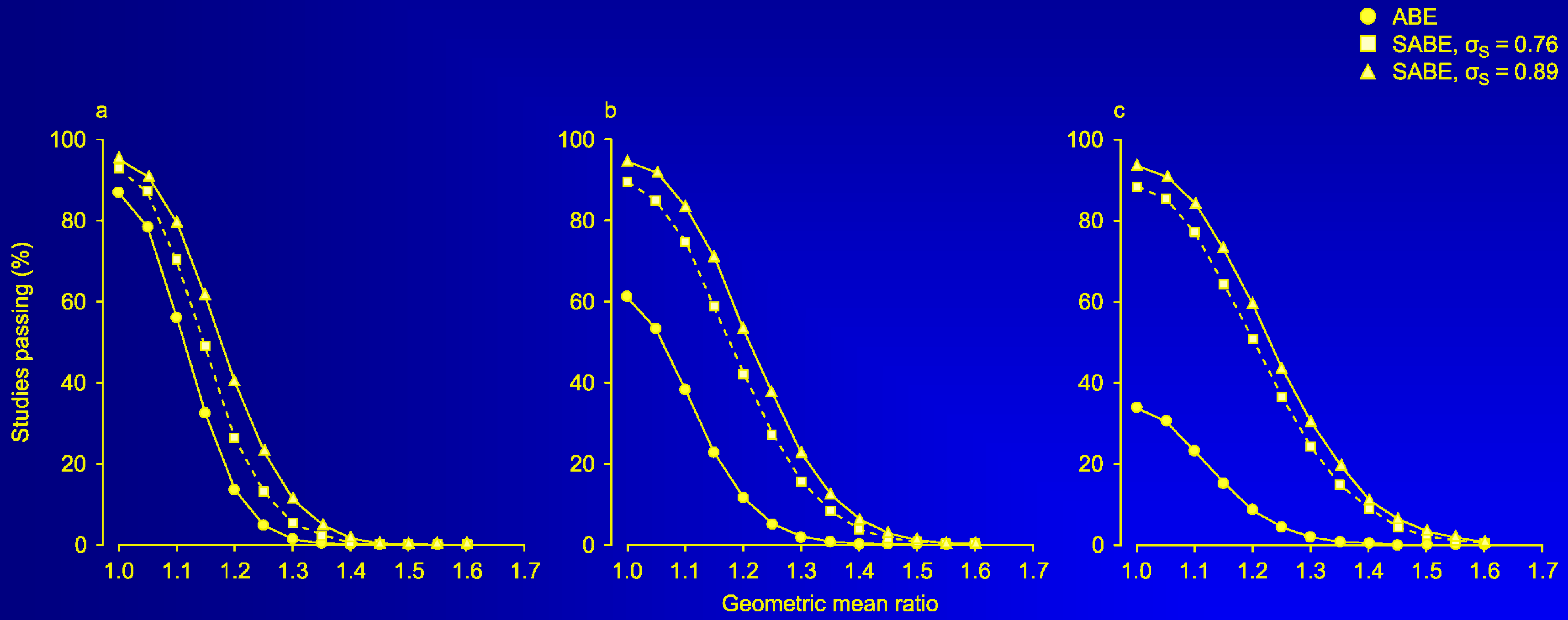
    Sample size to obtain the same power as a 2×2×2 study: ~50%
  - and many others… (FDA: TRR | RTR | RRT, aka ‘partial replicate’)
  - The statistical model is quite complicated – and dependent on the actual design!

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

HVDs/HVDPs

- Replicate designs
  - 4-period replicate designs:
    sample size = ~½ of 2×2 study's sample size.
  - 3-period replicate designs:
    sample size = ~¾ of 2×2 study's sample size.
  - Number of treatments (and biosamples) ~conventional 2×2 cross-over.
  - Allow for a safety margin – expect a higher number of drop-outs due to additional period(s).
  - Consider increased blood loss (ethics!); eventually improved bioanalytics required.

