Successfully Overcoming Sample Size Challenges in BE Studies

Helmut Schütz
BEBAC

Sample Size (Limits)

- Minimum
  - 12: WHO, EU, CAN, NZ, AUS, AR, MZ, ASEAN States, RSA
  - 12: USA ‘A pilot study that documents BE can be appropriate, provided its design and execution are suitable and a sufficient number of subjects (*e.g.*, 12) have completed the study.’
  - 20: RSA (MR formulations)
  - 24: Saudia Arabia (12 to 24 if statistically justifiable)
  - 24: Brazil
  - Sufficient number: JPN

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.’

Sample Size (Limits?)

- Reminder
  - Generally power is set to at least 80 % (β, error type II: producers's risk to get no approval for a bioequivalent formulation; power = 1 – β).

    1 out of 5 studies will fail just by chance!
  - If you plan for power of less than 70 %, problems with the ethics committee are likely (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 – especially for studies with more than two periods or multiple-dose studies.

EU

- EMEA NfG on BA/BE (2001)
  - Detailed information (data sources, significance level, expected deviation, desired power).
- EMA GL on BE (2010)
  - Batches must not differ more than 5%.
  - The number of subjects to be included in the study should be based on an appropriate sample size calculation.

Cookbook?

Coefficient(s) of Variation

- The more ‘sophisticated’ the design is, the more information one may extract.
  - Hierarchy of designs:
    Full replicate (TRTR | RTRT)
    Partial replicate (TRR | RTR | RRT)
    Standard 2×2 cross-over (RT | RT)
    Parallel (R | T)
  - Variances which can be resolved:
    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!)
  - Total CV% from a 2×2 cross-over used in planning a parallel design study:
    - Intra-subject CV% (within) → $CV_{\text{intra}}\,\% = 100 \cdot \sqrt{e^{MSE_W} - 1}$
    - Inter-subject CV% (between) → $CV_{\text{inter}}\,\% = 100 \cdot \sqrt{e^{\frac{MSE_B - MSE_W}{2}} - 1}$
    - Total CV% (pooled) → $CV_{\text{total}}\,\% = 100 \cdot \sqrt{e^{\frac{MSE_B + MSE_W}{2}} - 1}$

Coefficient(s) of Variation

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

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

    - … pilot study unavoidable

Hints

- 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} not given (quite often), a little algebra helps. All you need is the 90% geometric confidence interval and the sample size.

Algebra…

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

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

  - Estimate the number of subjects / sequence (example 2×2 cross-over)
    - If total sample size (N) is an even number, *assume* (!)

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

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

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

  - 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)

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

  - 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)

