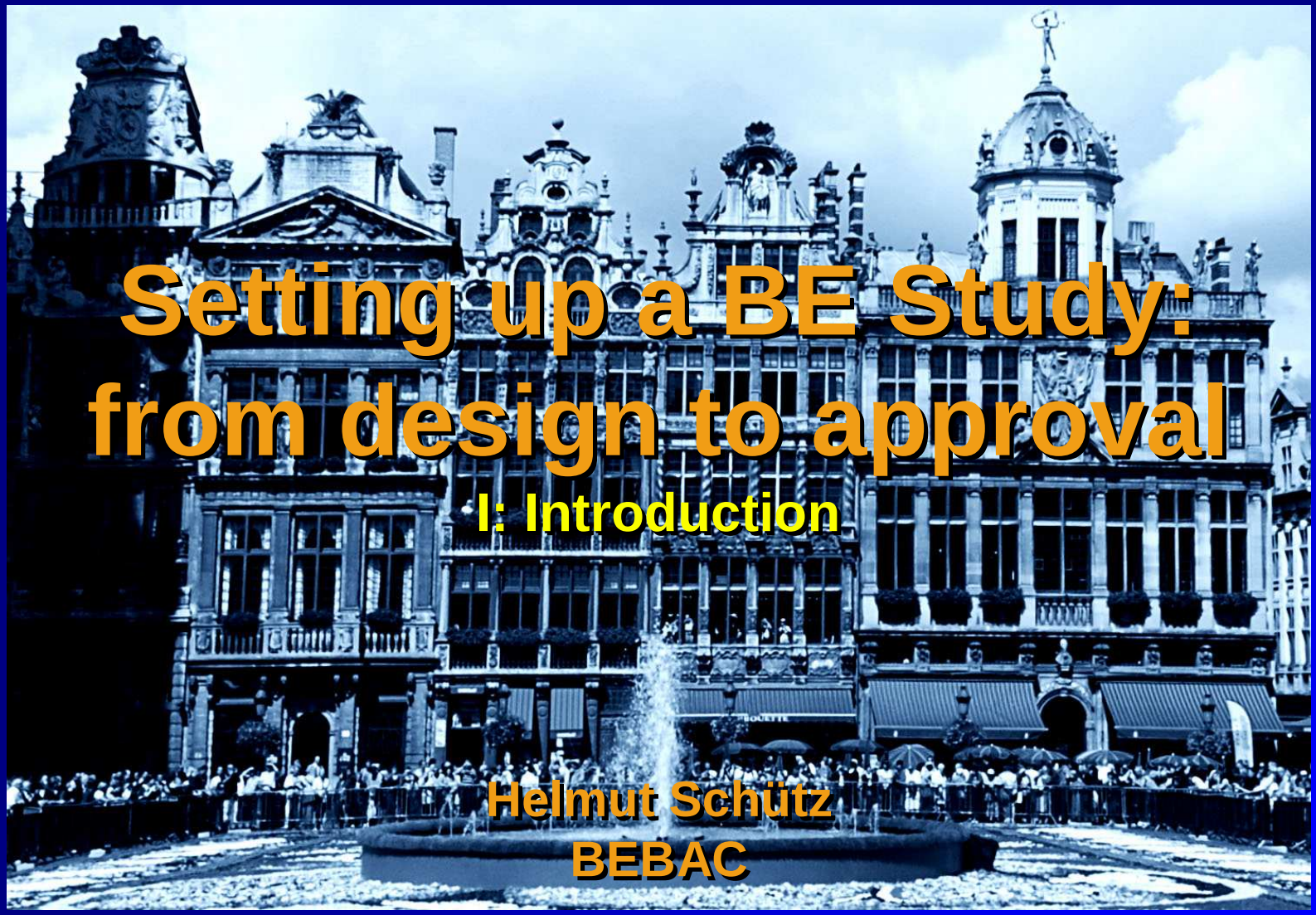


Setting up a BE Study: from design to approval

I: Introduction

Helmut Schütz
BEBAC

To bear in Remembrance...

Whenever a theory appears to you as the only possible one, take this as a sign that you have neither understood the theory nor the problem which it was intended to solve.

Karl R. Popper

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

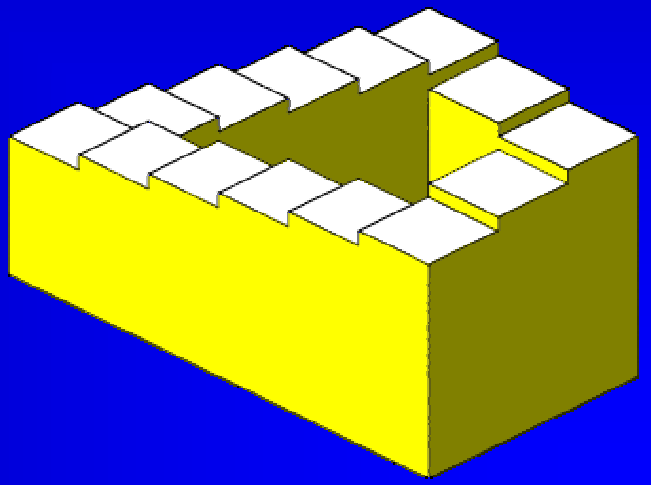
Leslie Z. Benet

Statistics* – A subject which most statisticians find difficult but in which nearly all physicians are expert.** ***Stephen Senn

Bioequivalence Studies

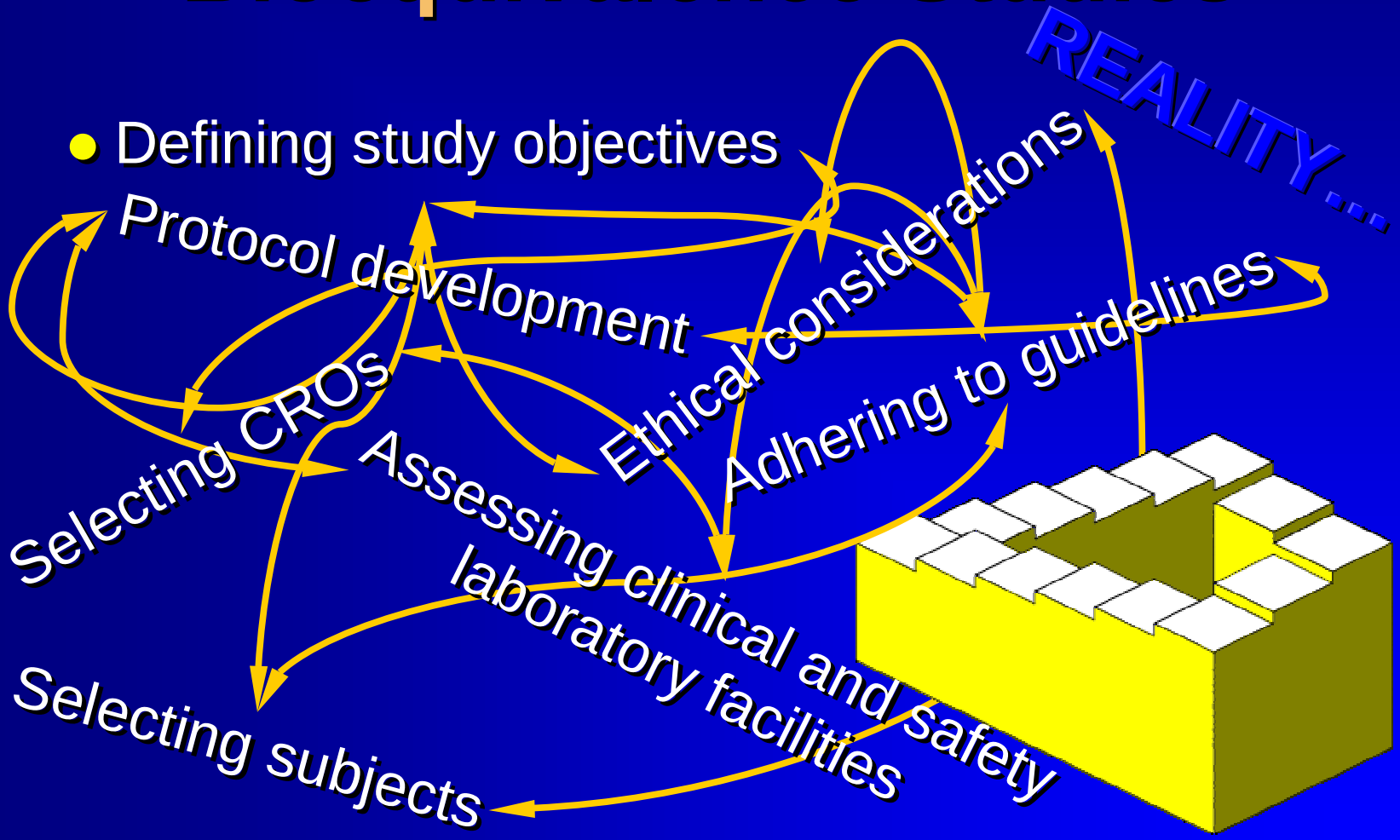
DREAM...

- Defining study objectives
- Selecting CROs
- Protocol development
- Ethical considerations
- Assessing clinical and safety laboratory facilities
- Selecting subjects
- Adhering to guidelines



Bioequivalence Studies

- Defining study objectives



Overview

- Bioequivalence
  - Surrogate of clinical equivalence or
  - Measure of pharmaceutical quality?
- Types of studies
  - Pharmacokinetic (PK)
  - Pharmacodynamic (PD)
  - Clinical (equivalence and/or safety/efficacy)
  - Healthy Subjects *vs.* patients
  - Single dose *vs.* multiple dose
  - Parallel / cross-over / replicate

Overview

- Types of studies (cont'd)
  - Food effect
  - PK interaction
- Design Issues
  - Reference product / batch, dose regimen
  - Fasted / fed state
  - Standardization
- Bioanalytics (*not* GLP!)
  - Parent drug / metabolite(s) / enantiomers / pro-drugs
  - Validation / routine application

Overview

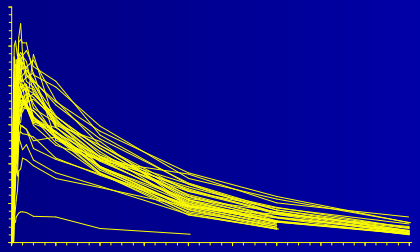
- Ethics (GCP!)
  - Dose levels / number of administered doses
  - Number / volume of blood samples
  - Drug and/or adverse effects
- Clinical performance (GCP!)
  - CRO selection
  - Responsibilities of sponsor / investigator
  - Audits / monitoring

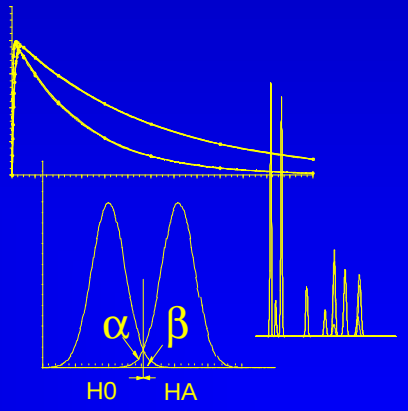
Overview

- NCA / PK (PD)
  - Sampling schedule
  - Metrics (AUC, C_{max}; $AUEC$, Ae_{max}, …)
  - Design, methods, evaluation
- Sample size
  - Estimation from previous and/or pilot studies, literature
  - Two-stage designs, scaling appoaches (HVDs)
- Biostatistics
  - Models & assumptions
  - Protocol, evaluation, report

