Power and intra-subject variability in 2 stage approaches to bioequivalence approval

Helmut Schütz
BEBAC

Wikimedia Commons 2010 • EmDee • Creative Commons Attribution-ShareAlike 3.0 Unported

Overview

- ‘Classical’ sample size estimation in BE
  - Patient’s & producer’s risk
  - Power in study planning
- History / early approaches
  - Add-on studies
  - Problems with α-inflation
- Uncertainties
  - Variability
  - Test/Reference-ratio
  - Sensitivity analysis

Overview

- Recent developments
  - Review of guidelines
  - Multi-sequential designs
  - Two-stage sequential designs
- Open issues
  - Feasibility / futility rules
  - Arbitrary PE and/or power; adaption for stage 1 PE
  - Dropping a candidate formulation from a higher-order X-over
  - Application to replicated designs (for HVDs/HVDPs)

α- vs. β-Error

- All formal decisions are subjected to two types of error:
  - Error Type I (*α*-Error, Risk Type I)
  - Error Type II (*β*-Error, 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- vs. β-Error

- 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**

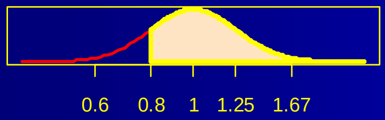
- In BE-testing the null hypothesis is bioinequivalence ($\mu_1 \neq \mu_2$)!

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- vs. β-Error

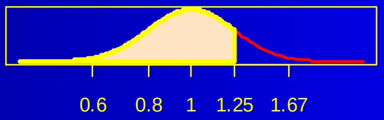
- α-Error: Patient's Risk to be treated with a bioinequivalent formulation (H_0 falsely rejected)
  - BA of the test compared to reference in a *particular* patient is risky *either* below 80% *or* above 125%.
  - If we keep the risk of particular patients at 0.05 (5%), the risk of the entire population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

95% one-sided CI



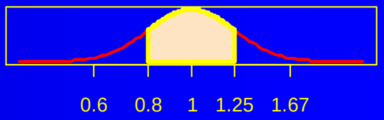
particular patient

95% one-sided CI



particular patient

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



population of patients

α- vs. β-Error

- β-Error: Producer’s Risk to get no approval for a bioequivalent formulation (H_0 falsely not rejected)
  - *Set* in study planning to ≤0.2, where power = 1 – β = ≥80%
  - If power is set to 80 %

One out of five studies will fail just by chance!

