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

- 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.

CV_{intra}

α

$1-\beta$

EU

- NfG on the Investigation of BA/BE (2001)
  - Problems/solutions
    - … the error variance associated with the *primary characteristic* to be studied …
      - Since BE must be shown both for AUC and C_{max}, and,
      - if you plan your sample size only for the 'primary characteristic' (*e.g.*, AUC), in many cases you will fail for the secondary parameter (*e.g.*, C_{max}), which most likely shows higher variability – your study will be 'underpowered'.
      - Based on the assumption, that CV is identical for test and reference (what if only the reference formulation has high variability, *e.g.*, some formulations of PPIs?).

EU

- NfG on the Investigation of BA/BE (2001)
  - Problems/solutions
    - … as estimated from
      - a *pilot experiment*,
      - *previous studies*, or
      - *published data*,
    - The correct order should read:

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

EU

- NfG on the Investigation of BA/BE (2001)
  - Problems/solutions
    - … the *significance level desired* …
      - Throughout the NfG the significance level (α, error type I: patient's risk to be treated with a bio***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)

EU

- NfG on the Investigation of BA/BE (2001)
  - Problems/solutions
    - … the *required power*.
      - Generally the power is set to at least 80 % (β, error type II: producers's risk to get no approval for a bioequivalent drug; power = 1 – β).
        Remember: *1 out of 5 studies will fail just by chance!*
      - If you plan for power of less than 70 %, problems with the ethics committee are likely (ICH E9).
      - If you plan for power of more than 90 % (especially with low variability drugs), problems with the regulator are possible ('forced bioequivalence').
      - Add subjects ('alternates') according to the expected drop-out rate!

EU

- NfG on the Investigation of BA/BE (2001)
  - Problems/solutions
    - … the *expected deviation (Δ) from the reference* …
      - Reliable estimate only from a previous full-sized study
      - If you are using data from a pilot study, allow for a safety margin
      - If no data are available, commonly a GMR (geometric test/reference-ratio) of 0.95 (Δ = 5%) is used
      - If more than Δ = 10% is expected, questions from the ethics committee are likely
      - BE GL (2010) batches must not differ more than 5%.

EU

- EMA BE Guideline (2010)
  - 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?

α- vs. β-Error

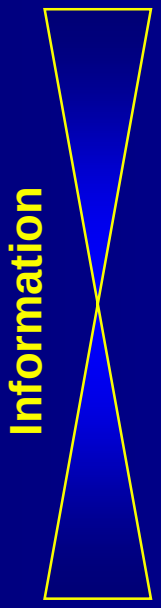
- α-Error: Patient's risk to be treated with a bioinequivalent formulation.
  - Although α is generally set to 0.05, sometimes $<$0.05 (*e.g.*, NTDIs in Brazil, multiplicity, interim analyses).
- β-Error: Producer's risk to get no approval for a bioequivalent formulation.
  - Generally *set* in study planning to $\leq$0.2, where power = 1 – β = $\geq$80%.
  - There is no *a posteriori* (aka *post hoc*) power!

Either a study has demonstrated BE or not.

Phoenix/WinNonlin's output is statistical nonsense!

Hierarchy of Designs

- The more ‘sophisticated’ a design is, the more information can be extracted.
  - Hierarchy of designs:



Full replicate (TRTR | RTRT)
Partial replicate (TRR | RTR | RRT)
Standard 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!)
  - 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)