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

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

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

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

Algebra…

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

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

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

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

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

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

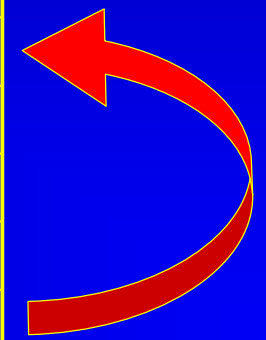
Sensitivity to Imbalance

- 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

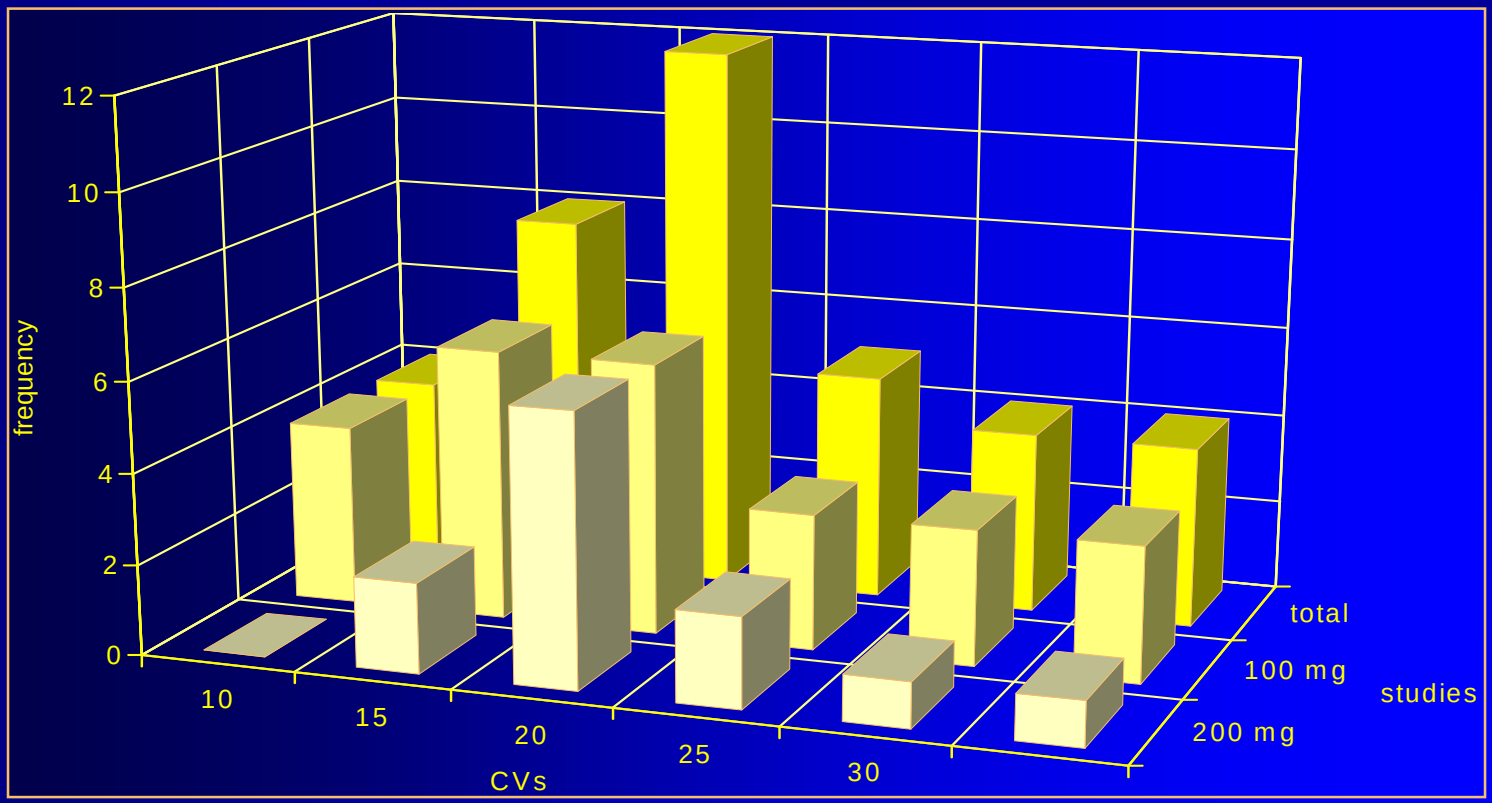
Balanced Sequences assumed…

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

Sequences in study



Literature data



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

Pooling of CV%

- Intra-subject CV from different studies can be pooled
  - Do not use the arithmetic mean (or the geometric mean either) of CVs.
  - In the parametric model of log-transformed data, additivity of variances (not of CVs!) apply.
  - Before pooling variances must be weighted acccording to the studie's sample size.

Pooling of CV%

- Intra-subject CV from different studies
  - Calculate the variance from CV

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

  - Calculate the total variance weighted by df

$$\sum \sigma_w^2 df$$

  - Calculate the pooled CV from total variance

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

  - Optionally calculate an upper (1-α) % confidence limit on the pooled CV (recommended α=0.25)

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

Pooling of CV%

- Example 1: $n_1=n_2$;
  $CV_{Study1} < CV_{Study2}$

studies	N
2	24

df (total)	α	1-α	total	CV_{pooled}	CV_{mean}
20	0.2	0.8	1.2540	**0.254**	~~0.245~~
		$\chi^2_{(1-\alpha,df)}$	14.578	0.300	+17.8%

CV_{intra}	n	seq.	df (mj)	σw	σ^2w	$\sigma^2w \times df$	$CV_{intra\ /\ pooled}$	$>CL_{upper}$
0.200	**12**	**2**	10	0.198	0.0392	0.3922	78.6%	no
0.300	**12**	**2**	10	0.294	0.0862	0.8618	117.9%	yes

Pooling of CV%

- Example 2: $n_1 < n_2$;
  $CV_{Study1} < CV_{Study2}$

studies	N
2	36

df (total)	α	1-α	total	CV_{pooled}	CV_{mean}
32	0.2	0.8	2.2881	**0.272**	~~0.245~~
		$\chi^2_{(1-\alpha,df)}$	25.148	0.309	+13.4%

CV_{intra}	n	seq.	df (mj)	σw	$\sigma^2 w$	$\sigma^2 w \times df$	$CV_{intra\ /\ pooled}$	$>CL_{upper}$
0.200	**12**	**2**	10	0.198	0.0392	0.3922	73.5%	no
0.300	**24**	**2**	22	0.294	0.0862	1.8959	110.2%	no

Pooling of CV%

- Example 3: $n_1 > n_2$;
  $CV_{Study1} < CV_{Study2}$

studies	N
2	36

df (total)	α	1-α	total	CV_{pooled}	CV_{mean}
32	0.2	0.8	1.7246	**0.235**	~~0.245~~
		$\chi^2_{(1-\alpha,df)}$	25.148	0.266	+13.2%

CV_{intra}	n	seq.	df (mj)	σ_w	σ^2_w	$\sigma^2_w \times df$	$CV_{intra\ /\ pooled}$	$>CL_{upper}$
0.200	**24**	**2**	22	0.198	0.0392	0.8629	85.0%	no
0.300	**12**	**2**	10	0.294	0.0862	0.8618	127.5%	yes

α- *vs.* β-Error

- α-Error (aka error type I): patient's risk to be treated with a bioinequivalent formulation
  - Reminder: BA in a particular patient can be *either* below 80% *or* above 125%.
  - If we keep the risk of particular patients at 0.05 (5%), the risk of entire the population of patients (<80% *and* >125%) is 2×α (10%)
    That's where the 90% confidence interval comes from (CI = 1 – 2×α = 0.90)
  - Although α is generally set to 0.05, sometimes <0.05 (*e.g.*, NTDIs in Brazil, multiplicity, interim analyses).