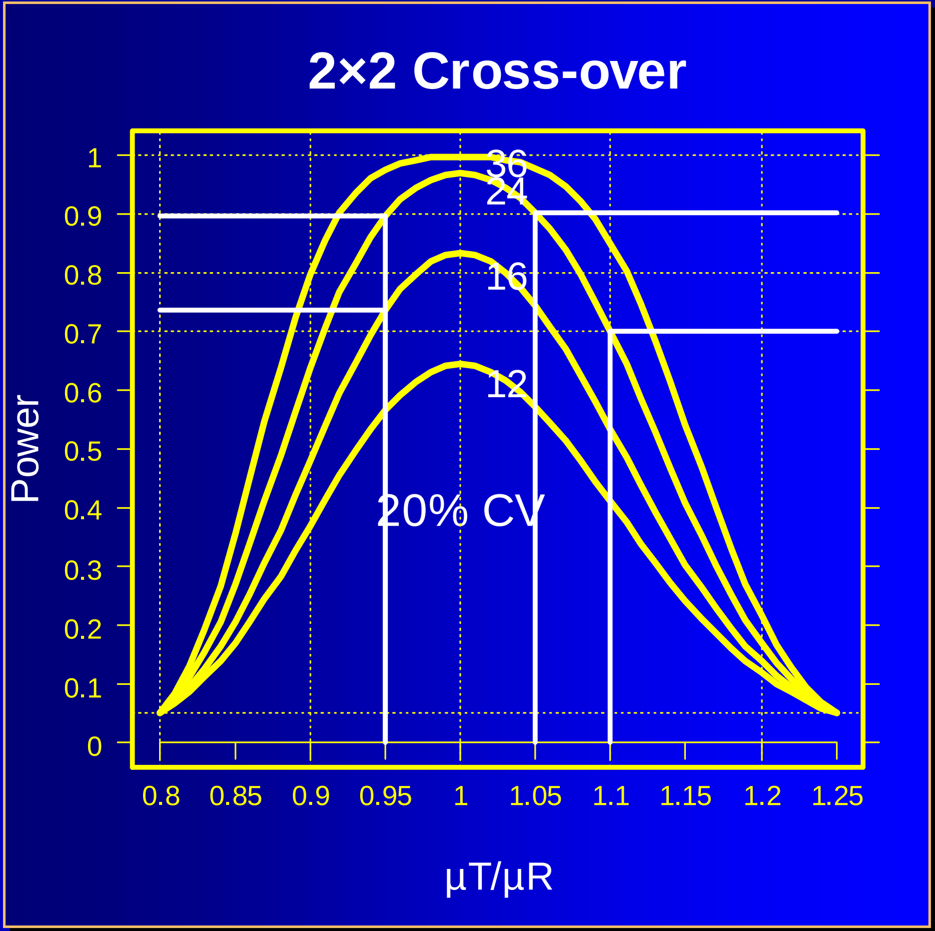
α 0.05	**BE**
not BE	**β 0.20**

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



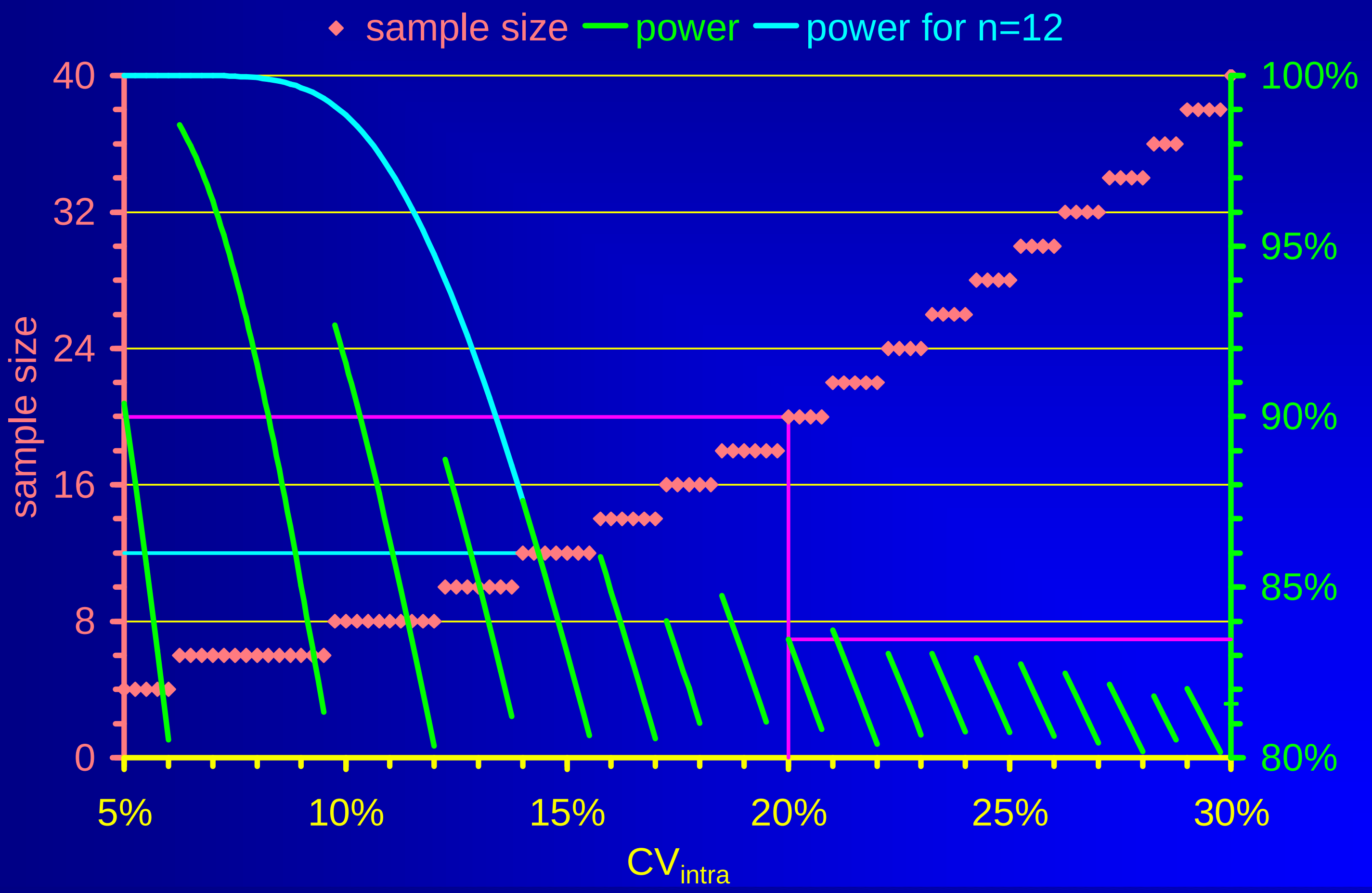
Power *vs.* Sample Size

- It is not possible to calculate the required sample size *directly*.
- Power is calculated 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 (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%



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 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)

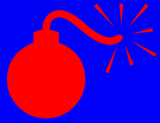
Patient's risk α 0.05, Power 80% (Producer's risk β 0.2), AR [0.80 – 1.25], CV 0.2 (20%), T/R 0.95

1. 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

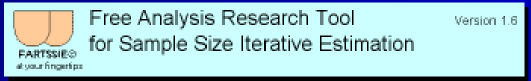
1. Δ = ln(T/R) – ln(1.25) = 0.1719
2. Start with e.g. n=8/sequence
    1. df_{α} = roundup(2·n-2)·2-2 = (2×8-2)×2-2 = 26
    2. df_{β} = roundup(4·n-2) = 4×8-2 = 30
    3. $t_{\alpha,df}$ = 1.7056
    4. $t_{\beta/2,df}$ = 0.8538
    5. new n = $\beta^2 \cdot [(t_{\alpha,df} + t_{\beta/2,df})^2/\Delta^2$ = $0.2^2 \times (1.7056+0.8538)^2 / 0.1719^2$ = 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 the 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+ →

```
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")
```