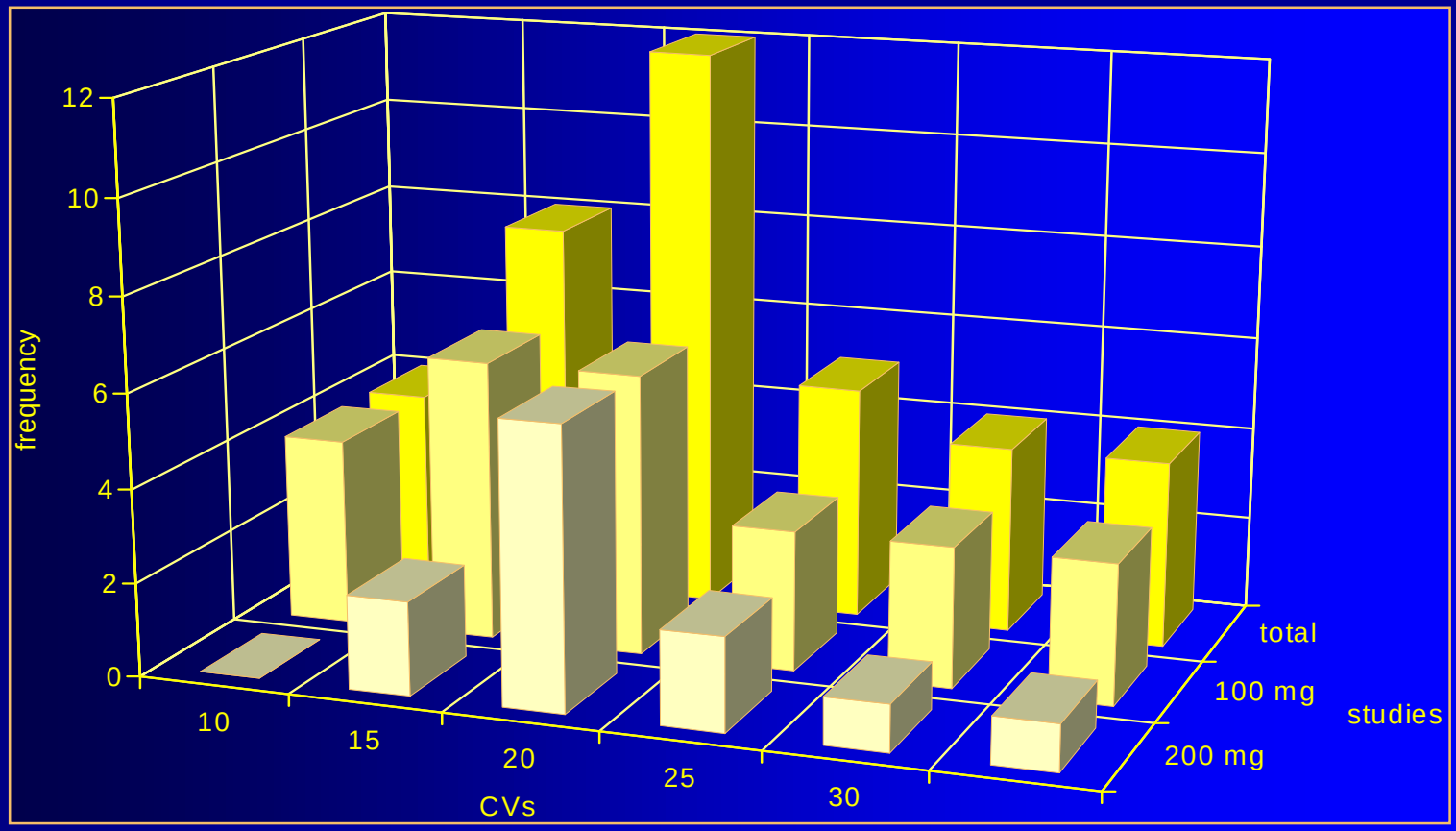
drug, formulation	design	n	metric	CV_{intra}	CV_{inter}	CV_{total}	$\%_{intra/total}$
methylphenidate MR	SD	12	AUC_t	7.00	19.1	20.4	34.3
paroxetine MR	MD	32	AUC_τ	25.2	55.1	62.1	40.6
lansoprazole DR	SD	47	C_{max}	47.0	25.1	54.6	86.0

    - … pilot study unavoidable

Literature Data

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

Literature Data



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

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

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

Pooling of CV%

- Intra-subject CV from different studies can be pooled (LA Gould 1995, Patterson and Jones 2006)
  - In the parametric model of log-transformed data, additivity of variances (not of CVs!) apply.
  - Do not use the arithmetic mean (or the geometric mean either) of CVs.
  - Before pooling variances must be weighted acccording to the studies' sample size – larger studies are more influentual than smaller ones.