HVDPs (EMA/Russia *vs.* FDA)



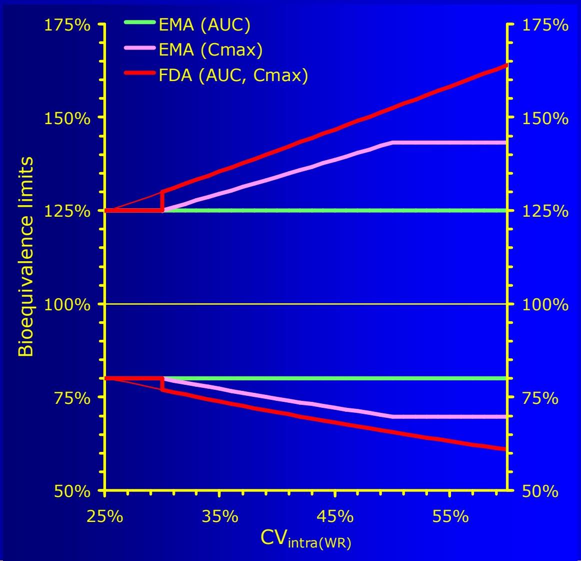
Tóthfalusi *et al.* (2009), Fig. 3

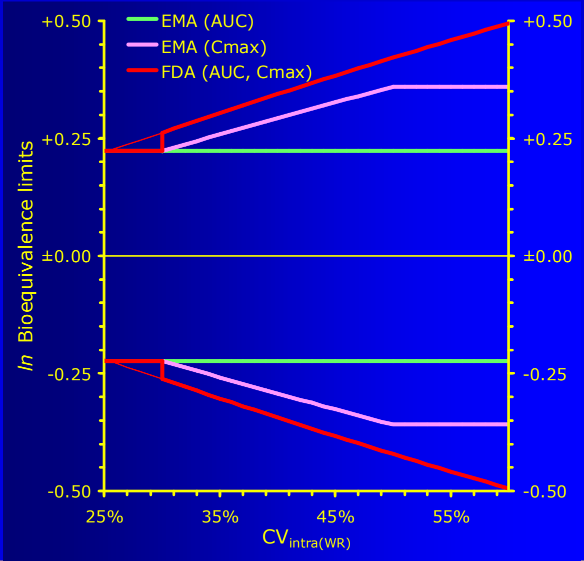
Simulated (n = 10 000) three-period full replicate design studies (TRT | RTR) in 36 subjects; GMR restriction 0.80–1.25. (a) CV = 35%, (b) CV = 45%, (c) CV = 55%.

ABE: Conventional Average Bioequivalence, SABE: Scaled Average Bioequivalence, 0.76: EMA/Russia criterion, 0.89: FDA criterion.

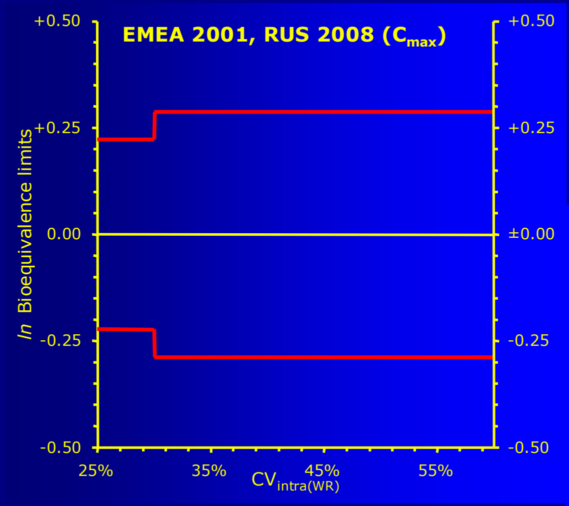
HVDPs (EMA/Russia *vs.* FDA)