- 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)
```

History / early approaches

- Sometimes properly planned studies fail due to
  - Pure chance (producer's risk hit)
  - False 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

History / early approaches

- *'The primary concern in bioequivalence assessment is to limit the risk of erroneously accepting bioequivalence. Only statistical procedures which do not exceed the nominal risk of 5% can be approved, and among them the one with the smallest risk of erroneously rejecting bioequivalence should be selected.'* *
- Performing a second study and pooling data with the first's not acceptable
- Performing a (much larger) second study and base BE on this study *alone* was (and is) acceptable

* **CPMP Working Party**
Investigation of Bioavailability and Bioequivalence: Note for Guidance
Section 3.6 Data analysis, Document Ref. III/54/89-EN (1 May 1992)

History / early approaches

- Inflation/preservation of patient’s risk
  - Repeated tests increase the overall significance level. For two tests the overall level is ~ 8%[1]
  - With two repeated tests at 2.94% overall α ~ 5%[2]
  - Derived for tests assuming normally distributed data with known variances. Approximately valid if sample size not too small.

[1] **Armitage P, McPherson K, and BC Rowe**
Repeated significance tests on accumulating data
J R Statist Soc A 132, 235–44 (1969)

[2] **SJ Pocock**
Group sequential methods in the design and analysis of clinical trials
Biometrika 64, 191–9 (1977)

History / early approaches

- However naïve pooling (*without* α-adjustment) was performed in the past
  - Statistical model modified in order to include a formulation-by-study interaction factor.
  - Test for homogeneity of error variances between studies
  - Pooling only acceptable if both tests not significant*

* **H Mellander**
Problems and Possibilities with the Add-On Subject Design, in:
Midha KK, Blume HH (eds.)
Bio-International. Bioavailability, Bioequivalence and Pharmacokinetics
medpharm Scientific Publishers, Stuttgart, pp. 85–90 (1993)

Add-on Design

- According to Canadian guidances (1992+)
  - Pooling of two *or more* [sic!] studies may be allowed
  - Model:

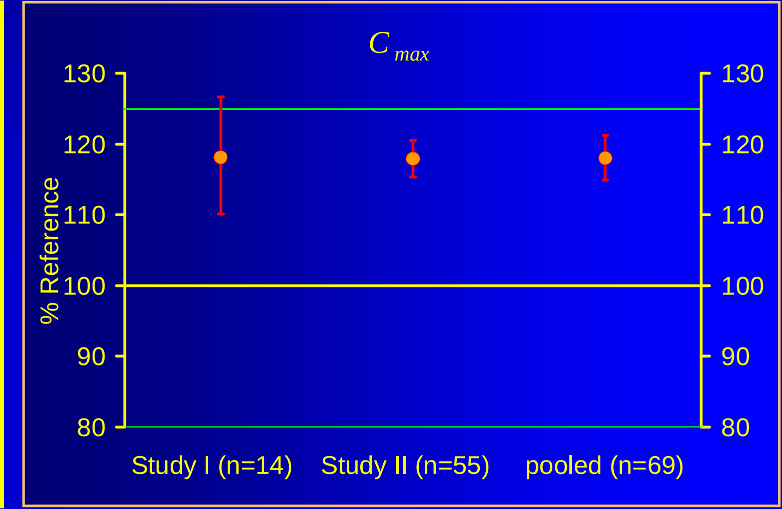
```
Study + Subject(Study) + Period(Study) +
Treatment + Treatment × Study
```

  - Consistency tests
    - Test for equality of residual mean squares: Ratios of MSE of the 1st study to all others; smaller value used as denominator. *F*-test at 5%.
    - Formulation-by-study interaction. *F*-test at 5%.
    - If *both* tests not significant, pooling without (!) α-adjustment

Add-on Design

- Example (C_{max}, SD fasting studies)

Study	I	II	pooled
n	14	55	69
PE%	118.14	117.93	118.04
CI%	110.16 126.70	115.40 120.53	114.91 121.25
MSE	0.01078	0.004645	0.005777
CV_{intra}	10.41	6.82	7.61



- MSE-ratio 2.3198: p($F_{12,53}$) 0.01812
- Study-by-formulation interaction: p($F_{1,65}$) 0.9573
- Pooling not allowed due to lacking equality of MSEs

Problems with α-inflation

- Patient's risk likely is not preserved
  - The probability to obtain at least one significant result with k independent (!) t-tests (at level α) is

$$P(k) = 1 - (1-\alpha)^k$$

$$P(2) = 1 - (1-0.05)^2 = 0.0975$$

  - Bonferroni-correction for two studies would mandate calculation of a 95% confidence interval