Assumptions

World *'Truth'*

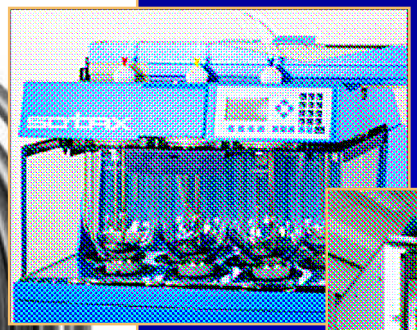


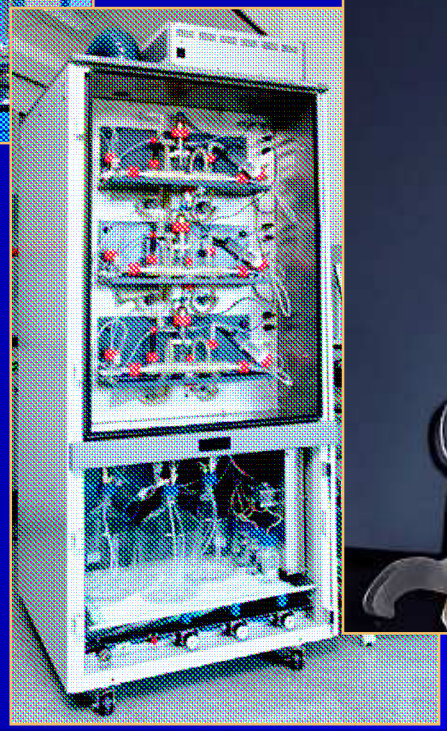


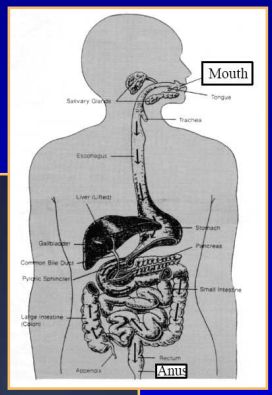
Model *'Data'*

Theory *'Reality'*

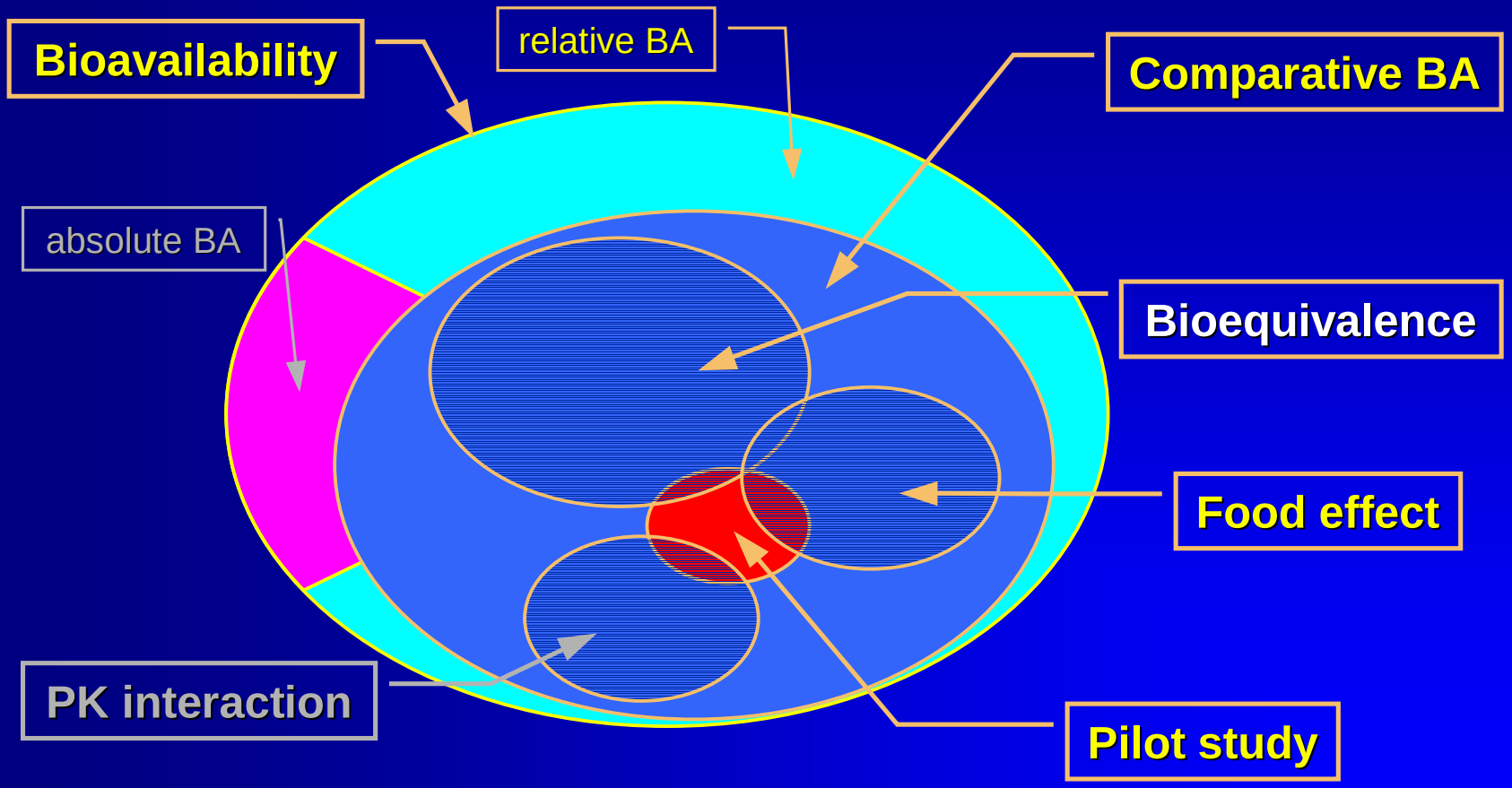
Models *vs.* Reality







Terminology



Definition

- According to ‘old’ EU NfG (3. Design and Conduct of Studies, paragraph 2):

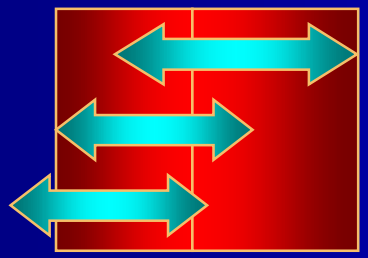
  A bioequivalence study is basically a comparative bioavailability study designed to establish equivalence between test and reference products.

  - Comparative BA,
  - designed to demonstrate BE,
  - reference = innovator’s product.

EMEA Human Medicines Evaluation Unit / CPMP
Note for Guidance on the Investigation of Bioavailability and Bioequivalence (2001)
http://bebac.at/downloads/140198enfin.pdf

Bioequivalence…

- Comparative BA
  - True experiment; no bibliographic comparison
- Designed to demonstrate BE
  - Variability,
  - Deviation of test from reference,
  - Drop-out rate, …
    - ➔to be able (statistical power!) to demonstrate BE
- Reference = Innovator's product



#1: BE [90%–125%]

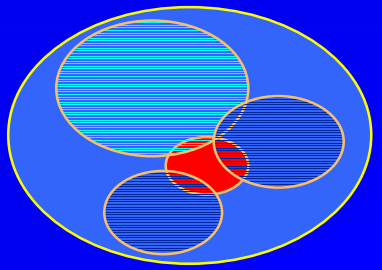
#2: BE [80%–110%]

#3: not BE [76%–103%]; (but 'BE' to #2)

Bioequivalence…

- EMA GL on BE (2010)

  Two medicinal products containing the same active substance are considered bioequivalent if they are pharmaceutically equivalent or pharmaceutical alternatives and their bioavail-abilities (rate and extent) after administration in the same molar dose lie within acceptable predefined limits. These limits are set to ensure comparable in vivo performance, i.e. similarity in terms of safety and efficacy.



Global Harmonization?

- In almost all regulations two metrics are necessary to demonstrate BE, namely
  - extent (AUC_t or AUC_∞) and
  - rate (C_{max}) of exposure.
- One exception: US-FDA (where AUC_t and AUC_∞ must demonstrate extent of exposure)
  - Although stated in the GL, such a requirement is statistically flawed.
    - Multiplicity issues (what is the patient's risk?)
    - Impossible α-adjustment (interdependence)

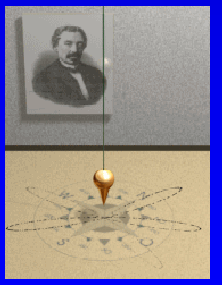
There can be only one!

History of BE

- Bioequivalence
  - Problems first noticed with NTIDs (Narrow Therapeutic Index Drugs) in the late 1970s
  - Intoxications (and even some fatallities!) were reported (warfarin, digoxin, phenytoin)
    - Warfarin, digoxin: Patients switched between formulations which were got approval solely based on *in vitro* data (innovator ↔ generic)
    - Phenytoin: The innovator's API was changed from a microcrystalline to an amorphous form resulting in 10× higher plasma concentrations in steady state

History of BE

- Bioequivalence
  - Surrogate of clinical equivalence (1980+)
    - Studies in steady state in order to reduce variability
    - Studies based on active metabolite
    - Wider acceptance range if clinical justifiable (not FDA!)
  - Measure of pharmaceutical quality (2000+)
    - Single dose studies preferred
    - Generally parent drug
    - Widening of acceptance range exceptional (except FDA HVDs and EMA C_{max} of HVDs)



Early 1980s

- First method
  - FDA's 75/75 Rule
    BE, if 75% of subjects show ratios of 75%-125%. Not a statistic, variable formulations may pass by chance…

	T	R	T/R	75%-125%
1	71	81	87.7%	yes
2	61	65	93.8%	yes
3	80	94	85.1%	yes
4	66	74	89.2%	yes
5	94	54	174.1%	no
6	97	63	154.0%	no
7	70	85	82.4%	yes
8	76	90	84.4%	yes
9	54	53	101.9%	yes
10	99	56	176.8%	no
11	83	90	92.2%	yes
12	51	68	75.0%	yes
				75.0%

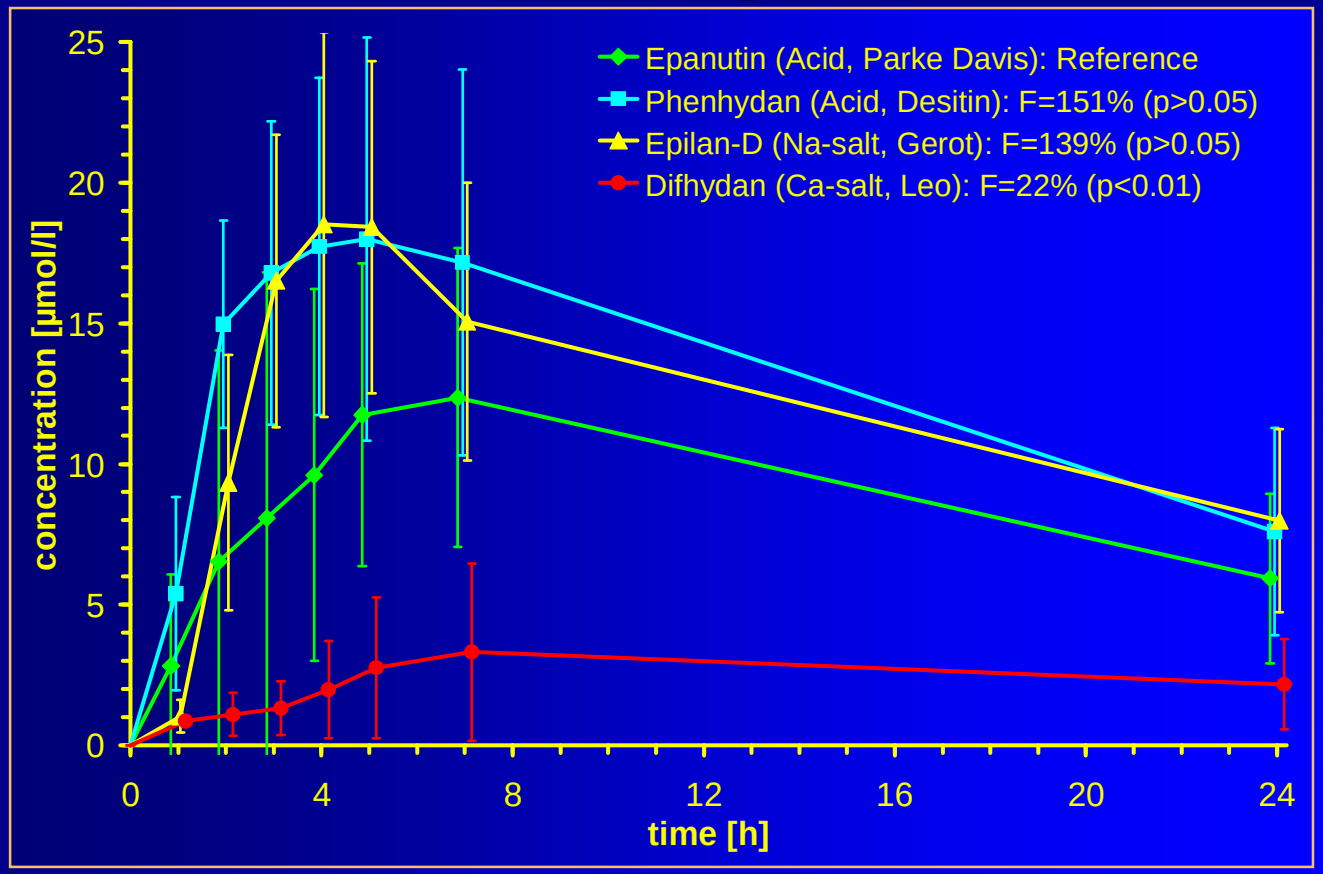
BE Cabana
Assessment of 75/75 Rule: FDA Viewpoint
J Pharm Sci 72, 98-99 (1983)
JD Haynes
FDA 75/75 Rule: A Response
J Pharm Sci 72, 99-100 (1983)

Mid 1980s I

- Early method
  - Testing for a significant difference (*t*-test) at α 0.05
    Problem:
    - High variability in differences
      $\rightarrow$ formulation will pass ($p \geq 0.05$)
    - Low variability in differences
      $\rightarrow$ formulation will fail ($p < 0.05$)
    - This is counterintuitive and the opposite of what we actually want!