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 $\alpha = 0.25$)

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

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	0.25	0.75	1.2540	**0.254**	~~0.245~~
		$\chi^2_{(\alpha,df)}$	15.452	0.291	+14.3%

CV_{intra}	n	seq.	df (mj)	σ_w	σ^2_w	σ^2_w × 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	$1-\alpha$	total	CV_{pooled}	CV_{mean}
32	0.25	0.75	2.2881	**0.272**	~~0.245~~
		$\chi^2_{(\alpha,df)}$	26.304	0.301	+10.7%

CV_{intra}	n	seq.	df (mj)	σ_w	σ^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	$1-\alpha$	total	CV_{pooled}	CV_{mean}
32	0.25	0.75	1.7246	**0.235**	~~0.245~~
		$\chi^2_{(\alpha,df)}$	26.304	0.260	+10.6%

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

Pooling of CV%

- *R* package *PowerTost* function *CVpooled*, data of last example.

```
require(PowerTOST)
CVs <- ("
  PKmetric |  CV  |  n | design | source
     AUC   | 0.20 | 24 |  2x2   | study 1
     AUC   | 0.30 | 12 |  2x2   | study 2
")
txtcon <- textConnection(CVs)
CVdata <- read.table(txtcon, header=TRUE, sep="|",
            strip.white=TRUE, as.is=TRUE)
close(txtcon)
CVsAUC <- subset(CVdata,PKmetric=="AUC")
print(CVpooled(CVsAUC, alpha=0.25), digits=3, verbose=TRUE)

Pooled CV = 0.235 with 32 degrees of freedom
Upper 75% confidence limit of CV = 0.260
```

Pooling of CV%

- Or you may combine pooling with an estimated sample size based on uncertain CVs (we will see later what that means).
  R package *PowerTost* function *expsampleN.TOST*, data of last example.
  CVs and degrees of freedom must be given as vectors:
  CV = c(0.2,0.3), dfCV = c(22,10)

Pooling of CV%

```
require(PowerTOST)
expsampleN.TOST(alpha=0.05,
  targetpower=0.8,
  theta1=0.8, theta2=1.25,
  theta0=0.95, CV=c(0.2,0.3),
  dfCV=c(22,10), alpha2=0.05,
  design="2x2", print=TRUE,
  details=TRUE)
```

```
++++++++ Equivalence test - TOST ++++++++
   Sample size est. with uncertain CV
------------------------------------------
Study design:  2x2 crossover
Design characteristics:
df = n-2, design const. = 2, step = 2

log-transformed data (multiplicative model)

alpha = 0.05, target power = 0.8
BE margins        = 0.8 ... 1.25
Null (true) ratio = 0.95
Variability data
  CV df
 0.2 22
 0.3 10
CV(pooled)          = 0.2353158 with 32 df
one-sided upper CL = 0.2995364 (level = 95%)

Sample size search
 n    exp. power
24   0.766585
26   0.800334
```