α- *vs.* β-Error

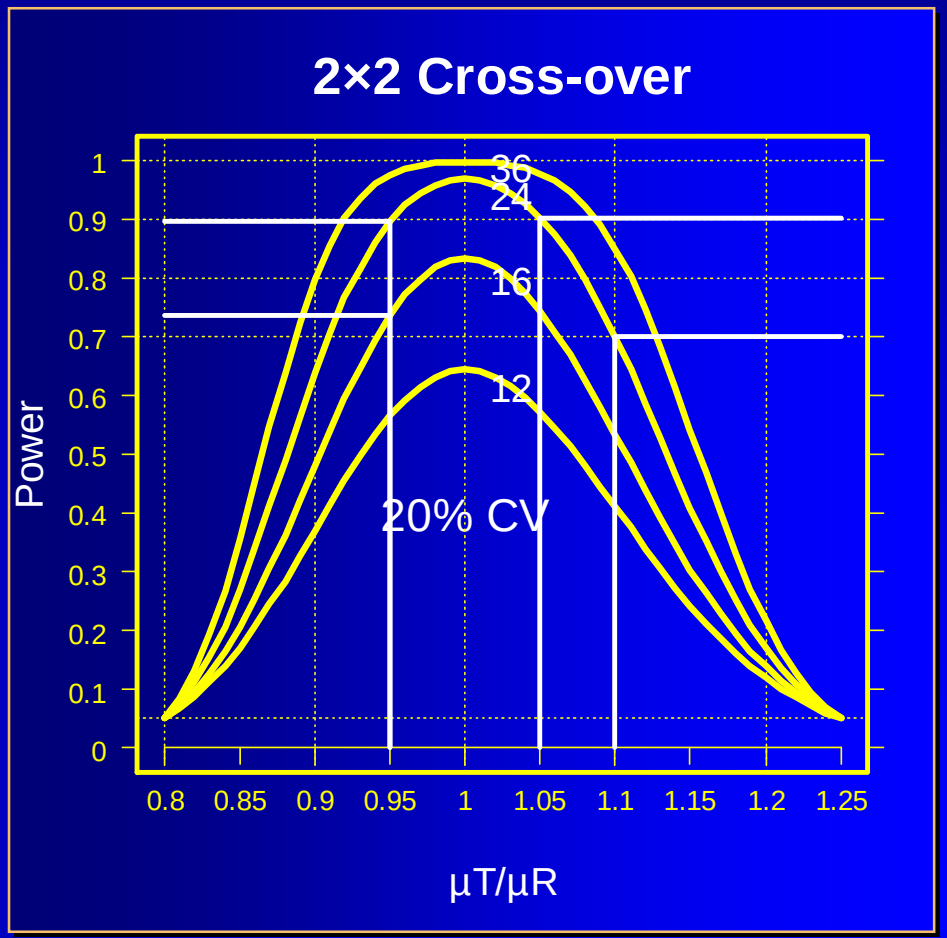
- β-Error (aka error type II): producer's risk to get no approval for a bioequivalent formulation
  - Generally *set* in study planning to ≤0.2, where power = 1 - β = ≥80%
  - No guidelines about power ('appropriate'), but
    - 70% only in exceptional cases
    - >90% *may* raise questions from the Ethics Committee (suspection of 'forced bioequivalence')
  - There is no *a posteriori* (aka *post hoc*) power!
    Either a study has shown BE or not.
    Phoenix'/WinNonlin's output is statistical nonsense!