	T	R	T–R
1	71	81	-10
2	61	65	-4
3	80	94	-14
4	66	74	-8
5	94	54	+40
6	97	63	+34
7	70	85	-15
8	76	90	-14
9	54	53	+1
10	99	56	+43
11	83	90	-7
12	51	68	-17
mean	75	73	+2
SD	16	15	23
CV%	21.4%	20.6%	940%
		t-table	2.2010
		t-calc	0.3687
			n.s.

Example



Nitsche V, Mascher H, and H Schütz
Comparative bioavailability of several phenytoin preparations marketed in Austria
Int J Clin Pharmacol Ther Toxicol 22(2), 104-107 (1984)

Mid 1980s II

- Later method
  - FDA's 80/20 rule
  - At least 80% power to be able to demonstrate a 20% difference (*t*-test) at α 0.05
    - Essentially the 75/75 rule in more statistical terms.
    - Power 71.5% < 80! (not BE)
    - In any study (even at 'true' T=R) with variability

      $s\sqrt{2/n} > 6.44$

      it is impossible to show BE!

	T	R	T–R
1	71	81	-10
2	61	65	-4
3	80	94	-14
4	66	74	-8
5	94	54	+40
6	97	63	+34
7	70	85	-15
8	76	90	-14
9	54	53	+1
10	99	56	+43
11	83	90	-7
12	51	68	-17
mean	75	73	+2
SD	16	15	23
		t-table	2.2010
		t-calc	0.3687
			n.s.
		power	71.59%

Late 1980s

- TOST (Two One-Sided Tests)
  - First formulation of the problem based on equivalence rather than a difference
    - Two One-Sided *t*-tests
    - Bioequivalent if
      $p(<80\%) + p(>120\%) \leq 0.05$
    - Equivalent to a 90% confidence interval within an acceptance range of 80% – 120%

	T	R	T–R
1	71	81	-10
2	61	65	-4
3	80	94	-14
4	66	74	-8
5	94	54	+40
6	97	63	+34
7	70	85	-15
8	76	90	-14
9	54	53	+1
10	99	56	+43
11	83	90	-7
12	51	68	-17
	p(<80%)		0.0069
	p(>120%)		0.0344
	p(total)		0.0414
		T/R	103.32%
	90% CI (lo)		88.35%
	90% CI (hi)		118.30%

DA Schuirmann
A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability
J Pharmacokin Biopharm 15, 657–680 (1987)

Excursion: 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

- … in more statistical terms:

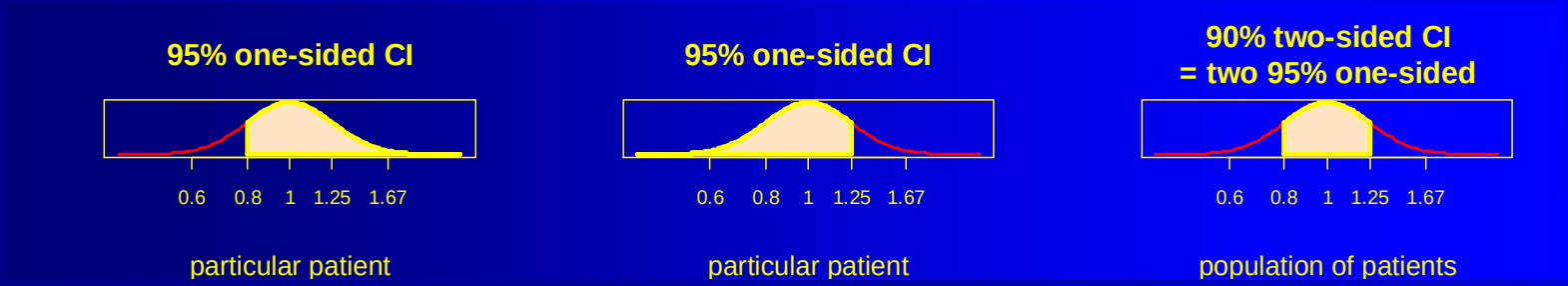
Decision	Null hypothesis true	Null hypothesis false
Null hypothesis rejected	**Error type I**	**Correct (H_a)**
Failed to reject null hypothesis	**Correct (H_0)**	**Error type II**

- 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%) is:
    90% CI = 1 – 2×α = 0.90



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

Human Guineapigs I

- BE studies as a surrogate for clinical efficacy / safety ('essential similarity')
  - We want to get unbiased estimates, *i.e.*, the point estimate from the study sample …

$$PE = \frac{\hat{X}_{Test}}{\hat{X}_{Reference}}$$

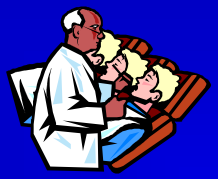
  - … should be representative for the population of patients.

$$F_{Pop} = \frac{\mu_{Test}}{\mu_{Reference}}$$

Human Guineapigs II

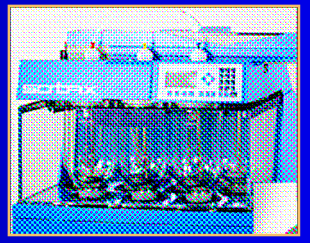
- BE studies as a special case of documented pharmaceutical quality
  - The *in vivo* release in the biostudy …

$$PE = \frac{\hat{X}_{Test}}{\hat{X}_{Reference}}$$



  - … should be representative for the *in vitro* performance.

$$f_2 = 50 \cdot \log\left[\frac{100}{\sqrt{1+\frac{\sum_{t=1}^{t=n}\left[\bar{R}(t)-\bar{T}(t)\right]^2}{n}}}\right]$$



Science → Regulations

- We can't compare bioavailabilities in the entire population of patients
  - Scientific Reductionism (based on assumptions)
    - 'Similar' concentrations in healthy subjects will lead to 'similar' effects in patients.
    - Equal doses and inter-occasion clearances!

$$AUC_T = \frac{D_T \cdot F_T}{CL_T}, AUC_R = \frac{D_R \cdot F_R}{CL_R}$$

$$[D_T \cong D_R, CL_T \cong CL_R]$$

$$F_{rel}(BA) = \frac{F_T}{F_R} \cong \frac{AUC_T}{AUC_R}$$

Another reminder

Rose
is a rose
is a rose
is a rose.

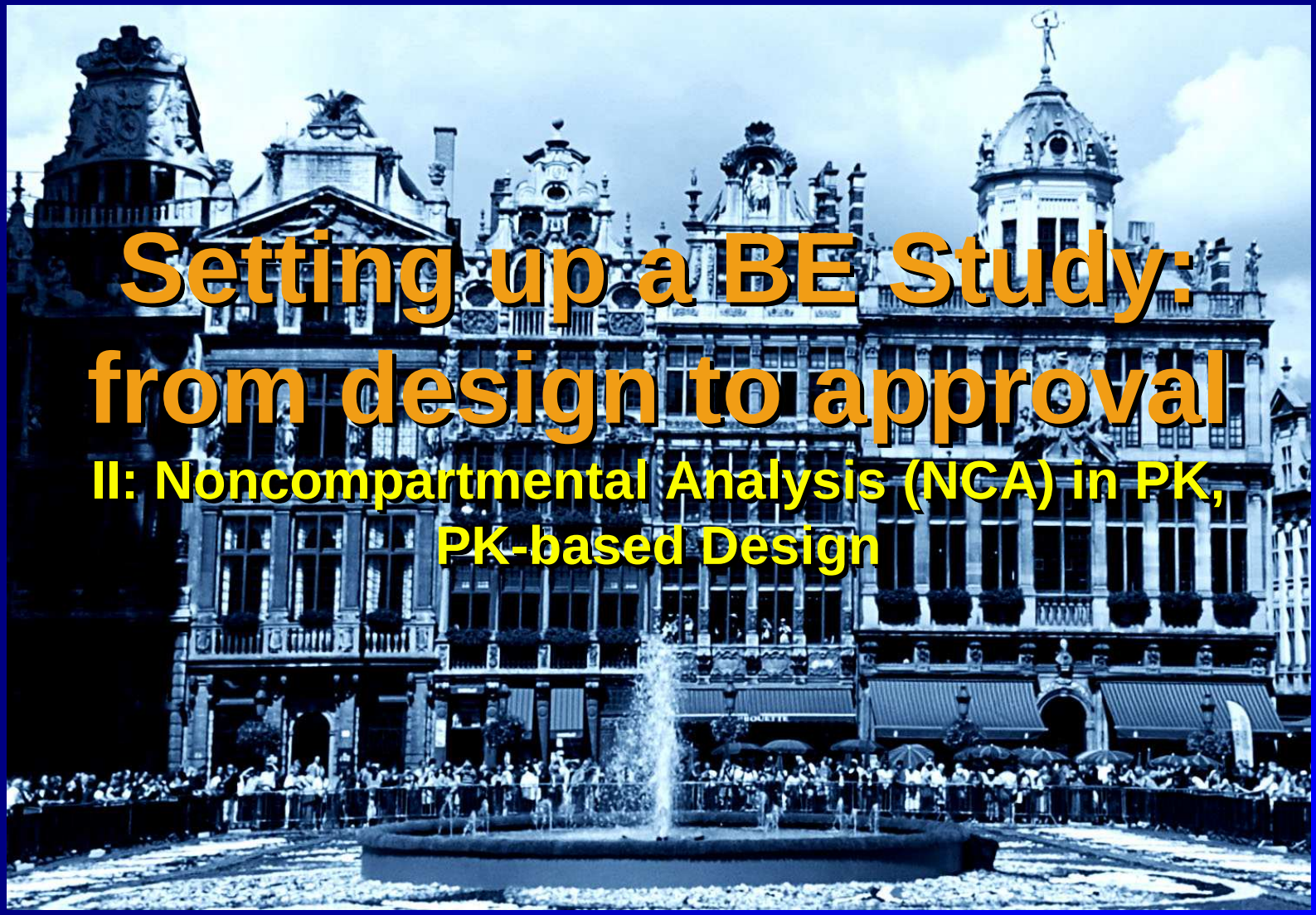
Gertrude Stein (1913)

Guidelines
are guidelines
are guidelines.

Henrike Potthast (ca. 2004)

No one wants to learn from mistakes,
but we cannot learn enough from successes
to go beyond the state of the art.

Henry Petroski



Setting up a BE Study: from design to approval

II: Noncompartmental Analysis (NCA) in PK, PK-based Design

NCA *vs.* PK Modeling

- Noncompartmental methods do not rely on a pharmacokinetic (=compartmental) model
- Also called SHAM (Shape, Height, Area, Moments)
  - Metrics (plasma)
    - Extent of absorption (EU…), total exposure (US): AUC
    - Rate of absorption (EU…), peak exposure (US): C_{max}
    - t_{max} (EU…)
    - Early exposure (US, CAN): $AUC_{t_{max}}$; partial AUC truncated at population (CAN: subject's) t_{max} of the reference
    - Others: C_{min}, Fluctuation, MRT, Occupancy time, t_{lag},…

NCA *vs.* PK Modeling

- Noncompartmental methods (cont'd)
  - Metrics (urine)
    - Extent of absorption (EU…), total exposure (US):
      Ae_t (cumulative amount excreted)
      rarely extrapolated to $t=\infty$
    - Rate of absorption, peak exposure (US):
      ΔAe_{max}, $t\Delta Ae_{max}$
    - EU: C_{max}, t_{max} from plasma!

NCA *vs.* PK Modeling

- Pharmacokinetic models
  - Useful for understanding the drug/formulation
    - Study design of BA/BE!
  - Drawbacks:
    - Almost impossible to validate (fine-tuning of side conditions, weighting schemes, software, …)
    - Still a mixture of art and science.
    - Impossible to recalculate any given dataset using different software – sometimes even different versions of the same software!
    - Not acceptable for evaluation of BA/BE studies!