$$\alpha_{adj} = \alpha / k$$

$$P_{adj}(2) = 1 - (1-0.025)^2 = 0.04938 < 0.05$$

  - Applicability doubtful since no *independent* tests!

Problems with α-inflation

- Patient's risk (cont'd)
  - For two repeated tests on accumulating data the overall level is ~8% (Armitage 1969)
  - In naïve pooling the variance will be underestimated[1]
  - Simulations of BE studies (sample sizes 24 – 48, CV_{intra} 19 – 37%, 1 – 3 interim looks, Lan-DeMets sequential method, 1540 studies in all combinations) showed empirical α of up to 5.97%[2]

[1] **Wittes J, Schabenberger O, Zucker D, Brittain E, and M Proschan**
Internal pilot studies I: type I error rate of the naïve t-test
Statistics in Medicine 18, 3481–91 (1999)

[2] **Hauck WW, Preston PE, and FY Bois**
A group sequential approach to crossover trials for average bioequivalence
Journal of Biopharmaceutical Statistics 7(1), 87–96 (1997)

Problems with α-inflation

- Patient's risk (cont'd)
  - Simulations of 1 Mio BE studies (12 subjects in 1st study, CV_{intra} 20%, sample size re-estimation based on PE 0.95 and CV_{intra} of 1st study) showed empirical α of 5.84%*
  - Naïve pooling without α-adjustment (Add-on designs, internal pilot designs) should be avoided!

* **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

Uncertainties

- CV_{intra} used in sample size estimation is not set in stone but an *estimate*!
  - Sample sizes for target power 90%, PE 0.95, CV_{intra} 20% → n=26
  - Not done yet! What if $CV_{intra} \neq$ 20%?

CV_{intra}	n	$power_n$	$power_{n=26}$
15	16	0.92602	0.99153
16	18	0.92685	0.98379
17	20	0.92601	0.97253
18	22	0.92400	0.95763
19	24	0.92114	0.93922
20	**26**	0.91763	0.91763
21	28	0.91362	0.89329
22	30	0.90919	0.86659
23	32	0.90443	0.83794
24	36	0.91451	0.80767
25	38	0.90889	0.77606

Uncertainties

- According to 2010 GL test and reference batches should not differ in measured content by >±5%
  - n=26, CV_{intra} 20%, PE 0.95 → power 91.76%
  - What about analytical error?

PE	power
0.90	0.66945
0.91	0.73684
0.92	0.79577
0.93	0.84547
0.94	0.88591
0.95	**0.91763**
0.96	0.94154
0.97	0.95867
0.98	0.97003
0.99	0.97646
1.00	0.97853

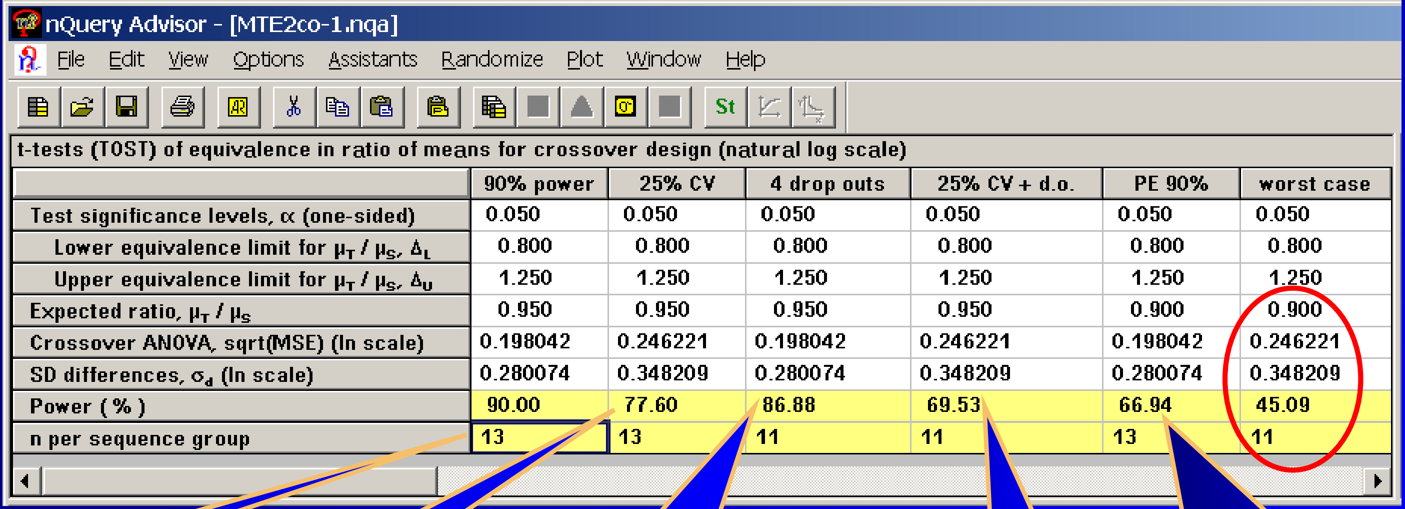
Sensitivity Analysis

- 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, 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 (*a posteriori* or *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
    Two Sample-Size Practices that I don't recommend (2000)
    JM Hoenig and DM Heisey
    The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis (2001)
    P Bacchetti
    Current sample size conventions: Flaws, harms, and alternatives (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

Recent developments

- Review of guidelines
  - WHO (May 2006)
    - Add-on studies
      - Declared in the protocol
      - Appropriate statistical treatment
      - Japanese GL given as an example
  - South Africa (Jul 2007)
    - Add-on studies
      - Declared in the protocol
      - Maximum sample size *a priori*
      - No recommendations about statistical analysis

Recent developments

- Review of guidelines
  - Japan (Nov 2006); no essential change to Dec 1997
    - Add-on studies
      - Sample size at least 50% of 1st study
      - ‘Study’ as a factor in the analysis
      - No consistency tests
      - No Bonferroni-correction
      - If sample size of 1st study ≥20 or