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

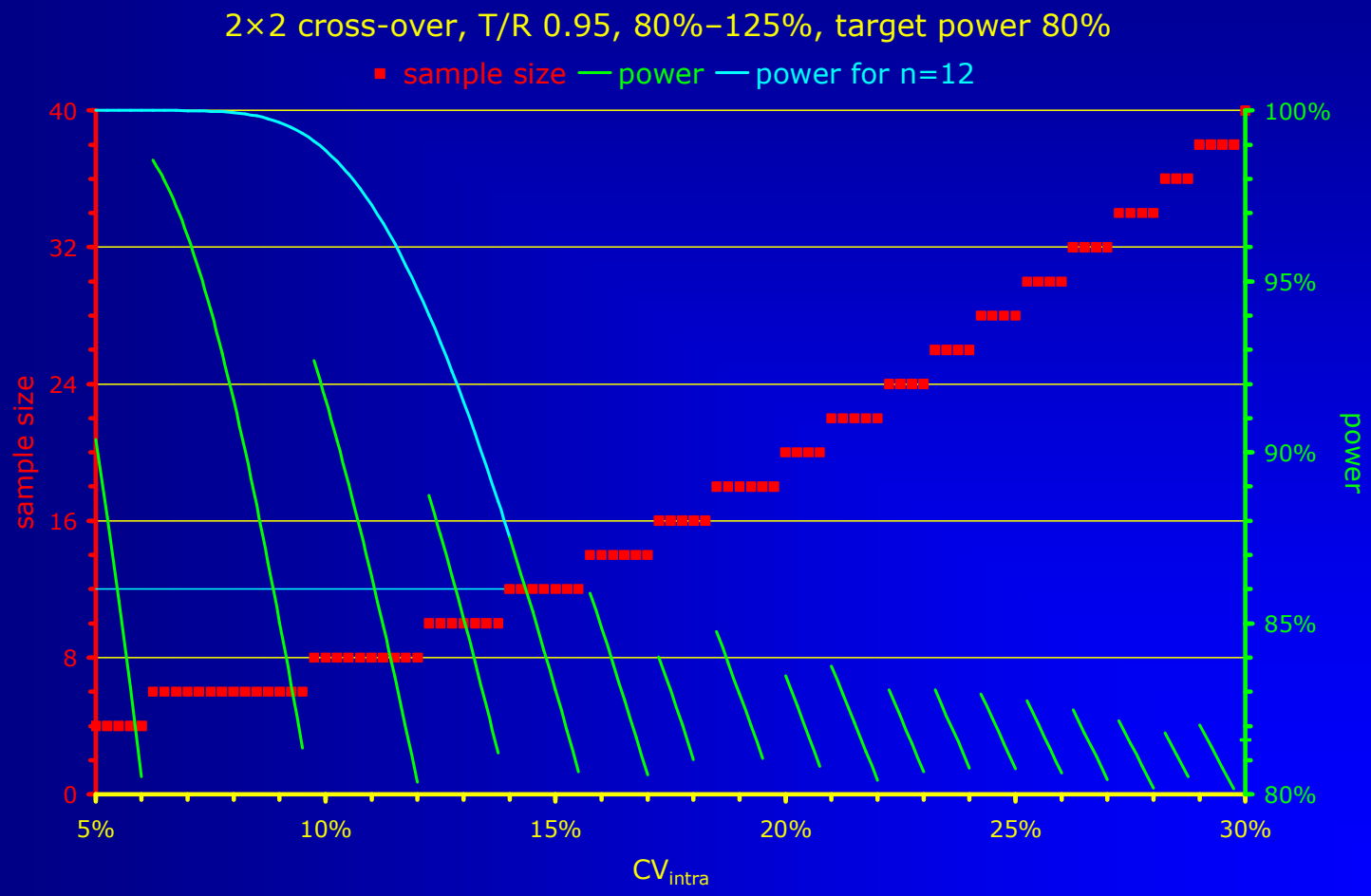


Power *vs.* Sample Size

- 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% (β=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 2×2 study (power 83%).

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

Power *vs.* Sample Size



Tools

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

Background

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

Background

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

Comparison

original values	Method	Algorithm	CV% 5.	7.5	10.	12.	12.5	14.	15.	16.	17.5	18.	20.	22.
PowerTOST 0.5 (2010)	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*	?	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

original values	Method	Algorithm	CV% 22.5	24.	25.	26.	27.5	28.	30.	32.	34.	36.	38.	40.
PowerTOST 0.5 (2010)	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*	?	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

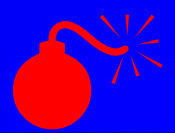
1. $\Delta = \ln(0.8)-\ln(T/R) = -0.1719$
2. Start with e.g. n=8/sequence
   1. $df = n \cdot 2 - 1 = 8 \times 2 - 1 = 14$
   2. $t_{\alpha,df} = 1.7613$
   3. $t_{\beta,df} = 0.8681$
   4. new $n = [(t_{\alpha,df} + t_{\beta,df})^2 \cdot (CV/\Delta)]^2 =$ $(1.7613+0.8681)^2 \times (-0.2/0.1719)^2 = 9.3580$
3. Continue with n=9.3580/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

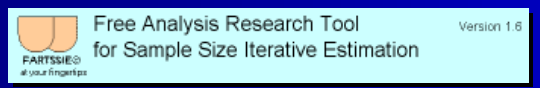
1. $\Delta = \ln(T/R) - \ln(1.25) = 0.1719$
2. Start with e.g. n=8/sequence
   1. $df_\alpha = \text{roundup}(2 \cdot n-2) \cdot 2-2 = (2 \times 8-2) \times 2-2 = 26$
   2. $df_\beta = \text{roundup}(4 \cdot n-2) = 4 \times 8-2 = 30$
   3. $t_{\alpha,df} = 1.7056$
   4. $t_{\beta/2,df} = 0.8538$
   5. new $n = \beta^2 \cdot [(t_{\alpha,df} + t_{\beta/2,df})^2/\Delta^2 =$ $0.2^2 \times (1.7056+0.8538)^2 / 0.1719^2 = 8.8723$
3. Continue with n=8.8723/sequence (N=17.7446 → 18)
   1. $df_\alpha = \text{roundup}(2 \cdot n-2) \cdot 2-2 = (2 \times 8.8723-2) \times 2-2 = 30$
   2. $df_\beta = \text{roundup}(4 \cdot n-2) = 4 \times 8.8723-2 = 34$
   3. $t_{\alpha,df} = 1.6973$
   4. $t_{\beta/2,df} = 0.8523$
   5. new $n = \beta^2 \cdot [(t_{\alpha,df} + t_{\beta/2,df})^2/\Delta^2 =$ $0.2^2 \times (1.6973+0.8538)^2 / 0.1719^2 = 8.8045$
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 the plausibility of software's results
- Approximations based on noncentral *t* (FARTSSIE17)