Power Curves

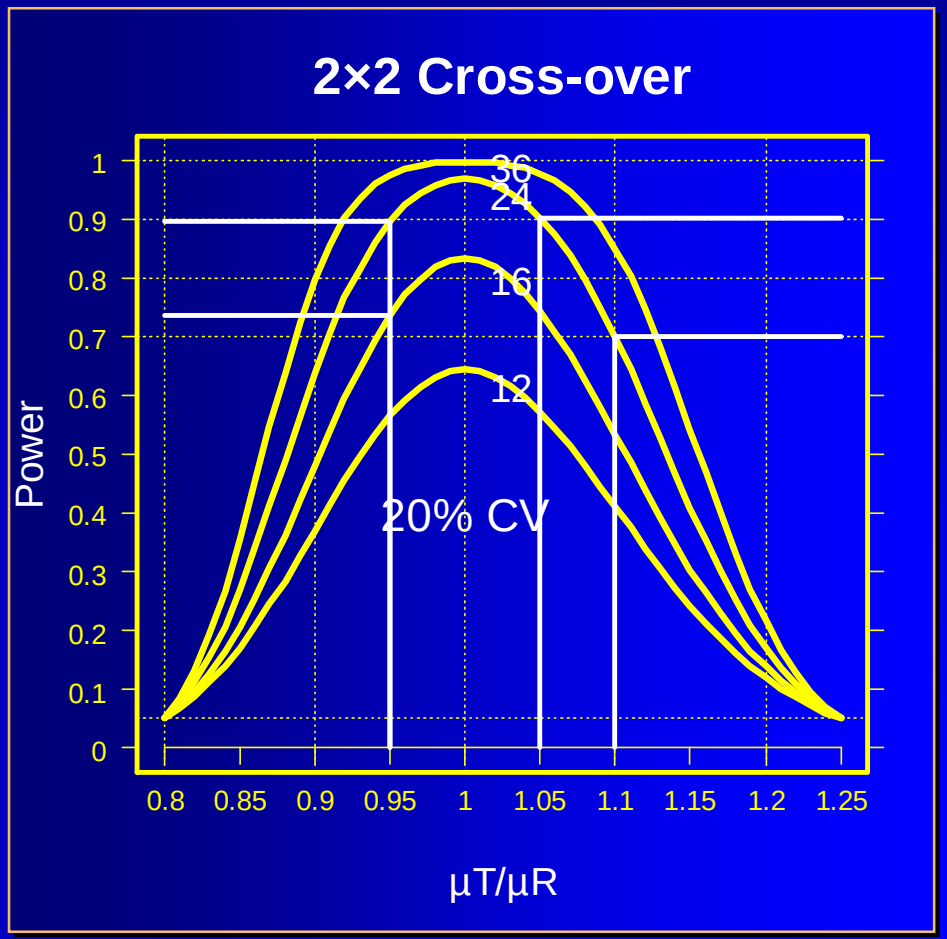
Power to show

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



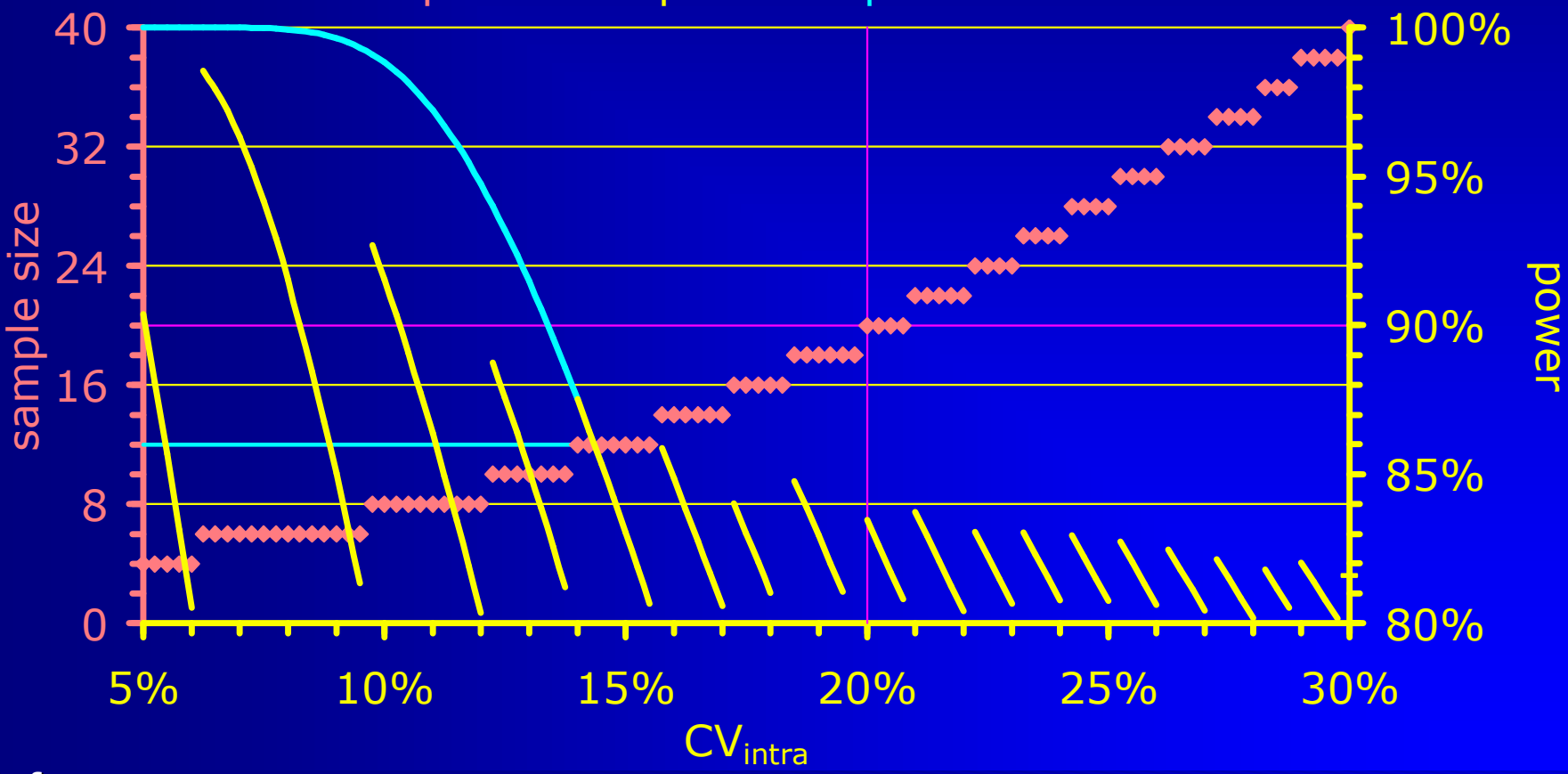
Power *vs.* Sample Size

- It is not possible to *directly* calculate the required 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

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



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*)

Detlew Labes
PowerTOST: Power and Sample size based on two one-sided t-tests (TOST) for bioequivalence studies
Version 0.8-5, 2011-05-16 http://cran.r-project.org/web/packages/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 0.8-2 (2011)	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

original values	Method	Algorithm	CV% 22.5	24	25	26	27.5	28	30	32	34	36	38	40
PowerTOST 0.8-2 (2011)	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

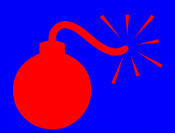
1. $\Delta = \ln(0.8)-\ln(T/R) = -0.1719$
2. Start with e.g. n=8/sequence
   1. $df = n \cdot 2 - 1 = 8 \times 2 - 1 = 14$
   2. $t_{\alpha,df} = 1.7613$
   3. $t_{\beta,df} = 0.8681$
   4. new $n = [(t_{\alpha,df} + t_{\beta,df})^2 \cdot (CV/\Delta)]^2 =$ $(1.7613+0.8681)^2 \times (-0.2/0.1719)^2 =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

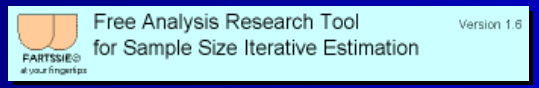
1. $\Delta = \ln(T/R) - \ln(1.25) = 0.1719$
2. Start with e.g. n=8/sequence
   1. $df_\alpha = \text{roundup}(2 \cdot n-2) \cdot 2-2 = (2 \times 8-2) \times 2-2 = 26$
   2. $df_\beta = \text{roundup}(4 \cdot n-2) = 4 \times 8-2 = 30$
   3. $t_{\alpha,df} = 1.7056$
   4. $t_{\beta/2,df} = 0.8538$
   5. 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 / S+ →

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

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

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

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

Sensitivity Analysis

- ICH E9 (1998)
  - Section 3.5 Sample Size, paragraph 3
    - The method by which the sample size is calculated should be given in the protocol […]. The basis of these estimates should also be given.
    - It is important to investigate the sensitivity of the sample size estimate to a variety of deviations from these assumptions and this may be facilitated by providing a range of sample sizes appropriate for a reasonable range of deviations from assumptions.
    - In confirmatory trials, assumptions should normally be based on published data or on the results of earlier trials.

Sensitivity Analysis

- Example

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

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

- Example

  PowerTOST, function *sampleN.TOST*

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

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

Sensitivity Analysis

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

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

Sensitivity Analysis

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

RV Lenth (2000)
JM Hoenig and DM Heisey (2001)
P Bacchetti (2010)

Data from Pilot Studies

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

Justification

- 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.