        sample size of pooled studies ≥30
        BE may be assessed on PE (within 0.90 – 1.11) and
        dissolution similarity (no CI)
  - Argentina (Sep 2006, Mar 3007)
    - Sequential Designs: not statisticals details

Recent developments

- Review of guidelines
  - New Zealand (Oct 2001)
    - Sequential Designs
      - Declared in the protocol
      - Maximum sample size *a priori* (≤40!)
      - ‘Appropriate statistical tests (e.g., sequential *t*-test)’
  - FDA
    - Sequential Designs: not mentioned in guidances but acceptable (pers. comm. Barbara Davit, May 2010)
  - EMA (Jan 2010)
    - Sequential Designs: fairly detailed informations given

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.

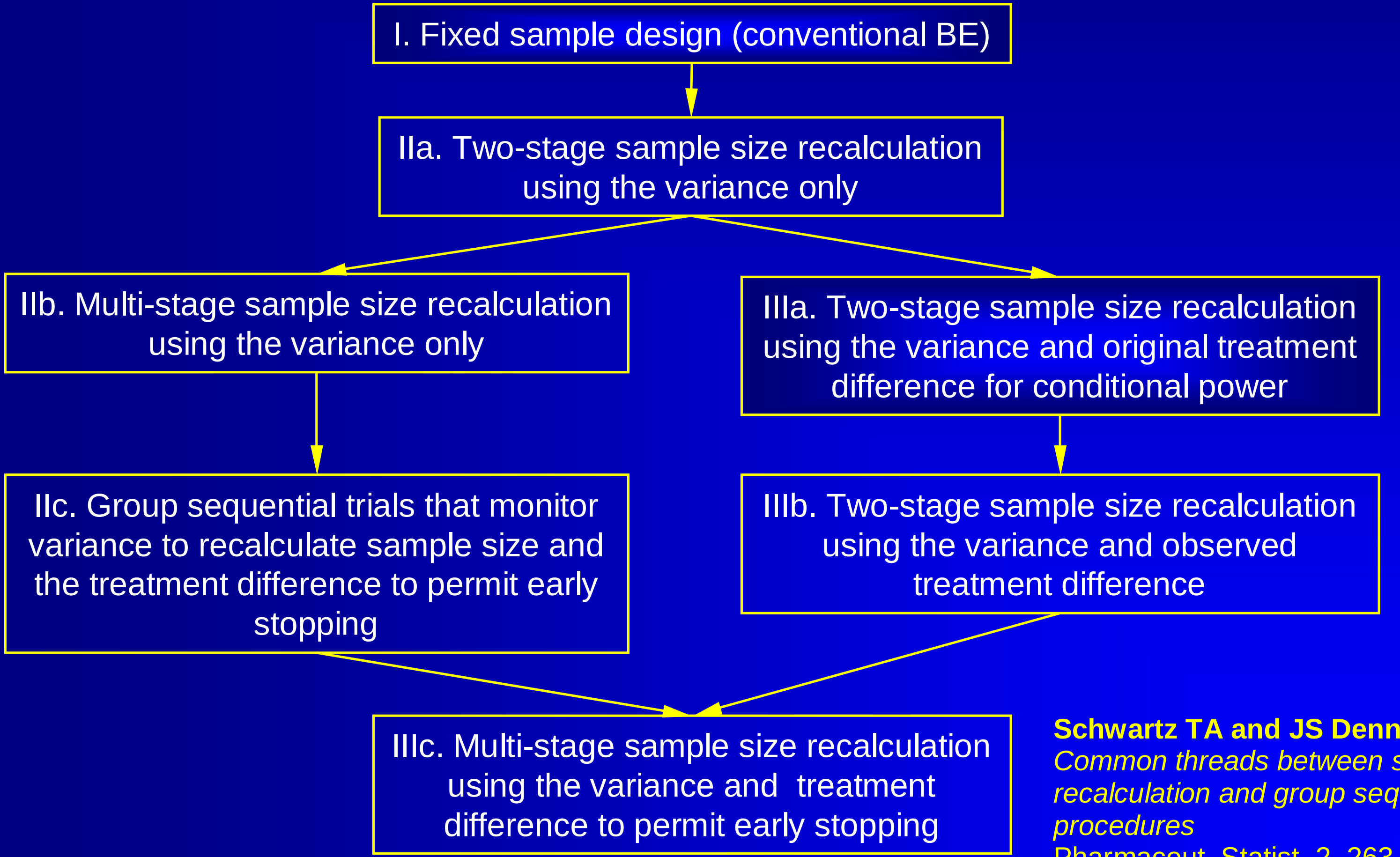
Two-Stage Design

- 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 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.

Classification



I. Fixed sample design (conventional BE)

IIa. Two-stage sample size recalculation using the variance only

IIb. Multi-stage sample size recalculation using the variance only

IIIa. Two-stage sample size recalculation using the variance and original treatment difference for conditional power

IIc. Group sequential trials that monitor variance to recalculate sample size and the treatment difference to permit early stopping

IIIb. Two-stage sample size recalculation using the variance and observed treatment difference

IIIc. Multi-stage sample size recalculation using the variance and treatment difference to permit early stopping

Schwartz TA and JS Denne
Common threads between sample size recalculation and group sequential procedures
Pharmaceut. Statist. 2, 263–71 (2003)

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
    - stated in Canadian draft guidance (2010) and EMA's BE guideline (2010).

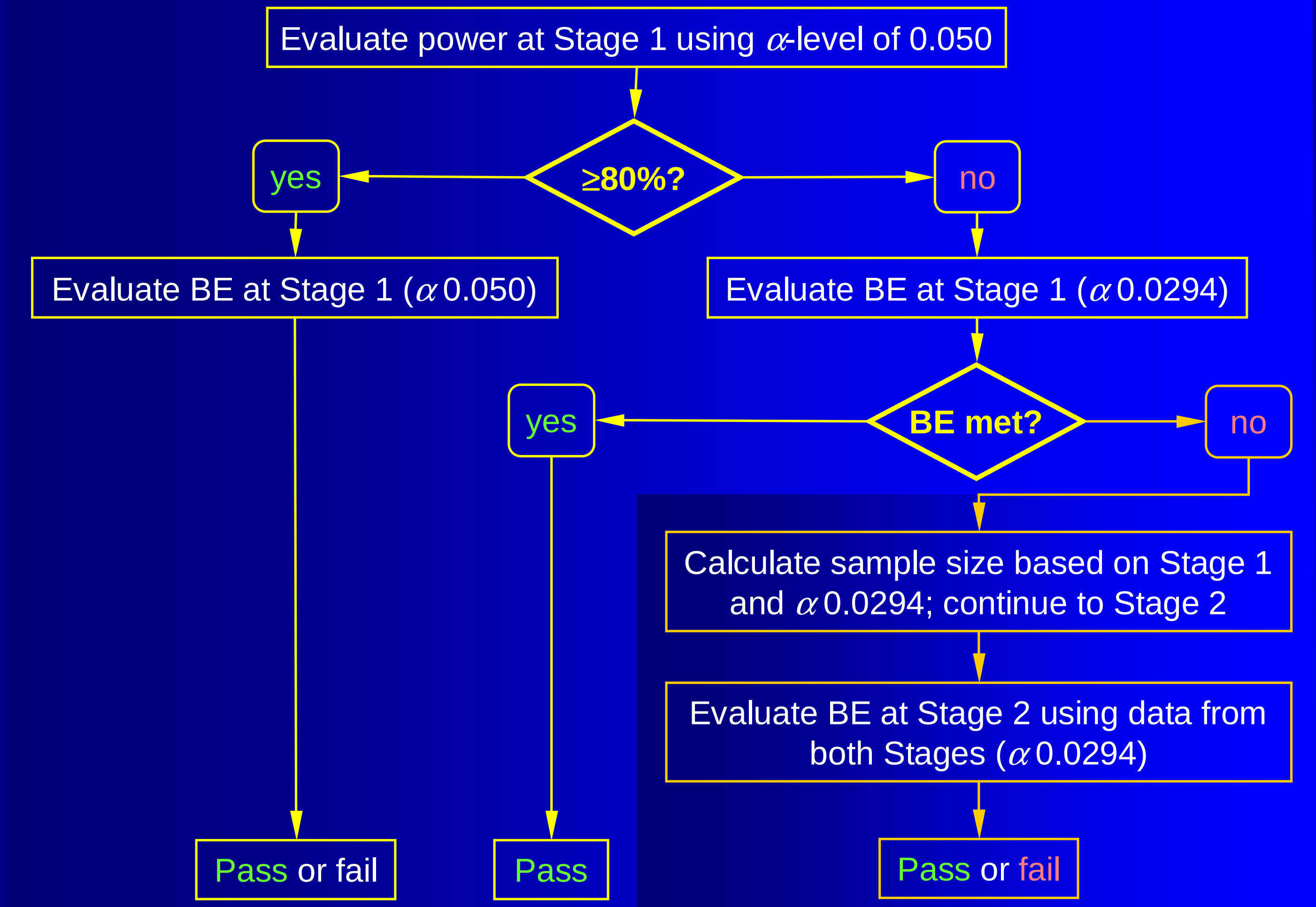
AL Gould
Group Sequential Extension of a Standard Bioequivalence Testing Procedure
J Pharmacokin Biopharm 23/1, 57–86 (1995)

Sequential Designs

- Methods by Potvin *et al.* (2008) promising
  - Supported by 'The Product Quality Research Institute' (members: FDA/CDER, Health Canada, USP, AAPS, PhRMA, …)
    - Acceptable by US-FDA
    - Canada? Or Gould (1995) mandatory?
    - 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, 245–62 (2008), DOI: 10.1002/pst.294
http://www3.interscience.wiley.com/cgi-bin/abstract/115805765/ABSTRACT

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

Calculate sample size based on Stage 1 and α 0.0294; continue to Stage 2

Evaluate BE at Stage 2 using data from both Stages (α 0.0294)

Pass or fail

Potvin *et al.* (Method C)

- Technical Aspects
  - Only *one* Interim Analysis (after Stage 1)
  - If possible, use software (too wide step sizes in Diletti's tables), preferrable the exact method (avoid approximations)
  - Should be termed '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

Potvin *et al.* (Method C)

- Technical Aspects (cont'd)
  - No stop criterion (*'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)
  - If power <80% in Stage 1 or in the pooled analysis (data from Stages 1 + 2), Pocock's α 0.0294 is used (*i.e.*, the 1 – 2×α = 94.12% CI is calculated)
  - Overall patient's risk preserved at ~≤0.05

Potvin *et al.* (Method C)

- 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)
```