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, logscale = TRUE,
  theta1 = 0.80, diff = 0.95, CV = 0.30,
  design = "2x2", exact = TRUE)
```

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

HVDs/HVDPs

- EU GL on BE (2010)
  - The regulatory switching condition θ_s is derived from the regulatory standardized variation σ_0. For CV_{WR} = 30% one gets

$$\sigma_0 = \sqrt{\ln(0.3^2 + 1)} = 0.2936$$

and

$$\theta_s = \frac{\ln(1.25)}{\sigma_0} = -\frac{\ln(0.80)}{\sigma_0} = 0.7601$$

Tothfalusi *et al.* (2009)

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 (θ_s=0.7601):

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

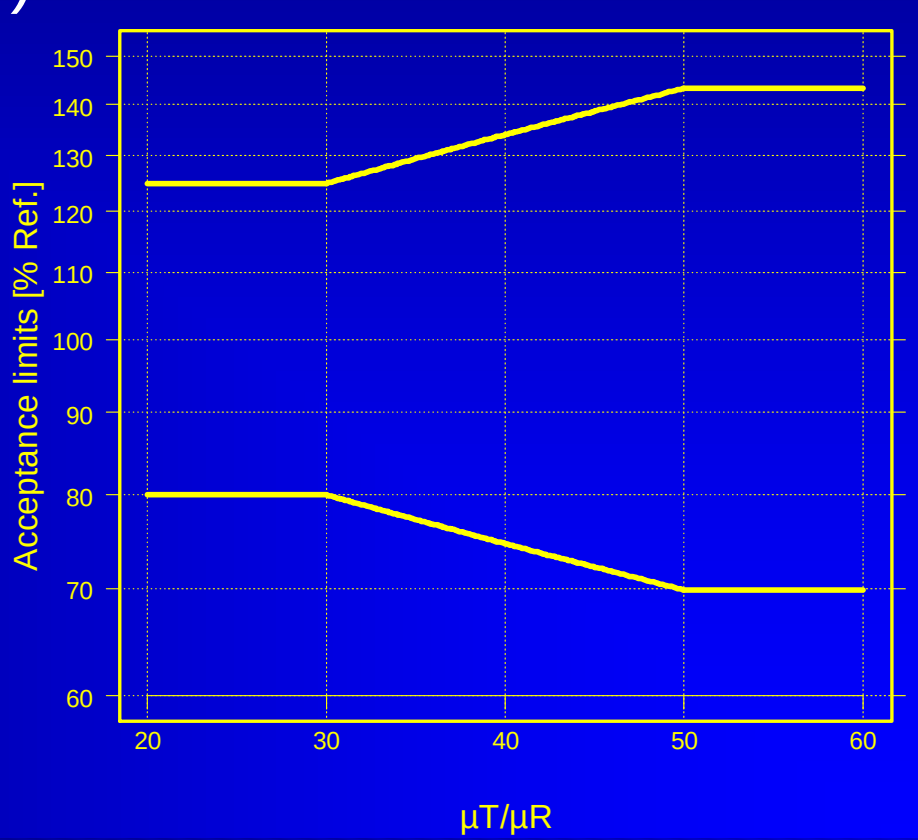
HVDs/HVDPs

- 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 with 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…

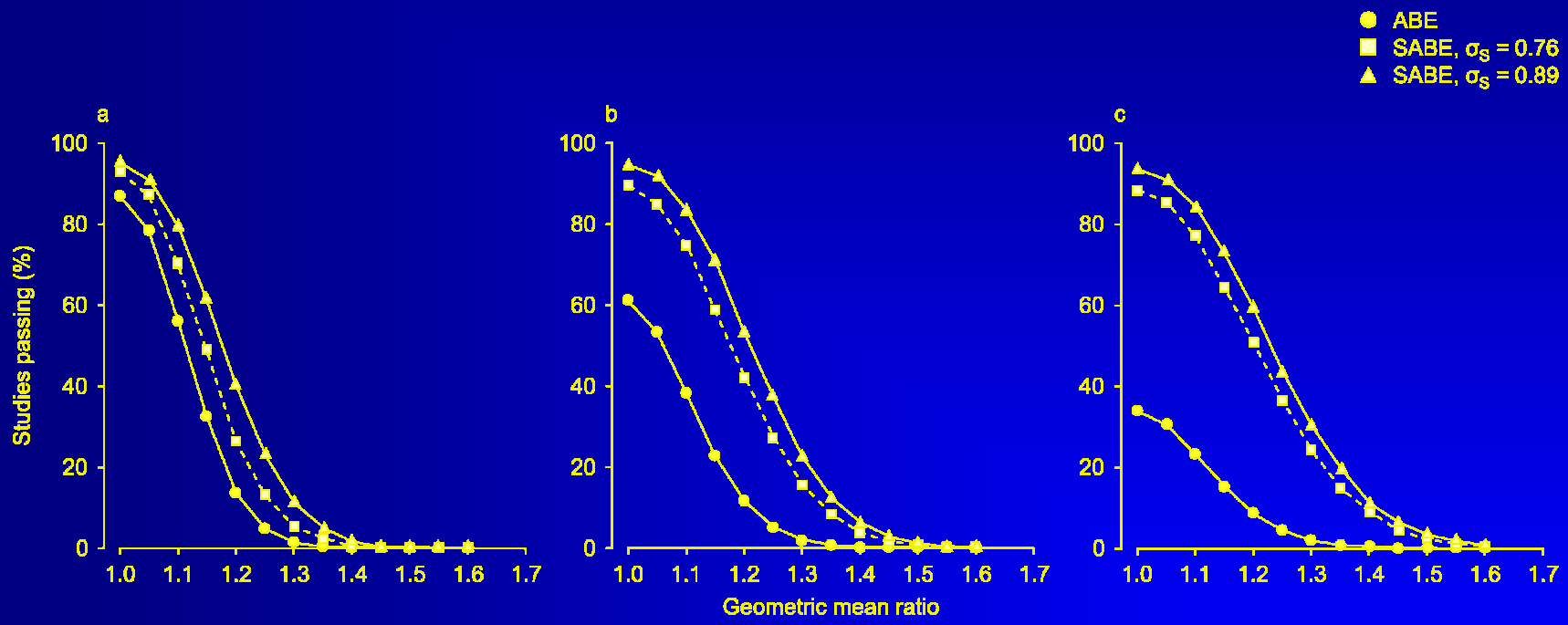
HVDs/HVDPs

- 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



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

- 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
  - 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

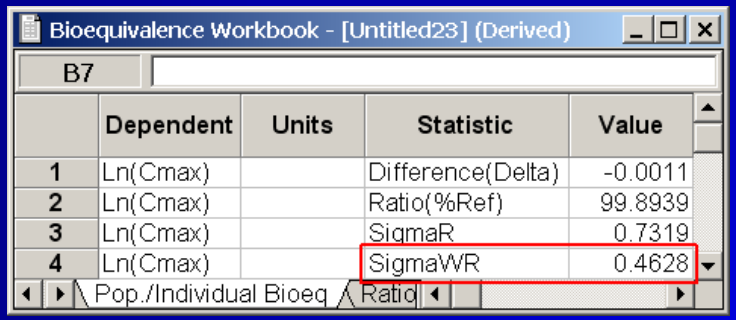
Subj	Seq	Per	Trt	Cmax
1	1	1	R	209.91
1	1	2	T	111.05
1	1	3	R	116.36
2	1	1	R	101.16
2	1	2	T	100.31
2	1	3	R	31.71
3	1	1	R	14.83
3	1	2	T	57.10
3	1	3	R	21.47
4	1	1	R	118.71
4	1	2	T	37.34
4	1	3	R	52.29
5	1	1	R	36.11
5	1	2	T	83.95
5	1	3	R	17.76
6	1	1	R	146.44
6	1	2	T	40.45
6	1	3	R	38.34

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)



Calculate the scaled acceptance range based on the regulatory constant k (0.7601) and the limiting CV_{WR}:

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

σ_{WR}	0.4628
CV_{WR}	0.4887
L	0.7154
U	1.3977

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]
  - 71.54, 139.77
  - 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
```