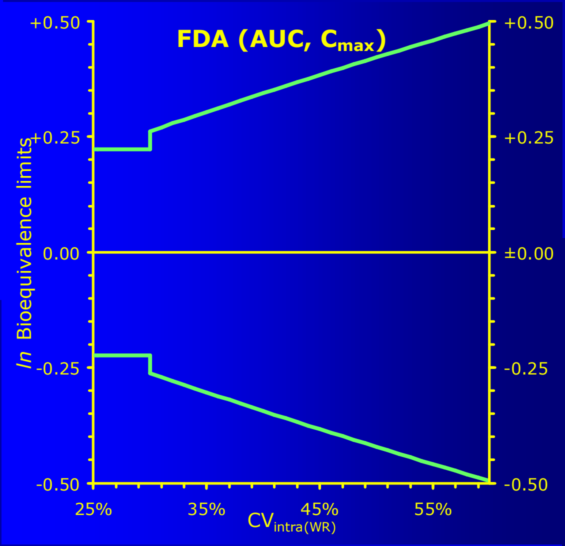
- EMA's/Russia's and FDA's approaches differ; FDA's leads to a discontinuity of the acceptance range at *CV* 30%, because FDA's scaling *CV* is 25.83% (σ_{WR} 0.294) – but *applied* at *CV* ≥30%.

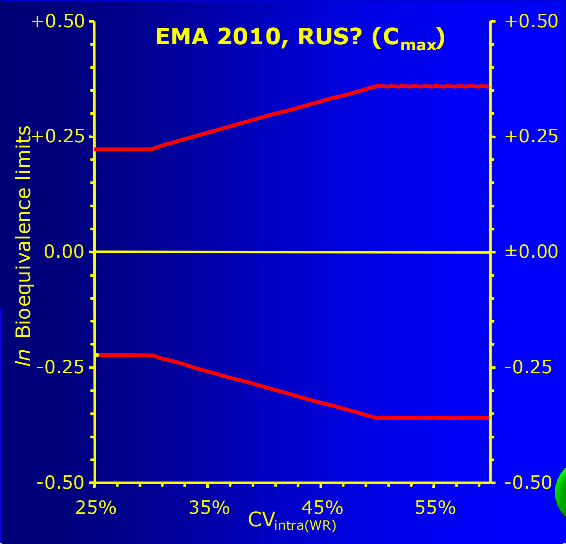




HVDPs (No Global Harmonization!)


EMEA 2001, RUS 2008 (C_{max})


FDA (AUC, C_{max})


EMA 2010, RUS? (C_{max})

HVDs/HVDPs (Reg. models)

- Common to EMA and FDA

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 EMA and FDA

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 EMA and FDA

  EMA/Russia: Regulatory regulatory switching condition θ_S avoids the discontinuity.

$$CV_W = 0.30$$

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

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

HVDs/HVDPs (FDA)

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

  Starting from the SABE model

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

  Rearrangement leads to a linear form

  $$\left(\mu_T - \mu_R\right)^2 - \theta_S^2 \cdot \sigma_W^2 \le 0$$

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

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

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

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

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

$$C_s = \frac{\theta_S^2 \cdot (N - S) \cdot s_W^2}{\chi^2_{\alpha,N-S}}$$

t and χ^2 are the inverse cumulative distribution functions at α 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

$$L_m = \left(C_m - E_m\right)^2$$

$$L_s = \left(C_s - E_s\right)^2$$

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

The CL of the rearranged SABE criterion is evaluated at the 95% level. If the upper 95% is positive, RSABE is rejected, and accepted otherwise.

HVDs/HVDPs (EMA, Russia)

- EU GL on BE (2010), Russia ?
  - Average Bioequivalence (ABE) with Expanding Limits (ABEL)
    - The regulatory switching condition θ_S at CV_{WR} 30% would be 0.7601228297680…
    - According to the GLs and the EMA's 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), Russia?
  - Average Bioequivalence (ABE) with Expanding Limits (ABEL)
    - Based on σ_{WR} (the *intra*-subject standard deviation of the reference formulation) calculate the scaled acceptance range based on the regulatory constant k (θ_s=0.760); limited at CV_{WR} 50%.