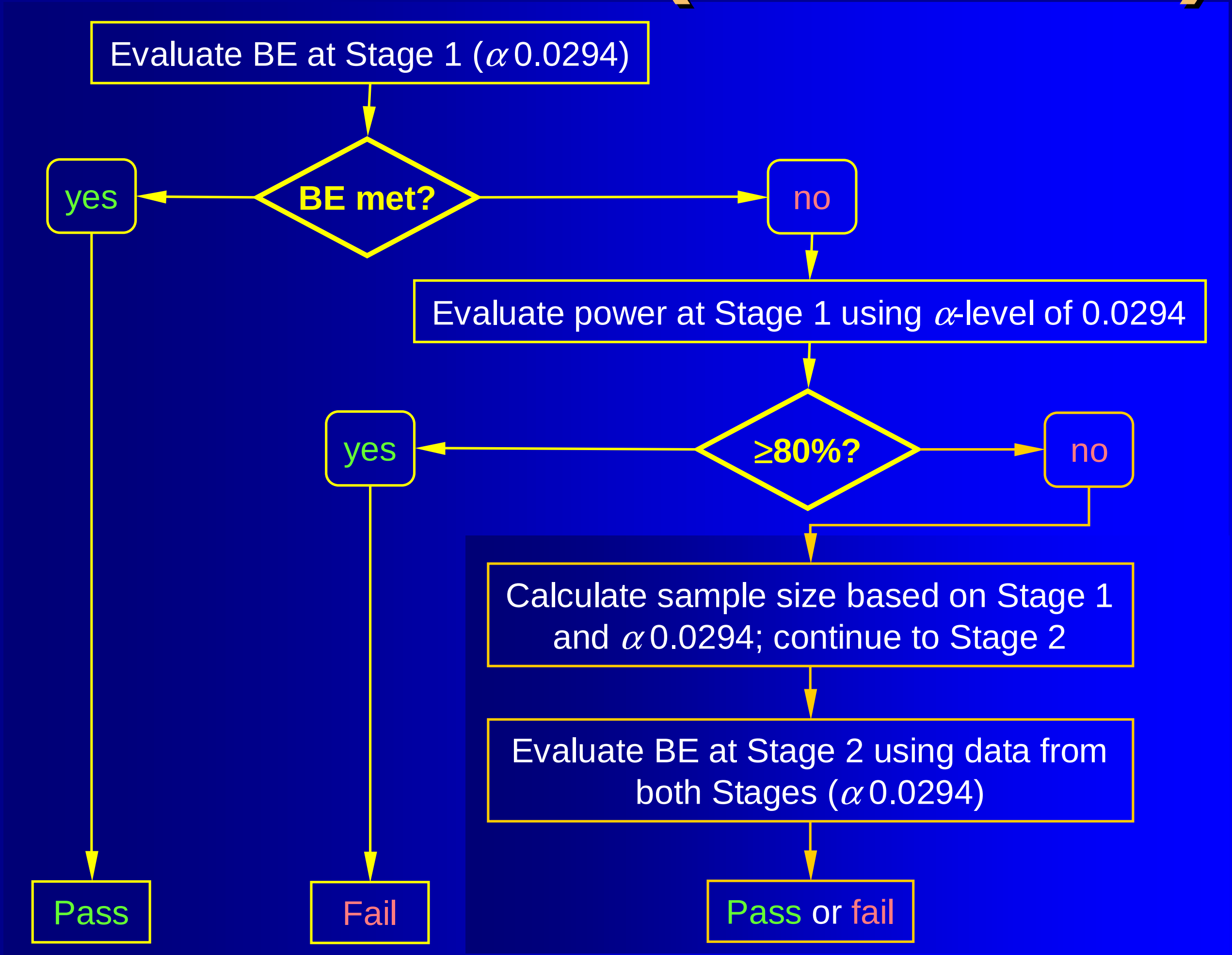
  - No poolability criterion; combining is *always allowed* – even for significant differences between Stages