NCA (Methods)

- Single dose
  - Calculation of Moments of Curve (AUC_t, MRT_t)
    - Linear trapezoidal rule, loglinear trapezoidal rule, or combination (lin-up, log-down).
  - Calculation of half life ($t_{½}$) from elimination rate (λ_z)
    - Unweighted (!) log-linear regression
  - Extrapolation from time point of last quantified concentration to infinity

$$AUC_\infty = AUC_t + \frac{C_t}{\hat{\lambda}_z}$$ or better: $$AUC_\infty = AUC_t + \frac{\hat{C}_t}{\hat{\lambda}_z}$$

  - C_{max} / t_{max} directly from profile

NCA (Methods)

- Single dose
  - Method of estimation of λ_z stated in protocol!
    - One-compartment model: TTT-method *)
      (Two times t_{max} to t_z)
    - Maximum adjusted R^2 (Phoenix/WinNonlin, Kinetica)

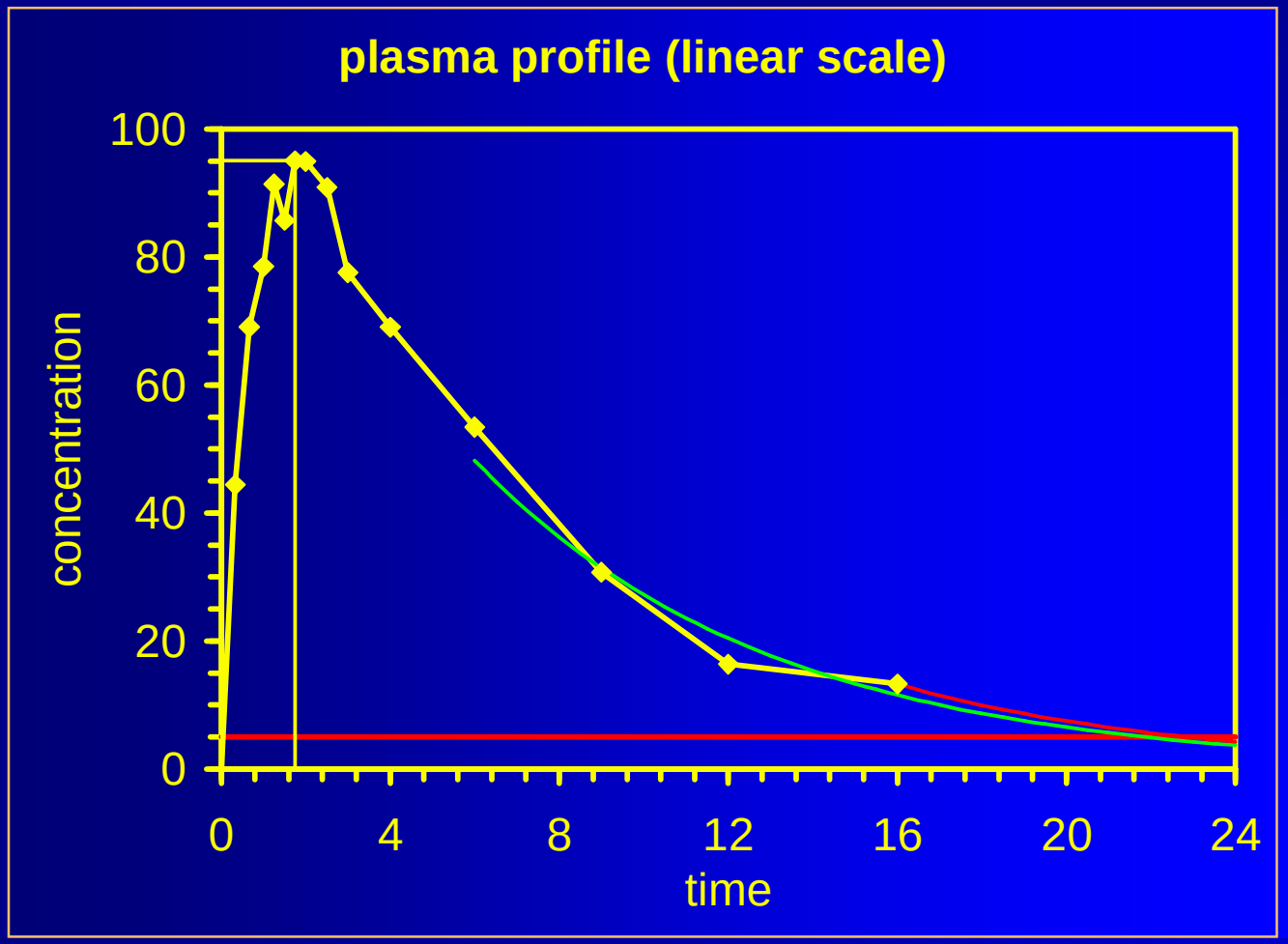
$$R^2_{adj} = 1 - \frac{(1 - R^2) \cdot (n - 1)}{n - 2}$$

WinNonlin ≤5.3: C_{max} **in**cluded
Phoenix/WNL ≥6.0: C_{max} **ex**cluded

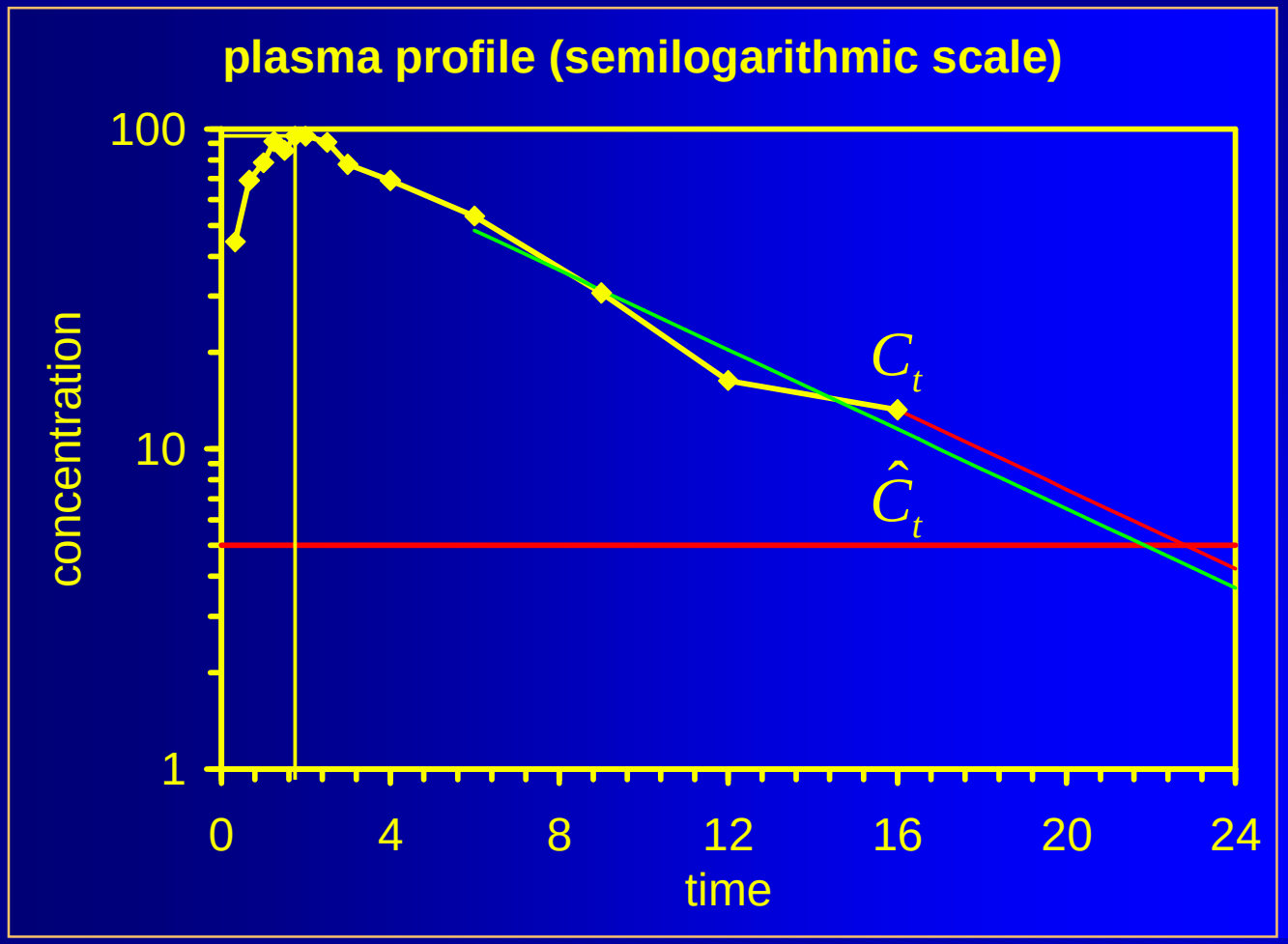
    - Multi-compartment models: starting point = last inflection
    - Minimum AIC $AIC = n \cdot [\ln(2 \cdot \pi) + 1] + n \cdot \ln(RSS/n) + 2 \cdot p$
    - Visual inspection of fit mandatory!

*) **Scheerans C, Derendorf H and C Kloft**
Proposal for a Standardised Identification of the Mono-Exponential Terminal Phase for Orally Administered Drugs
Biopharm Drug Dispos 29, 145–157 (2008)

NCA (Methods)


plasma profile (linear scale)
concentration
100
80
60
40
20
0
0
4
8
12
16
20
24
time

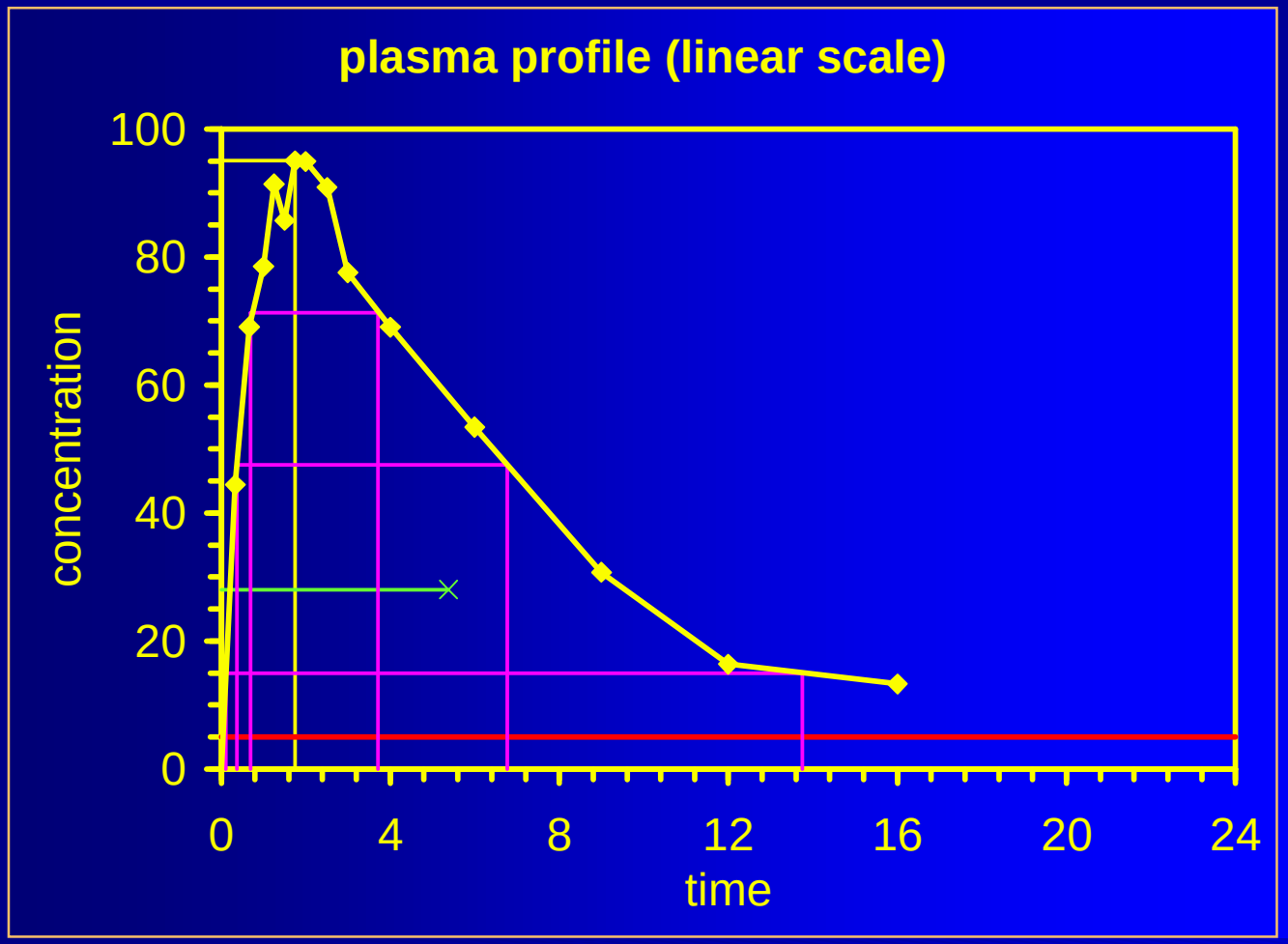
NCA (Methods)


plasma profile (semilogarithmic scale)
100
10
1
concentration
C_t
$\hat{C}_t$
0
4
8
12
16
20
24
time

NCA (Methods)