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

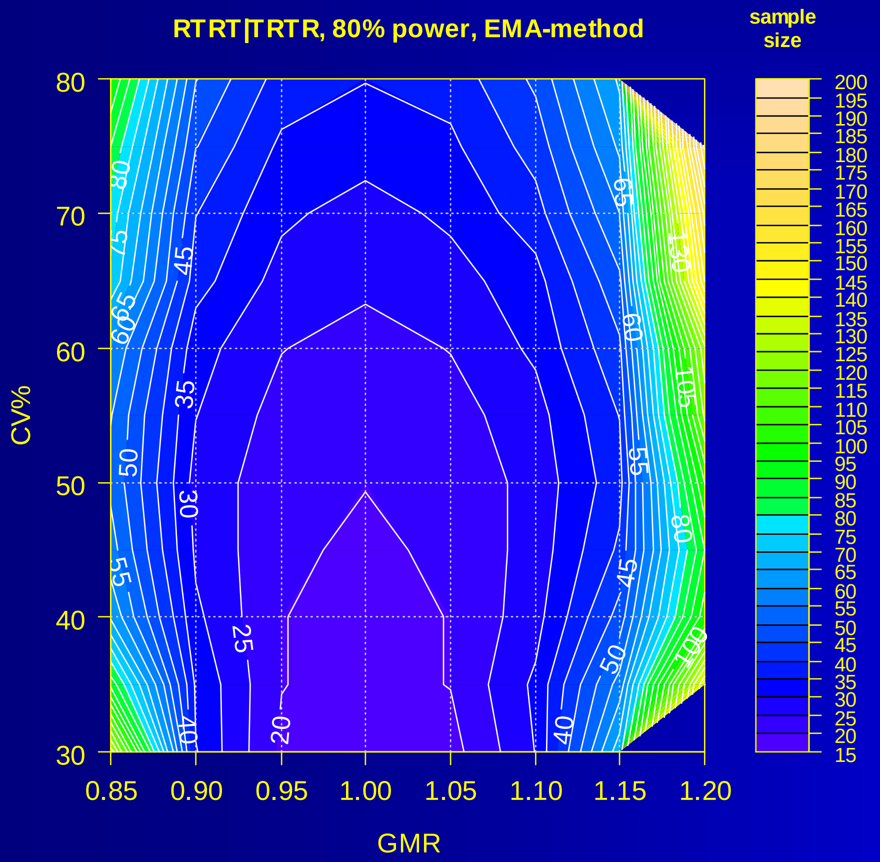
CV_{WR}	$L - U$
≤30	80.00 – 125.00
35	77.23 – 129.48
40	74.62 – 143.02
45	72.15 – 138.59
≥50	69.84 – 143.19

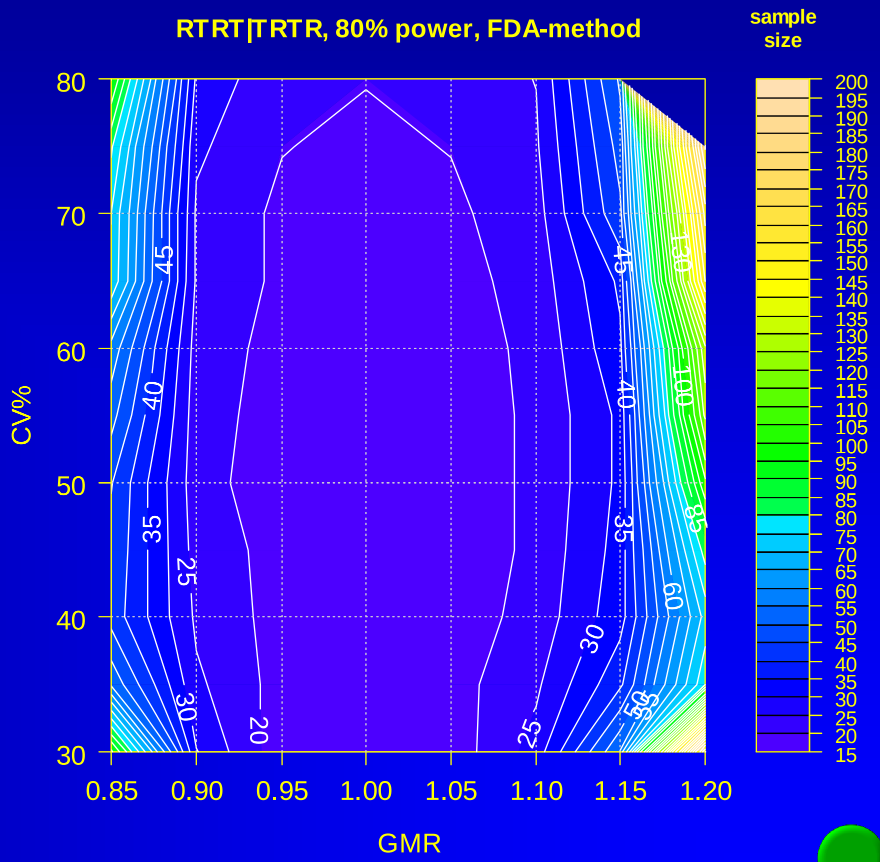
HVDs/HVDPs (EMA)