Read Only | Line 187 / 189

Sensitivity Analysis

- ICH E9
  - Section 3.5 Sample Size, paragraph 3
    - The method by which the sample size is calculated should be given in the protocol […]. The basis of these estimates should also be given.
    - It is important to investigate the sensitivity of the sample size estimate to a variety of deviations from these assumptions and this may be facilitated by providing a range of sample sizes appropriate for a reasonable range of deviations from assumptions.
    - In confirmatory trials, assumptions should normally be based on published data or on the results of earlier trials.

Sensitivity Analysis

- Example

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

nQuery Advisor - [MTE2co-1.nqa]

File Edit View Options Assistants Randomize Plot Window Help

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

	90% power	25% CV	4 drop outs	25% CV + d.o.	PE 90%	worst case
Test significance levels, α (one-sided)	0.050	0.050	0.050	0.050	0.050	0.050
Lower equivalence limit for 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

- Must be done *before* the study *(a priori)*
- **(The Myth of) *a posteriori* Power…**
  - High values do not support the claim of already demonstrated bioequivalence
  - Low values do not invalidate a bioequivalent formulation
  - Further reader:

RV Lenth (2000)
JM Hoenig and DM Heisey (2001)

'***Power****: That which statisticians are always calculating but never have.'*
Stephen Senn, Statistical Issues in Drug Development
Wiley, Chichester, p 197 (2nd ed. 2007)

Pilot Studies

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

Pilot Studies

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

Pilot Studies

- Best description by 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.