Potvin *et al.* (Method C)

- Technical Aspects (cont'd)
  - Potvin *et al.* used a simple approximative power estimation based on the shifted *t*-distribution (to increase speed in their simulations?)
  - If possible use the exact method (Owen; package *PowerTOST* exact = TRUE) or at least the one based on the noncentral *t*-distribution (*PowerTOST* exact = FALSE)
  - Power obtained in Stage 1:

method	power
approx. (shifted *t*)	64.94%
approx. (noncentral *t*)	66.45%
exact	66.47%

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

Calculate sample size based on Stage 1 and α 0.0294; continue to Stage 2

Evaluate BE at Stage 2 using data from both Stages (α 0.0294)

Pass

Fail

Pass or fail

Potvin *et al.* (example B/C)

```
Model Specification and User Settings
       Dependent variable : Response
                Transform : LN
              Fixed terms : int+Sequence+Treatment+Period
    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  90% = (   95.1474,   124.3162)
    CI User = (   92.9291,   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 (if power <80%)

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

Potvin *et al.* (example B/C)

```
require(PowerTOST)
power.TOST(alpha=0.05, logscale=TRUE,
           theta1=0.8, theta2=1.25, theta0=0.95,
           CV=0.182132, n=12,
           design = "2x2", exact = TRUE)
```

α 0.05 (C), α 0.0294 (B), expected ratio 95% – *not* 108.76% obs. in stage 1! CV_{intra} 18.2%, 12 subjects in Stage 1

```
[1] 0.6646934
```

Power 66.5% – 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", exact = TRUE,
            print = TRUE)
```

Calculate total sample size: expected ratio 95%, CV_{intra} 18.2%, 80% power

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

Total sample size 20: include another 8 for Stage 2

Potvin *et al.* (example B/C)

8 subjects in Stage 2 (20 total), modified model for 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.* (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 goes 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 (94.12% CI) or stage 2 as the 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 ~20% higher.

Sequential Designs

- Methods by Potvin *et al.* (2008) limited to point estimate of 0.95 and 80% power
  - Follow-up paper
    - Slight inflation of patient's risk (α 0.0547) observed in Methods B/C if PE 0.90 instead of 0.95 was used
    - Method D (like C, but α 0.0280 instead of α 0.0294)
    - Might be usefull if PE 0.95 and power 90% as well; *not validated yet!*

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

Sequential Designs

- Caveats
  - Methods for ‘classical’ group-sequential designs derived based on
    - Test for differences (superiority, parallel groups)
    - Large samples (Z test of normal distributed data with known variance)
    - Fixed total sample size (interim analysis at N/k)
    - Balanced case (no drop outs)
  - Don’t apply any published procedure unquestioned (*i.e.*, if not validated for bioequivalence)
  - *Simulations mandatory* to derive an empirical α (≤0.052)!

Open Issues

- Feasibility / futility rules
  - It would be desirable to stop a study after stage 1 under certain circumstances
    1. (1) BE is unlikely to be shown in even very high sample sizes (*e.g.*, CI outside acceptance range) → reformulate
    2. (2) It turns out that the drug/formulation is highly variable → replicate design study in order to perform scaling required
    3. (3) The calculated sample size exceeds the budget of the project by far

Open Issues

- Feasibility / futility rules
  - These points are not covered by Potvin *et al.*
  - If you decide to include a rule for early stopping, it's not part of the statistical procedure any more
  - (1) and (2) are ethically justifiable
  - (3) Acceptance?

Open Issues

- Arbitrary PE and/or power
  - Simulations mandatory
    - Set desired PE and power
    - Define maximum α-inflation (≤0.052?)
    - Simulate sufficiently large number of studies (N)
      - Count number of studies accepted BE at 1.25 (n_1) and number of studies rejected BE at the desired PE (n_2)
      - Empirical $\alpha = n_1/N$
      - Empirical $\beta = n_2/N$; power = $1 - \beta$
    - Start with Pocock's nominal α 0.0294 and decrease stepwise if empirical α too high
    - Compiled language almost necessary (speed!)

Open Issues

- Adaption for stage 1 PE (full adaptive design)
  - If applied naïvely, α-inflation of up to 30%!*
  - Various methods for superiority trials, but nothing in the area of BE published
  - Simulations mandatory

* **Cui L, Hung MJ, and S-J Wang**
Modification of sample size in group sequential clinical trials
Biometrics 55, 853–7 (1999)

Open Issues

- Dropping a candidate formulation from a higher-order cross-over design

Stage 1		
I	II	III
T_1	T_2	R
T_2	R	T_1
R	T_1	T_2
T_1	R	T_2
T_2	T_1	R
R	T_2	T_1
…	…	…

Stage 2	
I	II
R	T_2
T_2	R
…	…

How to decide *which* formulation to drop?

- Statistical model of BE assumes IID (common σ^2)
  - Let’s assume to continue with T_2
  - If $\sigma^2_{T1} > \sigma^2_{T2}$ and/or σ^2_{R}, the pooled variance in Stage 1 will be inflated. The estimated total sample size will be too high. Expensive, but no influence on α expected.
  - If $\sigma^2_{T1} < \sigma^2_{T2}$ and/or σ^2_{R}, power will be lower – increasing the producer’s risk only.

Don’t try this at home!

- Data of 6×3 dose proportionality study
  R 20 mg, T_1 30 mg, T_2 40 mg; CV_{intra} 8.76%
  - ⅔T_1, ¾T_2: fixed effects (EMA), Method D_B, PE 90%, α 0.028

Stage 1					
I		II		III	
R	146.05	T_2	133.26	T_1	269.51
T_1	86.83	R	55.08	T_2	52.58
T_2	52.78	T_1	75.51	R	60.57
T_1	99.57	T_2	57.29	R	74.45
T_2	80.61	R	94.62	T_1	121.39
R	57.10	T_1	80.58	T_2	52.08
T_1	109.79	R	59.20	T_2	55.99
T_2	44.07	T_1	79.76	R	57.25

Stage 2			
I		II	
R	74.45	T_2	61.72
T_2	54.93	R	54.71
T_2	43.17	R	37.49
R	54.64	T_2	47.32

Extremely imbalanced due to arbitrary ‘cut’ of original dataset! N=6 (single balanced block) would have zero df for sequences.

Don’t try this at home!