- Single dose
  - Unconventional parameters describing the shape of the profile
    - C_{max}/AUC
    - HVD (Half value duration: time interval where $C(t) \geq$ 50% of C_{max})
    - $t_{75\%}$ (Plateau time: interval where $C(t) \geq$ 75% of C_{max})
    - Occupancy time, $t \geq MIC$ (time interval where $C(t)$ is above some limiting concentration)

NCA (Methods)


plasma profile (linear scale)
concentration
100
80
60
40
20
0
0
4
8
12
16
20
24
time

NCA (Methods)

- Multiple dose
  - Calculation of AUC_{τ} (dosage interval τ); $AUC_{ss,24h}$ if more than *o.a.d.* and chronopharmaco-logical variation)
  - No extrapolation!
  - $C_{ss,max}$ directly from profile
  - $C_{ss,min}$ from profile *or* (if missing values / time dev's)

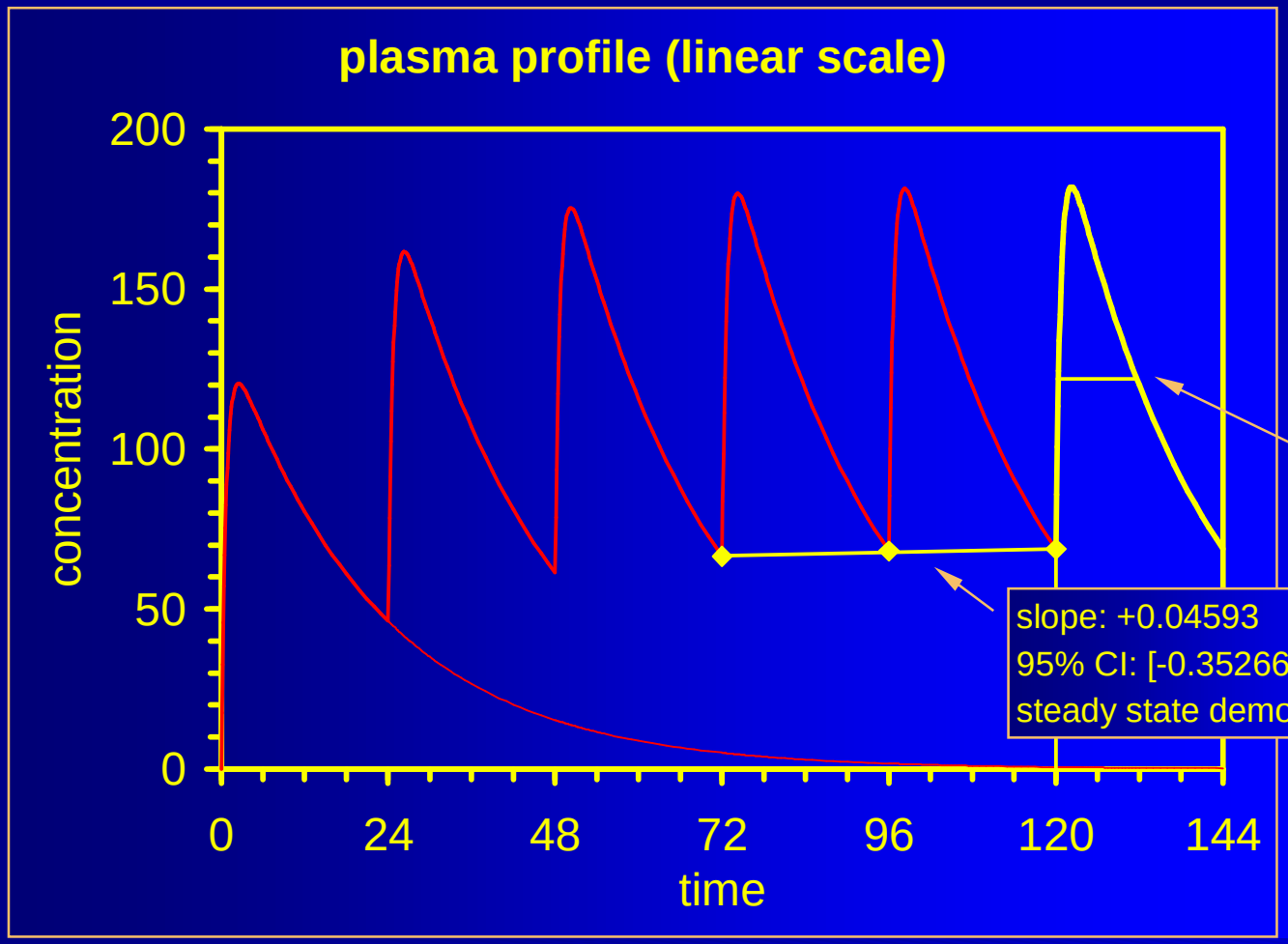
$$\hat{C}_{ss,min} = C_z e^{-\hat{\lambda}_z(\tau - t_z)}$$

  - Peak-Trough-Fluctuation $(C_{ss,max} - C_{ss,min}) / C_{ss,av}$, where $C_{ss,av} = AUC_{\tau} / \tau$
  - Swing $(C_{ss,max} - C_{ss,min}) / C_{ss,min}$

NCA (Methods)

- Multiple dose
  - Assessment whether steady state is reached (in a linear PK system: $AUC_{\tau} = AUC_{\infty}$)
    - No recommendations in GLs (except EU/US Veterinary)
    - Not required according to comments to EMA BE-GL
    - MANOVA-model (sometimes mentioned in Canada, rarely used)
    - t-test of last two pre-dose concentrations
    - Hotelling's T^2
    - Linear regression of last three pre-dose concentrations, individually for each subject/treatment
  - Only the last method allows the exclusion of subjects being not in stead state. Other methods give only a yes|no result!

NCA (Methods)

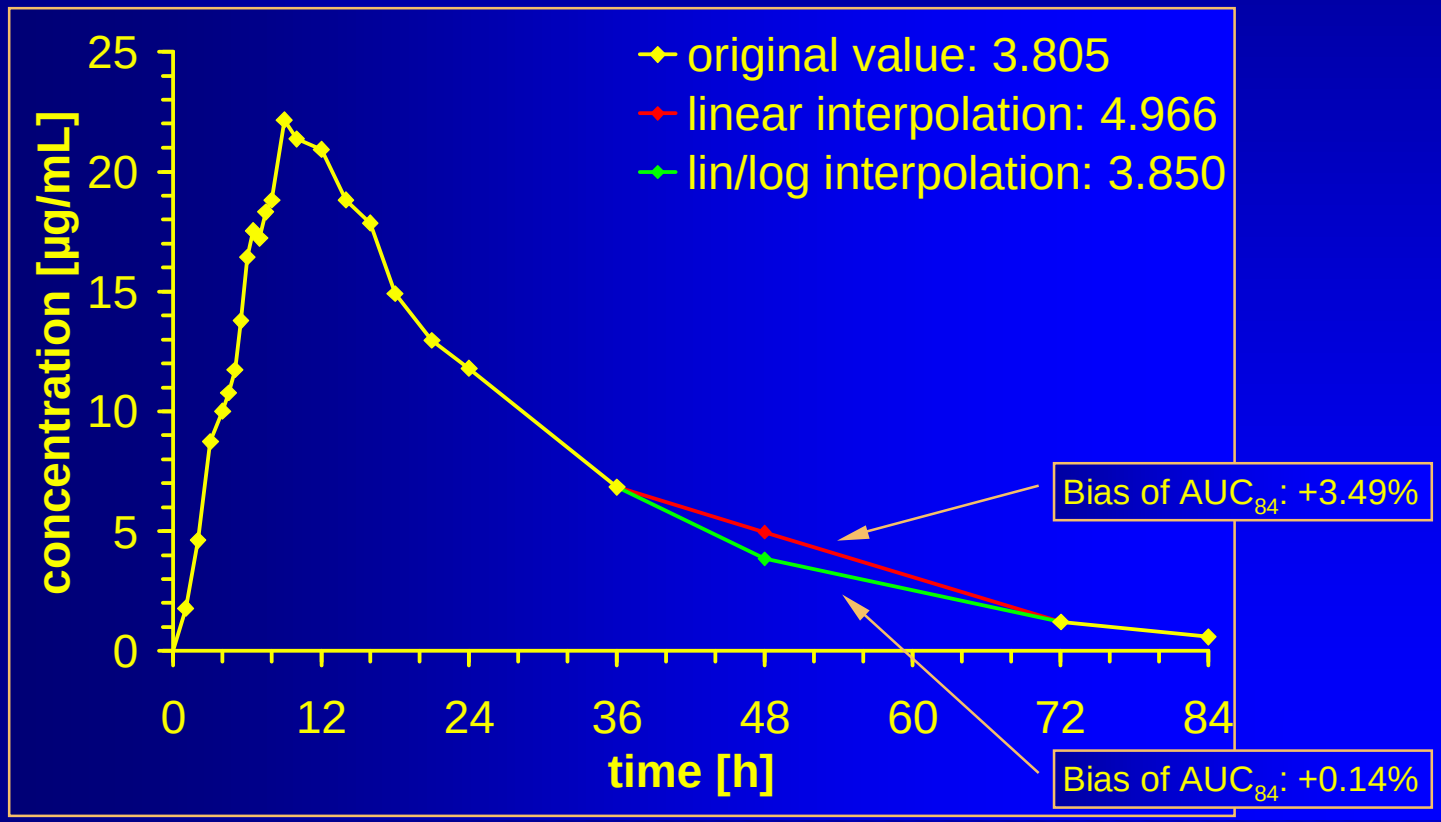

plasma profile (linear scale)
concentration
200
150
100
50
0
0
24
48
72
96
120
144
time
C_{av}
slope: +0.04593
95% CI: [-0.35266 | +0.44452]
steady state demonstrated

Some Problems…

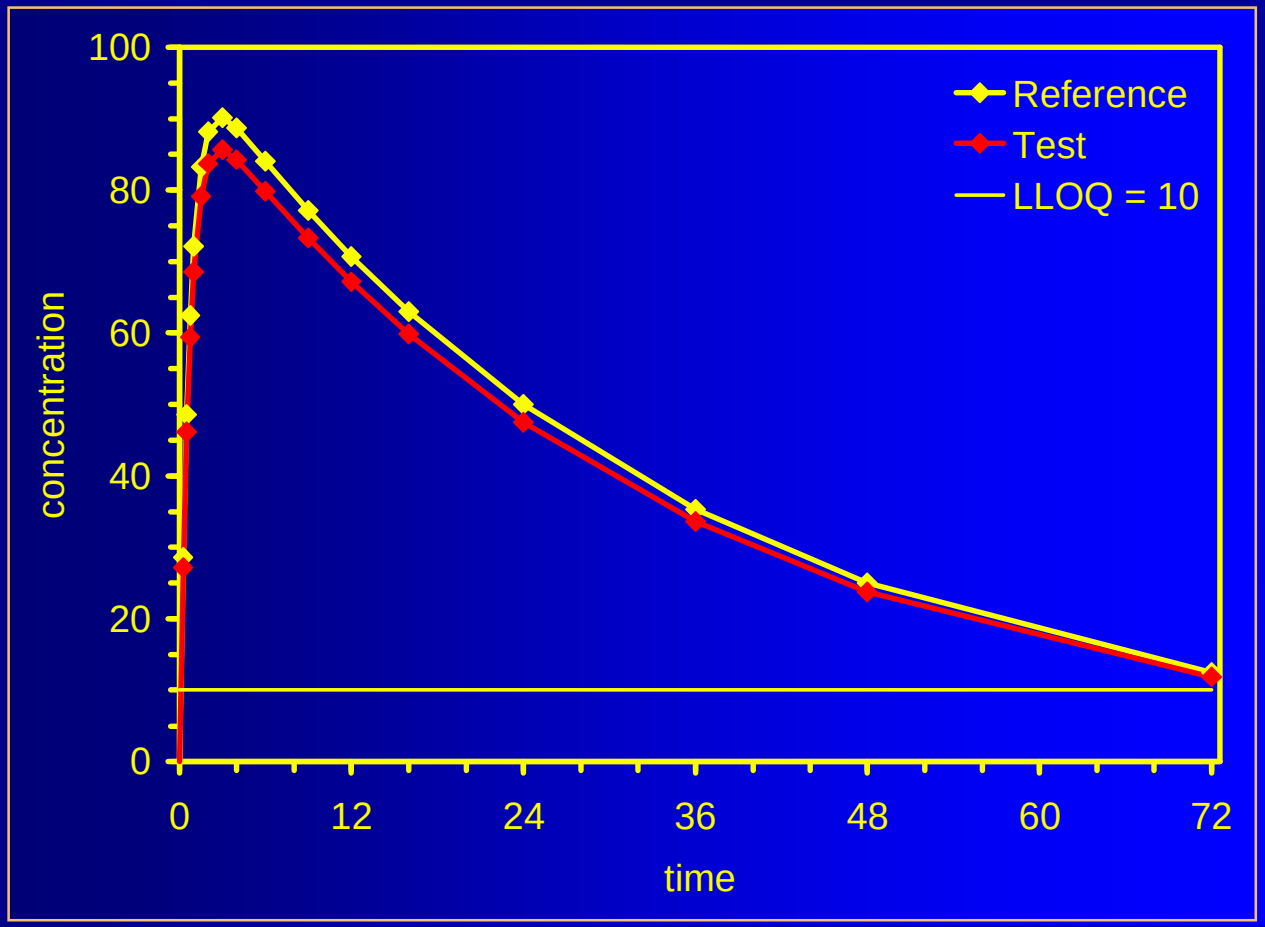
- Missing values
  - Procedure for Imputation must be stated in the Protocol; recommended:
    - in the Absorption Phase ($t < t_{max}$) by linear Interpolation of two adjacent values
    - in the Elimination Phase ($t \geq t_{max}$) by log/linear Interpolation of two adjacent values
    - estimated value must not be used in calculation of the apparent half life!
  - Don't rely on softwares' defaults!
    - Phoenix/WinNonlin interpolates linear – unless lin-up/log-down trapezoidal method is used
    - Kinetica interpolates log/lin within descending values