Pilot Studies

- Estimated CV has a high degree of uncer-tainty (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 ≥80%) based on a n=10 pilot study

```
require(PowerTOST)
   expsampleN.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%	fixed CV	uncertain CV	ratio
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 unnecessary 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 alter-natives and the choice of how much alpha to spend at the interim analysis is at the company's discretion.

Two-Stage Design

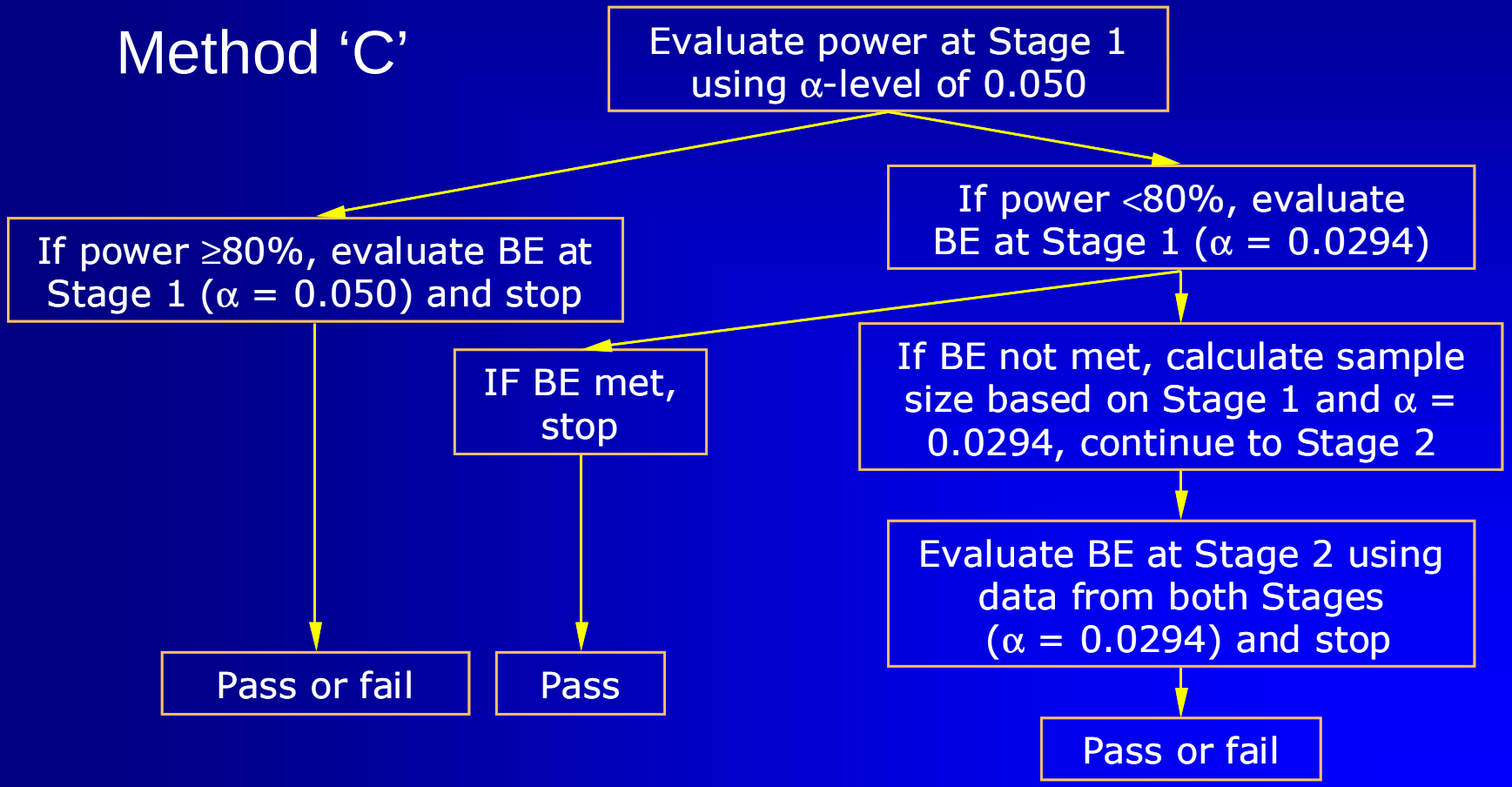
- EMA GL on BE (2010)
  - Section 4.1.8 (cont'd)
    - Plan to use a two-stage approach must be pre-specified in the protocol along with the adjusted significance levels to be used for each of the analyses.
    - When analysing the combined data from the two stages, a term for stage should be included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

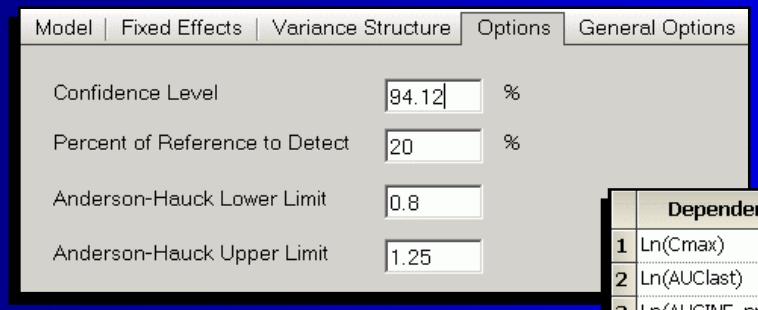


Potvin *et al.* (2007)

- 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 based on Pocock 1977
  - In the pooled analysis (data from Stages 1 + 2), α 0.0294 is used (*i.e.*, the 94.12% Confidence Interval is calculated)



	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 planing).
  - No stop criterion (EMA GL on BE?)
  - Not validated for any other design than the conventional 2×2 crossover (no higher order cross-overs, no replicate designs).

Thank You!

Sample Size Challenges in BE Studies

Open Questions?

(References in your handouts)

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

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

References

- Collection of links to global documents
  http://bebac.at/Guidelines.htm
- ICH
  - E9: Statistical Principles for Clinical Trials (1998)
- EMA-CPMP/CHMP/EWP
  - NfG on the Investigation of BA/BE (2001)
  - Points to Consider on Multiplicity Issues in Clinical Trials (2002)
  - BA/BE for HVDs/HVDPs: Concept Paper (2006); removed form EMEA's website in Oct 2007. Available at http://bebac.at/downloads/14723106en.pdf
  - Questions & Answers on the BA and BE Guideline (2006)
  - Draft Guideline on the Investigation of BE (2008)
  - Questions & Answers: Positions on specific questions addressed to the EWP therapeutic subgroup on Pharmacokinetics (2009)
  - Guideline on the Investigation of BE (2010)
- US-FDA
  - Center for Drug Evaluation and Research (CDER)
    - Statistical Approaches Establishing Bioequivalence (2001)