- At higher CVs the GMR is of increasing importance!
- $CV_{WR} > 50\%$ still requires large sample sizes.
- No software for sample size estimation (based on α, β, GMR, and CV) can deal with the GMR restriction.
- Recently sample size tables based on simulations were published (for EMA's and FDA's methods, full and partial replicate designs, CV_{WR} 30–80%, power 80 and 90%).

L Tóthfalusi and L Endrényi
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 (EMA/FDA; sample sizes)





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

Example datasets (EMA)

- Q&A document (March 2011)
  - Data set I
    RTRT | TRTR full replicate, 77 subjects, imbalanced, incomplete
    - FDA
      s_{WR} 0.446 ≥ 0.294 → apply RSABE (CV_{WR} 46.96%)
      a. critbound -0.0921 ≤ 0 and
      b. 80.00% ≤ pointest 115.46% ≤ 125.00% ✓
    - EMA
      - CV_{WR} 46.96% → apply RSABE (> 30%)
      - Scaled Acceptance Range: 71.23% – 140.40%
      - A: 71.23% ≤ 107.11% – 124.89% ≤ 140.40%, PE 115.66% ✓
      - B: 71.23% ≤ 107.17% – 124.97% ≤ 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 → apply ABE (CV_{WR} 11.43%)
      80.00% ≤ 97.05 – 107.76 ≤ 125.00% (CV_{intra} 11.55%) ✓
    - EMA
      - CV_{WR} 11.17% → apply ABE (≤ 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)

- EMA 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%

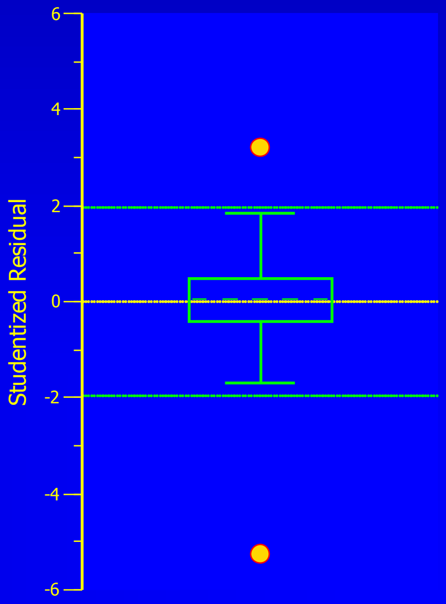
    Method B: 107.17% – 124.97%
  - But there *are* two outliers!

    Excluding subjects 45 and 52

    CV_{WR} drops to 32.16%.

    ABEL 78.79% – 126.93%

    Almost no more gain compared to conventional limits.



Спасибо!

Оценка числа добровольцев для исследований БЭ

Вопросы?

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

Dedicated to the memory of Dirk Maarten Barends (1945 – 2012).

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

In bioequivalence we must not forget the only important – *the patient*! He/she is living person, not just α 0.05.