Some Problems…

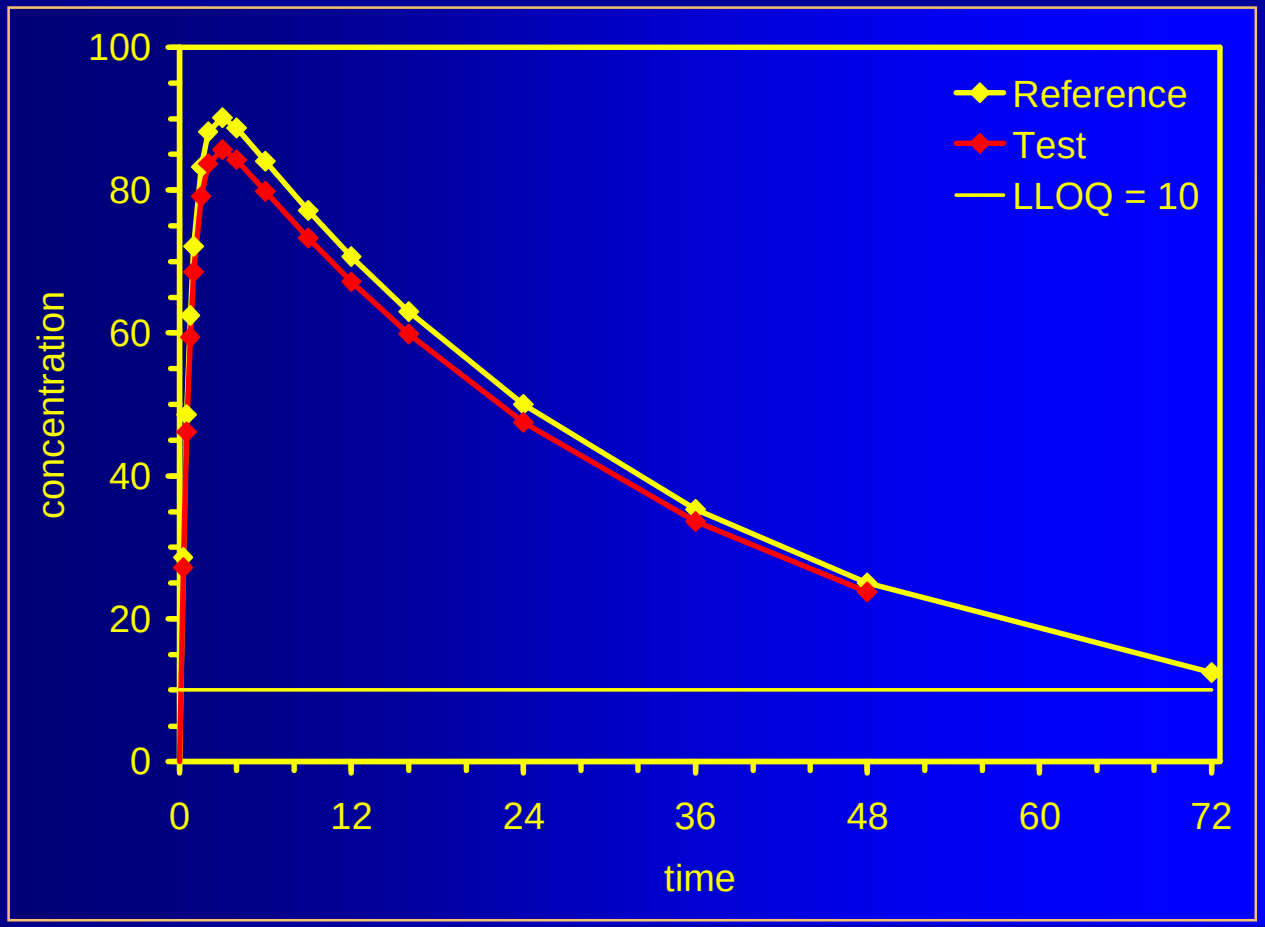
- Missing values



Some Problems…



Some Problems…

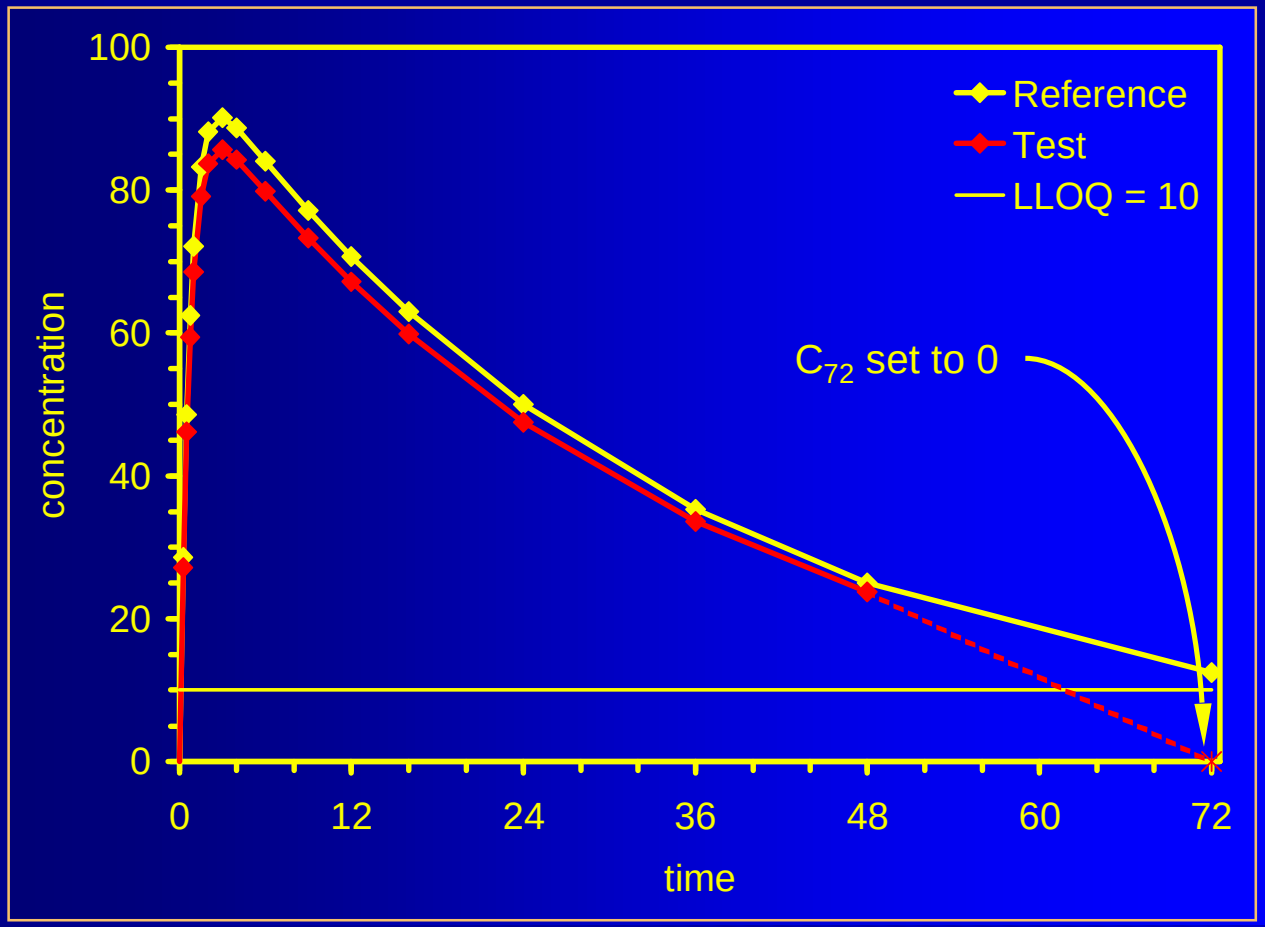


Some Problems…

- Missing values
  - Last value of test missing (*e.g.*, vial broken)
    - $AUC_{t_{last}}$ (48) T = 2407
      $AUC_{t_{last}}$ (72) R = 2984
      T/R = 80.67% biased!
    - Using AUC to t where $C \geq$ LLOQ for both formulations (48)
      AUC_{48} T = 2534
      AUC_{48} R = 2407
      T/R = 95% ✓
      - Not available in software
      - Regulatory acceptance?

	Reference		Test	
time	conc	AUC_{0-t}	conc	AUC_{0-t}
0	BLQ	0	BLQ	0
0.25	28.57	4	27.14	3
0.50	48.57	13	46.14	13
0.75	62.50	27	59.38	26
1.00	72.15	44	68.55	42
1.5	83.26	83	79.10	79
2	88.14	126	83.73	119
3	90.14	215	85.63	204
4	88.70	304	84.26	289
6	84.07	477	79.86	453
9	77.11	719	73.25	683
12	70.71	940	67.18	893
16	63.00	1208	59.85	1147
24	50.00	1660	47.50	1577
36	35.36	2172	33.59	2063
48	25.00	**2534**	23.75	**2407**
72	12.50	**2984**	Missing	NA

Some Problems…



Some Problems…

- Missing values
  - Last value of test missing (*e.g.*, vial broken)
    - Setting the first concentration in the profile where C<LLOQ to zero. AUC_{all}, 'invented' by Pharsight

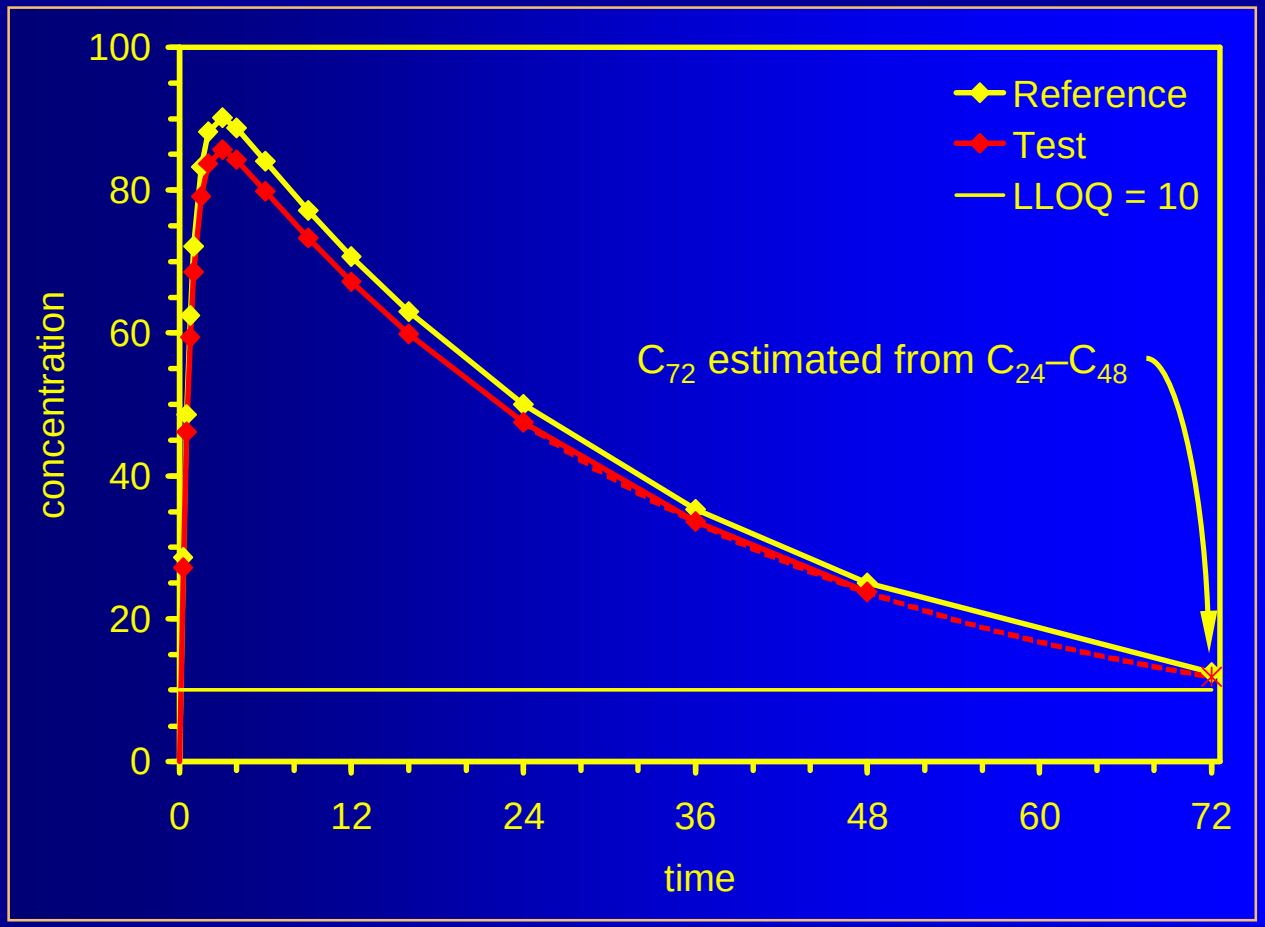
      AUC_{all} (72) T = 2692

      AUC_{all} (72) R = 2984

      T/R = 90.22% biased!
      - Available in Phoenix / WinNonlin, Kinetica
      - Regulatory acceptance?