8 subjects in Stage 1, all effects fixed (EMA)

```
Model Specification and User Settings
       Dependent variable : Response
                Transform : LN
              Fixed terms : int+sequence+treatment+period+subject(sequence)

Final variance parameter estimates:
            Var(Residual)    0.0068489

Bioequivalence Statistics
User-Specified Confidence Level for CI's = 94.4000
Percent of Reference to Detect for 2-1 Tests = 20.0%
A.H.Lower =  0.800   A.H.Upper =  1.250
Reference: Reference   LSMean=  4.332414  SE=  0.029948  GeoLSM=  76.127859
---------------------------------------------------------------------------
Test:      Test 1      LSMean=  4.726674  SE=  0.029948  GeoLSM= 112.919400
    Difference =      0.3943,  Diff_SE=   0.0417,  df= 12.0
    Ratio(%Ref) =   148.3286
    CI User = (135.8004, 162.0127)
    Average bioINequivalence shown for confidence=94.40 and percent=20.0.
---------------------------------------------------------------------------
Test:      Test 2      LSMean=  4.187643  SE=  0.029948  GeoLSM=  65.867359
    Difference =    -0.1448,   Diff_SE=   0.0417,  df= 12.0
    Ratio(%Ref) =   86.5220
    CI User = ( 79.2141,  94.5041)
    Failed to show average bioequivalence for confidence=94.40 and percent=20.0.
```

CV_{intra} 8.29%

α 0.028 (Method D/B)

Don’t try this at home!

```
require(PowerTOST)
power.TOST(alpha=0.0280, logscale=TRUE,
           theta1=0.8, theta2=1.25, theta0=0.90,
           CV=se2CV(sqrt(0.0068489)), n=8,
           design="3x6x3", exact=TRUE)
```

α 0.028, expected ratio 90%, MSE 0.06849 (CV_{intra} 8.29%), 8 subjects in Stage 1, 6×3 design

```
[1] 0.762231
```

Power 76.2% <80% – initiate Stage 2

```
sampleN.TOST(alpha=0.0280, targetpower=0.80, logscale=TRUE,
            theta1=0.8, theta2=1.25, theta0=0.90,
            CV=se2CV(sqrt(0.0068489)), design="3x6x3", exact=TRUE,
            print=TRUE)
```

Calculate total sample size: expected ratio 90%, CV_{intra} 8.29%, 80% power, keeping 6×3 design

```
+++++++++++ Equivalence test - TOST +++++++++++
            Sample size estimation
-----------------------------------------------
Study design:  3x6x3 crossover
log-transformed data (multiplicative model)

alpha = 0.0294, target power = 0.8
BE margins        = 0.8 ... 1.25
Null (true) ratio = 0.9,  CV = 0.0829

Sample size
 n     power
12   0.920990
```

Total sample size 12: include another 4 for Stage 2

Don’t try this at home!

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

Final variance parameter estimates:
            Var(Residual)    0.00667999

Bioequivalence Statistics
User-Specified Confidence Level for CI's = 94.4000
Percent of Reference to Detect for 2-1 Tests = 20.0%
A.H.Lower =  0.800   A.H.Upper =  1.250
Reference: Reference   LSMean=  4.045115  SE=  0.103862  GeoLSM=  57.117740
---------------------------------------------------------------------------
Test:      Test 1      LSMean=  4.455914  SE=  0.106556  GeoLSM=  86.134878
    Difference =      0.4108,  Diff_SE=   0.0394,  df= 14.985
    Ratio(%Ref) =   150.8023
    CI User = (138.9762, 163.6348)
    Average bioINequivalence shown for confidence=94.40 and percent=20.0.
---------------------------------------------------------------------------
Test:      Test 2      LSMean=  3.933423  SE=  0.103862  GeoLSM=  51.081521
    Difference =     -0.1117,  Diff_SE=   0.0335,  df= 14.985
    Ratio(%Ref) =    89.4320
    CI User = ( 83.4279,  95.8682)
    Average bioequivalence shown for confidence=94.40 and percent=20.0.
```

4 subjects in Stage 2 (12 total), modified model for pooled analysis

Don’t try this at home!

- Lessons learned, open questions
  - Not validated! Don’t think about using it at all!
  - Note that due to the massive imbalance the LSM of Test 1 (although not included in Stage 2) changed from Stage 1 in the pooled analysis!
    - Stage 1: 112.92
    - Pooled: 86.13
  - Drug has low CV$_{intra}$, but high CV$_{inter}$ – Apples and oranges?

CV%	T_1	T_2	R	model
Stage 1	26.86	34.15	37.32	period
Stage 2	–	18.08	24.79	period
Pooled	26.86	32.01	35.92	period

Don’t try this at home!

- Lessons learned, open questions
  - Must use software in the power calculation which can handle the degrees of freedom of a Williams’ design in Stage 1 correctly (*e.g.*, *PowerTOST*)
  - Obvious which formulation to drop in this example, but what if formulations are similar in PEs?
    Keep the one with smaller CV_{inter}?
  - Design in the sample size estimation of Stage 2?
    - 3×6 (block size 6 → 12)
    - 2×2 (block size 2 → 10)
    - The latter would have failed in the example

Don’t try this at home!

- Lessons learned, open questions
  - Tempting idea, but not recommended
    - until a statistical decision tree is developed and
    - suitable simulations have shown that the patient’s risk is not inflated

Open Issues

- Replicated designs (HVDs/HVDPs)
  - Nothing published!
  - Statistical model?
  - Although EMA assumes equal variances of formulations (Q&A document Jan 2010) that does not reflect the 'real world' (quite often $\sigma^2_{WR} > \sigma^2_{WT}$)
  - If you set up simulations allow for different variances of test and reference

Congratulations!

Power and intra-subject variability in 2 stage approaches to BE approval

Open Questions?

Helmut Schütz

BEBAC

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

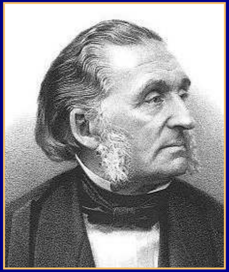
To bear in Remembrance...

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