    - Bioequivalence Recommendations for Specific Products (2007)

- Midha KK, Ormsby ED, Hubbard JW, McKay G, Hawes EM, Gavalas L, and IJ McGilveray
  Logarithmic Transformation in Bioequivalence: Application with Two Formulations of Perphenazine
  J Pharm Sci 82/2, 138-144 (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, 376-378 (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, Steinijans VW, Diletti E, and M Burke
  Sample Size Determination for Bioequivalence Assessment Using a Multiplicative Model
  J Pharmacokin Biopharm 20/5, 557-561 (1992)
- S-C Chow and H Wang
  On Sample Size Calculation in Bioequivalence Trials
  J Pharmacokin Pharmacodyn 28/2, 155-169 (2001)
  Errata: J Pharmacokin Pharmacodyn 29/2, 101-102 (2002)
- DB Owen
  A special case of a bivariate non-central t-distribution
  Biometrika 52, 3/4, 437-446 (1965)

References

- Tothfalusi L, Endrenyi L, and A Garcia Arieta
  Evaluation of Bioequivalence for Highly Variable Drugs with Scaled Average Bioequivalence
  Clin Pharmacokinet 48/11, 725-743 (2009)
- RV Lenth
  Two Sample-Size Practices that I don't recommend
  Joint Statistical Meetings, Indianapolis (2000)
  http://www.math.uiowa.edu/~rlenth/Power/2badHabits.pdf
- JM Hoenig and DM Heisey
  The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis
  The American Statistician 55/1, 19–24 (2001)
  http://www.math.uiowa.edu/~rlenth/Power/2badHabits.pdf
- P Bacchetti
  Current sample size conventions: Flaws, harms, and alternatives
  BMC Medicine 8:17 (2010)
  http://www.biomedcentral.com/content/pdf/1741-7015-8-17.pdf
- B Jones and MG Kenward
  Design and Analysis of Cross-Over Trials
  Chapman & Hall/CRC, Boca Raton (2nd Edition 2000)
- S Patterson and B Jones
  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, 1921-1986 (2004)
- SA Julious and RJ Owen
  Sample size calculations for clinical studies allowing for uncertainty about the variance
  Pharmaceutical Statistics 5/1, 29-37 (2006)
- SA Julious
  Sample Sizes for Clinical Trials
  Chapman & Hall/CRC, Boca Raton (2010)
- LA Gould
  Group Sequential Extension of a Standard Bioequivalence Testing Procedure
  J Pharmacokin Biopharm 23/1, 57-86 (1995)
- Potvin D, Diliberti CE, Hauck WW, Parr AF, Schuirmann DJ, and RA Smith
  Sequential design approaches for bioequivalence studies with crossover designs
  Pharmaceut Statist (2007), DOI: 10.1002/pst.294
  http://www3.interscience.wiley.com/cgi-bin/abstract/115805765/ABSTRACT