Justification

- 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

Application

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

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, theta0=0.95, CV=0.40,
  dfCV=24-2, alpha2=0.05, design="2x2")
```

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

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

Pilot Studies: Sample Size

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

Solutions

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

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

Two-Stage Design

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

'Internal Pilot Study Design'

Two-Stage Design

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

Two-Stage Design

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

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).

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

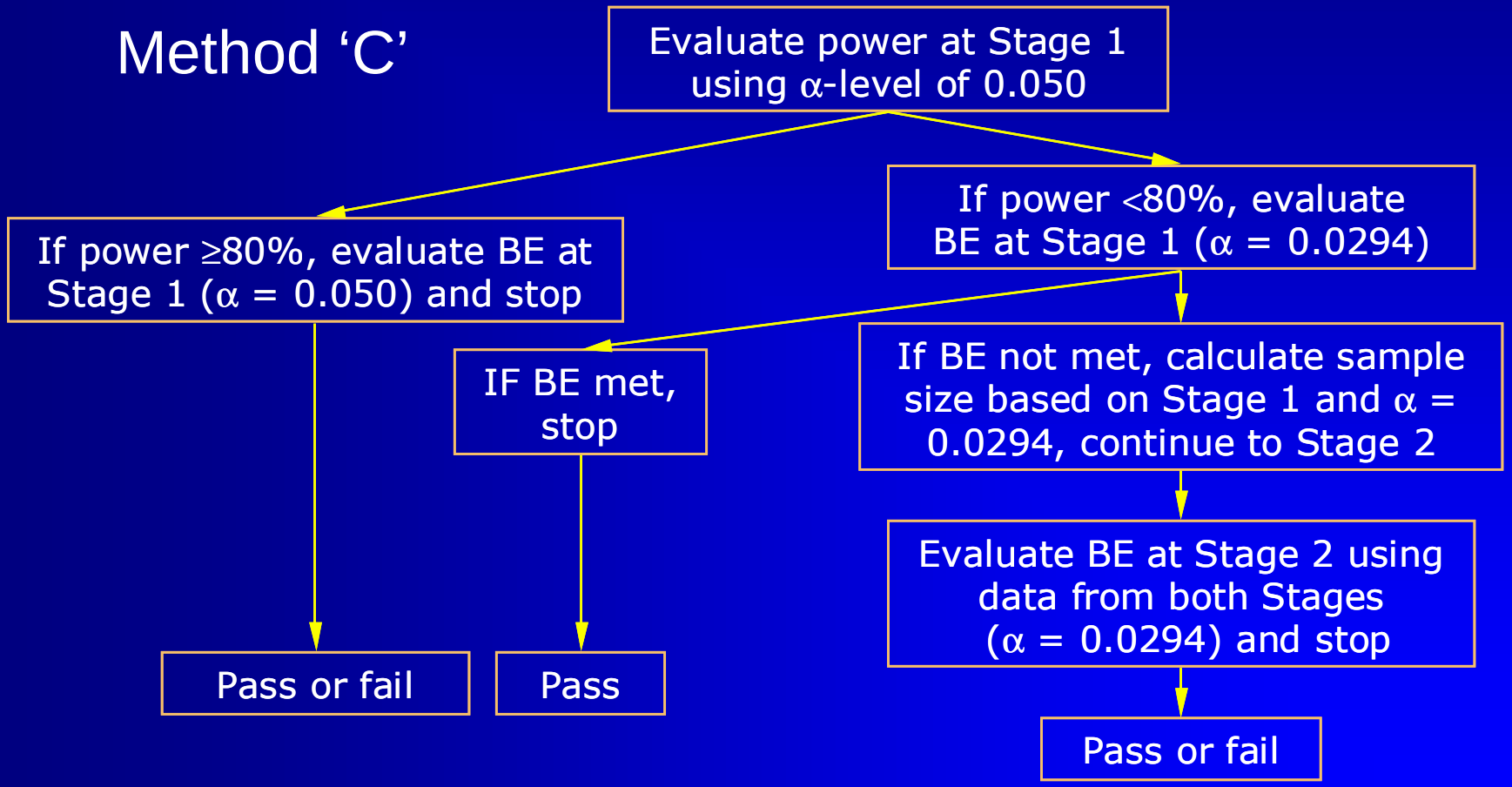
Sequential Designs

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

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, 245–262 (2008), DOI: 10.1002/pst.294
http://www3.interscience.wiley.com/cgi-bin/abstract/115805765/ABSTRACT

Potvin *et al.* (2008)

Method 'C'



Potvin *et al.* (2008)

- Technical Aspects
  - Only *one* Interim Analysis (after Stage 1)
  - If possible, use software (too wide step sizes in Diletti's tables)
  - Should be called '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.* (2008)

- Technical Aspects (cont'd)
  - Adjusted α of 0.0294 (Pocock 1977)
  - If power is <80% in Stage 1 and in the pooled analysis (data from Stages 1 + 2), α 0.0294 is used (*i.e.*, a 1–2×α = 94.12% CI is calculated)
  - Overall patients' risk is preserved at ≤0.0502

Potvin *et al.* (2008)

- 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 for significant differences between Stages.

Potvin *et al.* (2008)