	Reference		Test	
time	conc	AUC_{0-t}	conc	AUC_{0-t}
0	BLQ	0	BLQ	0
0.25	28.57	4	27.14	3
0.50	48.57	13	46.14	13
0.75	62.50	27	59.38	26
1.00	72.15	44	68.55	42
1.5	83.26	83	79.10	79
2	88.14	126	83.73	119
3	90.14	215	85.63	204
4	88.70	304	84.26	289
6	84.07	477	79.86	453
9	77.11	719	73.25	683
12	70.71	940	67.18	893
16	63.00	1208	59.85	1147
24	50.00	1660	47.50	1577
36	35.36	2172	33.59	2063
48	25.00	2534	23.75	2407
72	12.50	**2984**	= *0	**2692**

Some Problems…



Some Problems…

- Missing values
  - Last value of test missing (*e.g.*, vial broken)
    - Estimating the missing value from elimination phase.

      $AUC_{72}{}^{*}$ T = 2835

      AUC_{72} R = 2984

      T/R = 95% ✓
      - Not available in software
      - Regulatory acceptance ±

	Reference		Test	
time	conc	AUC_{0-t}	conc	AUC_{0-t}
0	BLQ	0	BLQ	0
0.25	28.57	4	27.14	3
0.50	48.57	13	46.14	13
0.75	62.50	27	59.38	26
1.00	72.15	44	68.55	42
1.5	83.26	83	79.10	79
2	88.14	126	83.73	119
3	90.14	215	85.63	204
4	88.70	304	84.26	289
6	84.07	477	79.86	453
9	77.11	719	73.25	683
12	70.71	940	67.18	893
16	63.00	1208	59.85	1147
24	50.00	1660	47.50	1577
36	35.36	2172	33.59	2063
48	25.00	2534	23.75	2407
72	12.50	**2984**	*11.88	***2835**

Some Problems…

- Missing values
  - Values below the lower limit of quantitation (LLOQ)
    - Example as before, but LLOQ = 12.5 (instead 10)

AUC_{72}: T = ?, R = 2984
T/R = ?

AUC_{48}: T = 2407, R = 2534
T/R = 95% ✓

AUC_{all}: T = 2692, R = 2984
T/R = 90.22% biased!

AUC_{72}^{*}: T = ?, R = 2984
T/R = ?

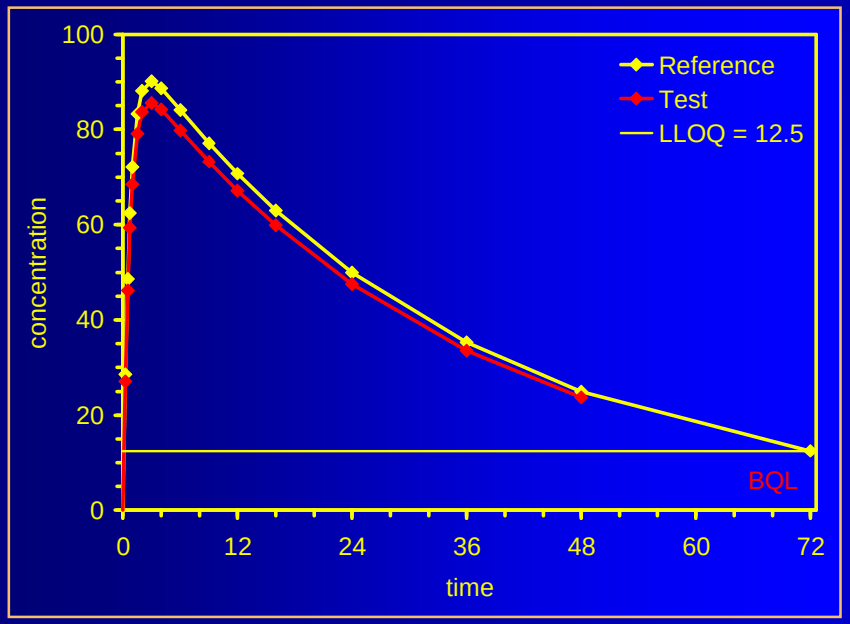
	Reference		Test	
time	conc	AUC_{0-t}	conc	AUC_{0-t}
24	50.00	1660	47.50	1577
36	35.36	2172	33.59	2063
48	25.00	**2534**	23.75	2407
72	12.50	**2984**	BLQ	NA

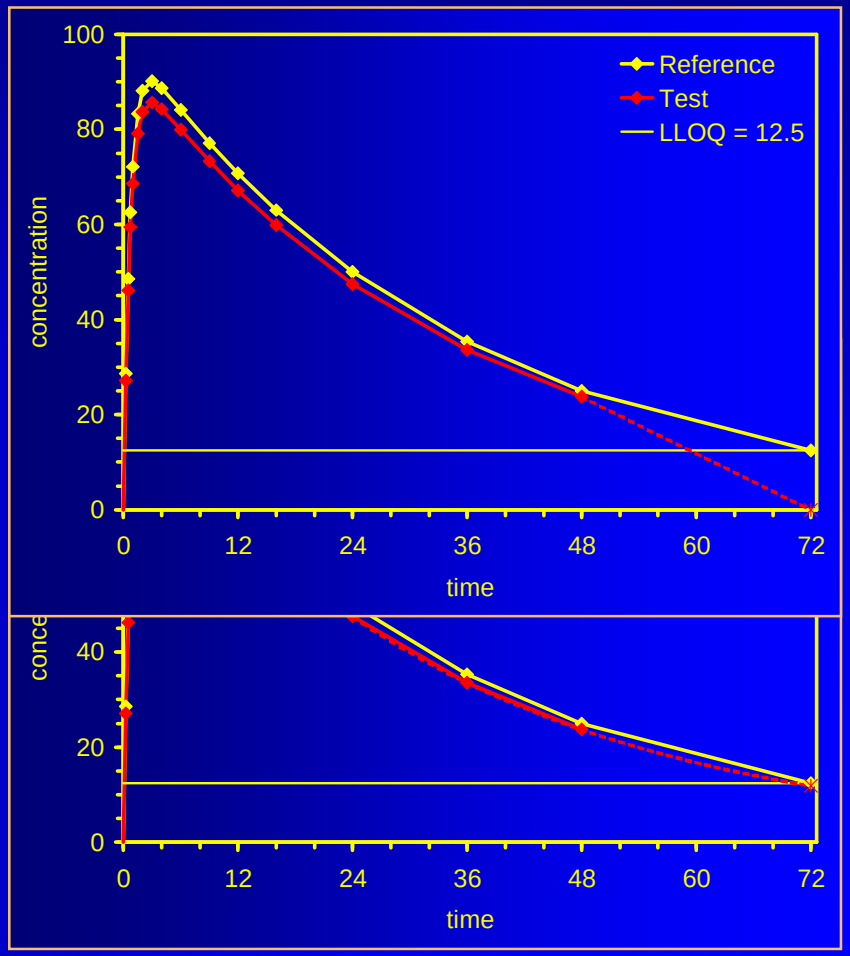
	Reference		Test	
time	conc	AUC_{0-t}	conc	AUC_{0-t}
24	50.00	1660	47.50	1577
36	35.36	2172	33.59	2063
48	25.00	2534	23.75	2407
72	12.50	**2984**	= *0	**2692**

	Reference		Test	
time	conc	AUC_{0-t}	conc	AUC_{0-t}
24	50.00	1660	47.50	1577
36	35.36	2172	33.59	2063
48	25.00	2534	23.75	2407
72	12.50	**2984**	*11.88	NA

Some Problems…

What would you do?



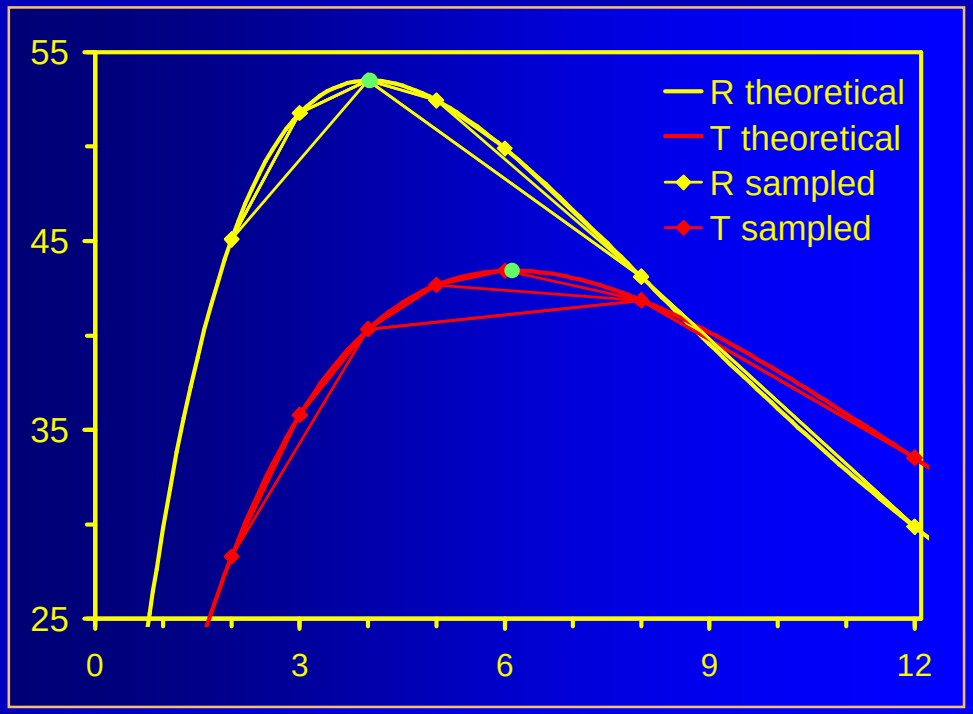


Sampling at C_{max}

- With *any (!)* given sampling scheme the *'true'* C_{max} is missed
  - It is unlikely that you sample *exactly* at the true C_{max} for any given subject
  - High inter- and/or intra-subject variability (single point metric)
  - Variability higher than *AUC's*
  - In many studies the win/loose metric!
  - Try to decrease variability
    - Increase sample size (more subjects)
    - Increase sampling *within* each subject (*maybe* better)