Dirk Marteen Barends

References

- ICH
  - E9: Statistical Principles for Clinical Trials (1998)
- EMA-CPMP/CHMP/EWP
  - Points to Consider on Multiplicity Issues in Clinical Trials (2002)
  - Guideline on the Investigation of BE (2010)
  - Questions & Answers: Positions on specific questions addressed to the EWP therapeutic subgroup on Pharmacokinetics (2011, 2012)
- US-FDA
  - Center for Drug Evaluation and Research (CDER)
    - Statistical Approaches Establishing Bioequivalence (2001)
    - Bioequivalence Recommendations for Specific Products (2007–2012):
      Guidance on Progesterone (Feb 2011)
      Guidance on Loteprednol (Jun 2012)
      Guidance on Dexamethasone/Tobramycin (Jun 2012)
- Midha KK *et al.*
  Logarithmic Transformation in Bioequivalence: Application with Two Formulations of Perphenazine
  J Pharm Sci 82/2, 138–44 (1993)
- 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, 376–8 (1994)
- Diletti E, Hauschke D, and VW Steinijans
  Sample size determination for bioequivalence assessment by means of confidence intervals
  Int J Clin Pharm Ther Toxicol 29/1, 1–8 (1991)
- Hauschke D *et al.*
  Sample Size Determination for Bioequivalence Assessment Using a Multiplicative Model
  J Pharmacokin Biopharm 20/5, 557–61 (1992)
- Chow S-C and H Wang
  On Sample Size Calculation in Bioequivalence Trials
  J Pharmacokin Pharmacodyn 28/2, 155–69 (2001)
  Errata: J Pharmacokin Pharmacodyn 29/2, 101–2 (2002)
- DB Owen
  A special case of a bivariate non-central t-distribution
  Biometrika 52, 3/4, 437–46 (1965)
- LA Gould
  Group Sequential Extension of a Standard Bioequivalence Testing Procedure
  J Pharmacokin Biopharm 23/1, 57–86 (1995)
  DOI: 10.1007/BF02353786

References

- RV Lenth
  Two Sample-Size Practices that I don't recommend
  Joint Statistical Meetings, Indianapolis (2000)
- Hoenig JM and DM Heisey
  The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis
  The American Statistician 55/1, 19–24 (2001)
- P Bacchetti
  Current sample size conventions: Flaws, harms, and alternatives
  BMC Medicine 8:17 (2010)
- Jones B and MG Kenward
  Design and Analysis of Cross-Over Trials
  Chapman & Hall/CRC, Boca Raton (2nd Edition 2000)
- Patterson S and B Jones
  Determining Sample Size, in:
  Bioequivalence and Statistics in Clinical Pharmacology
  Chapman & Hall/CRC, Boca Raton (2006)
- SA Julious
  Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data
  Statistics in Medicine 23/12, 1921–86 (2004)
- SA Julious
  Sample Sizes for Clinical Trials
  Chapman & Hall/CRC, Boca Raton (2010)
- Julious SA and RJ Owen
  Sample size calculations for clinical studies allowing for uncertainty about the variance
  Pharmaceutical Statistics 5/1, 29–37 (2006)
- D Labes
  Package 'PowerTOST', Version 0.9-11 (2012-08-07)
  http://cran.r-project.org/web/packages/PowerTOST/PowerTOST.pdf
- Potvin D *et al.*
  Sequential design approaches for bioequivalence studies with crossover designs
  Pharmaceut Statist 7/4, 245–62 (2008)
- Montague TH *et al.*
  Additional results for 'Sequential design approaches for bioequivalence studies with crossover designs'
  Pharmaceut Statist 11/1, 8–13 (2011)
- Tothfálusi L, Endrényi L, and A García Arieta
  Evaluation of Bioequivalence for Highly Variable Drugs with Scaled Average Bioequivalence
  Clin Pharmacokinet 48/11, 725–43 (2009)
- Tothfálusi L and L Endrényi
  Sample Sizes for Designing Bioequivalence Studies for Highly Variable Drugs
  J Pharm Pharmaceut Sci 15(1), 73–84 (2011)

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` ≥`0.294`

RSABE if

`a.` `critbound` ≤ 0 *and*

`b.` `0.8000` ≤`pointest` ≤`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 cl 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 cl 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