```
Model Specification and User Settings
      Dependent variable : Cmax (ng/mL)
               Transform : LN
           Fixed terms : int+Sequence+Treatment+Period
   Random/repeated terms : Sequence*Subject

Final variance parameter estimates:
    Var(Sequence*Subject)     0.0444152
          Var(Residual)       0.071194
       Intrasubject CV        0.271642

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=  1.593384  SE=  0.123689  GeoLSM=   4.920373
----------------------------------------------------------------------------
Test:      Test        LSMean=  1.471058  SE=  0.123689  GeoLSM=   4.353839

    Difference =     -0.1223,  Diff_SE=    0.1958,  df= 12.0
    Ratio(%Ref) =    88.4860

                     Classical
    CI  90% = (   62.4145,  125.4478)
    CI User = (   58.7888,  133.1845)
    Failed to show average bioequivalence for confidence=94.12 and percent=20.0.
```

14 subjects in Stage 1, conventional BE model

CV_{intra} 27.2%

α 0.0294 (if power <80%)

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

Potvin *et al.* (2008)

```
require(PowerTOST)
power.TOST(alpha=0.05, logscale=TRUE,
           theta1=0.8, theta2=1.25, theta0=0.95,
           CV=0.271642, n=14,
           design = "2x2", exact = TRUE)

[1] 0.3189318
```

Expected ratio 95% – not 88.5% observed in stage 1! CV_{intra} 27.2%, 14 subjects in Stage 1

Power 31.9% – initiate Stage 2

```
sampleN.TOST(alpha=0.0294, targetpower=0.8, logscale=TRUE,
             theta1=0.8, theta2=1.25, theta0=0.95,
             CV=0.271642, design = "2x2", exact = TRUE,
             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 = 0.271642

Sample size
 n     power
40   0.817146
```

Calculate total sample size: expected ratio 95%, CV_{intra} 27.2%, 80% power

Total sample size 40: 26 in Stage 2 (28 recruited)

Potvin *et al.* (2008)

```
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)   0.0430110
          Var(Residual)       0.0376772
        Intrasubject CV       0.1959489

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=  1.520255  SE=  0.047872  GeoLSM=  4.573390
-------------------------------------------------------------------------
Test:      Test        LSMean=  1.525145  SE=  0.047872  GeoLSM=  4.595809

    Difference =      0.0049,  Diff_SE=    0.0496,  df= 38.0
    Ratio(%Ref) =   100.4902

                     Classical
    CI  90% = (   92.4329,  109.2499)
    CI User = (   91.2387,  110.6797)
    Average bioequivalence shown for confidence=94.12 and percent=20.0.
```

27 subjects in Stage 2 (41 total), modified model for pooled analysis

α 0.0294 in pooled analysis

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

Sequential Designs

- Methods by Potvin *et al.* (2008) limited to point estimates of 0.95 and 80% power
  - Follow-up paper by Montague *et al.* (2011)
    - Slight inflation of patient's risk (α 0.0547) observed in Methods B/C if PE 0.90 was used
    - New Method D (α 0.028)
    - Similar α-adjustment might be usefull if PE 0.95 and power 90% as well, but is not validated yet!
  - Further work has to be done for arbitrary combinations of PE/power or even adjusting on the PE observed in stage (full adaptive design).

Thank You!

Determining Optimal Sample Size

Open Questions?

(References in the online PDF)

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

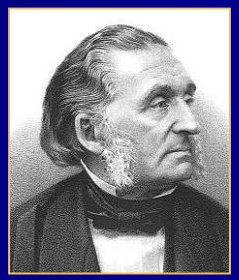
To bear in Remembrance...

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

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

Power Calculation – A guess masquerading as mathematics.

Stephen Senn



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

Armand Trousseau

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 Pharmacokinetics Working Party (2011)
  - 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)
- 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 (2008), DOI: 10.1002/pst.294
- 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. (2011), DOI: 10.1002/pst.483