Sampling at C_{max}

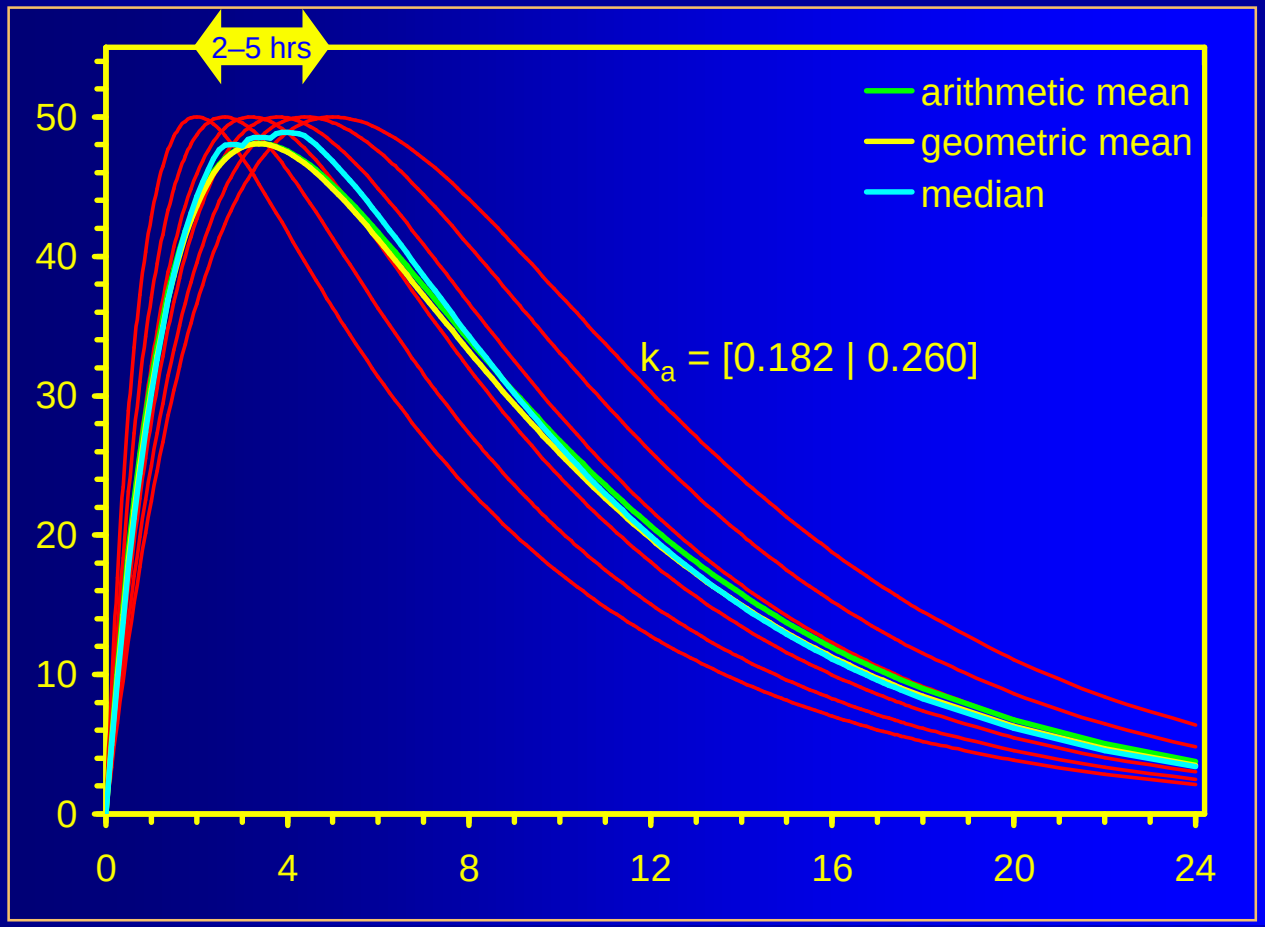
- Theoretical values (from PK simulation)
  C_{max}: 41.9/53.5 (81.2%), t_{max}: 6.11/4.02 (Δ 2.09)
  - # samples [2–12h]
    - 4
      - C_{max} 78.3%
      - t_{max} Δ 4
    - 5
      - C_{max} 78.3%
      - t_{max} Δ 4
    - 6
      - C_{max} 79.8%
      - t_{max} Δ 1
    - 7
      - C_{max} 81.2%
      - t_{max} Δ 2



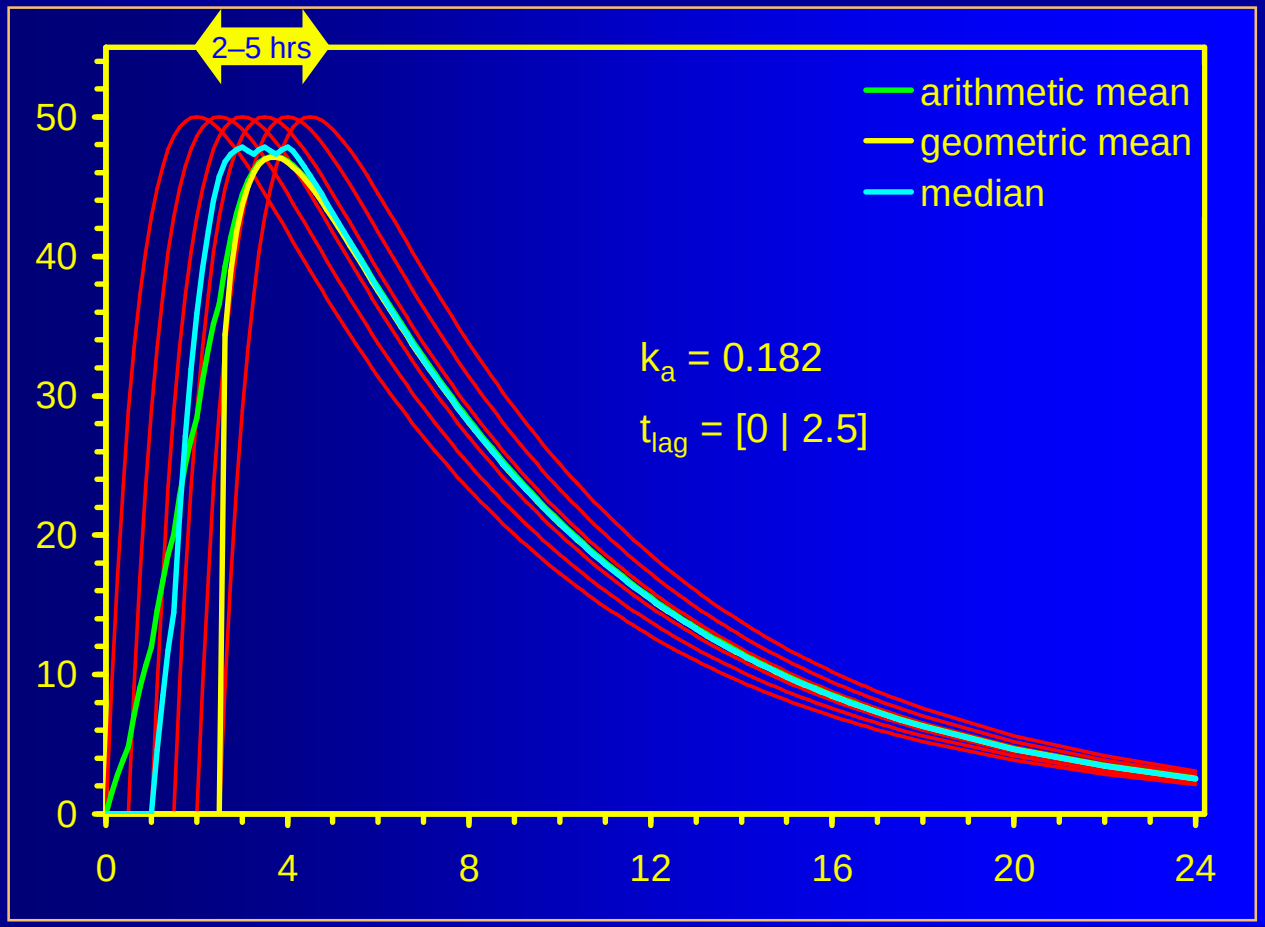
Sampling at C_{max}

- 'C_{max} was observed within two to five hours after administration …'
  - Elimination is drug specific,
  - but what about absorption?
    - Formulation specific (k_a and/or t_{lag})!
    - Dependent on the sampling schedule (in a strict sense study-specific)

Sampling at C_{max}



Sampling at C_{max}



Another Problem

- EMA GL on BE (2010)
  - Section 4.1.8 Reasons for exclusion 1)
    - A subject with lack of any measurable concentrations or only very low plasma concentrations for reference medicinal product. A subject is considered to have very low plasma concentrations if its AUC is less than 5% of reference medicinal product geometric mean AUC (which should be calculated without inclusion of data from the outlying subject). The exclusion of data […] will only be accepted in exceptional cases and may question the validity of the trial.

      Remark: Only possible after unblinding!

Another Problem

- EMA GL on BE (2010)
  - Section 4.1.8 Resons for exclusion 1) cont'd
    - The above can, for immediate release formulations, be the result of subject non-compliance […] and should as far as possible be avoided by mouth check of subjects after intake of study medication to ensure the subjects have swallowed the study medication […]. The samples from subjects excluded from the statistical analysis should still be assayed and the results listed.

Another Problem

- Gastro-resistant (enteric coated) preparations
  - Gastric emptying of single unit dosage forms non-disintegrating in the stomach is prolonged and highly erratic. The consequences of this effect on the enteric coating of delayed release formulations are largely unpredictable.
    - Sampling period should be designed such that measurable concentrations are obtained, taking into consideration not only the half-life of the drug but the possible occurrence of this effect as well. This should reduce the risk of obtaining incomplete concentration-time profiles due to delay to the most possible extent. These effects are highly dependent on individual behaviour.

Another Problem

- Gastro-resistant (enteric coated) preparations
  - Therefore, but only under the conditions that sampling times are designed to identify very delayed absorption and that the incidence of this outlier behaviour is observed with a comparable frequency in both, test and reference products, these incomplete profiles can be excluded from statistical analysis provided that it has been considered in the study protocol.

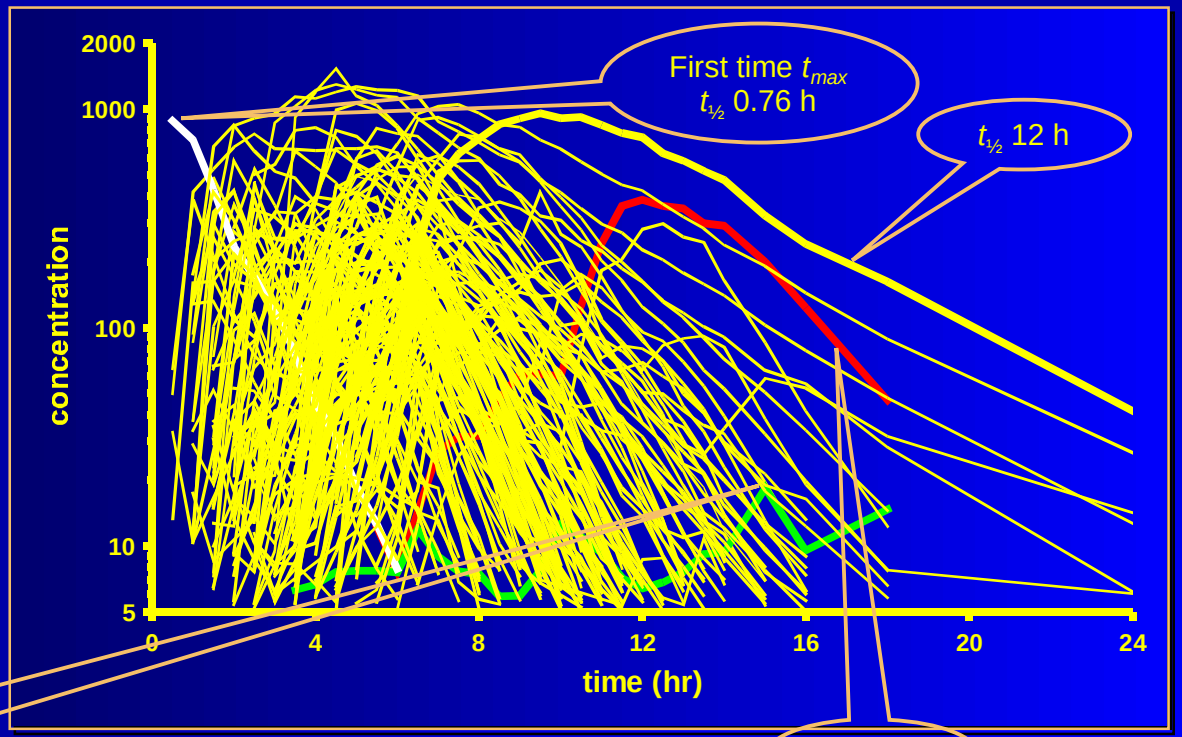
EMEA, CHMP Efficacy Working Party therapeutic subgroup on Pharmacokinetics (EWP-PK)
Questions & Answers: Positions on specific questions addressed to the EWP therapeutic subgroup on Pharmacokinetics
EMEA/618604/2008 Rev. 3, 26 January 2011
http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002963.pdf

What is *'comparable'*? For a study in 24 subjects, we get a significant difference for 5/0 (Fisher's exact test: *p* 0.0496).

Case Study (PPI)

- Attempt to deal with high variability in C_{max}

Powered to 90% according to CV from previous studies; 140 (!) subjects and to 80% for expect-ed dropout rate. Sampling every 30 min up to 14 hours (7785 total).



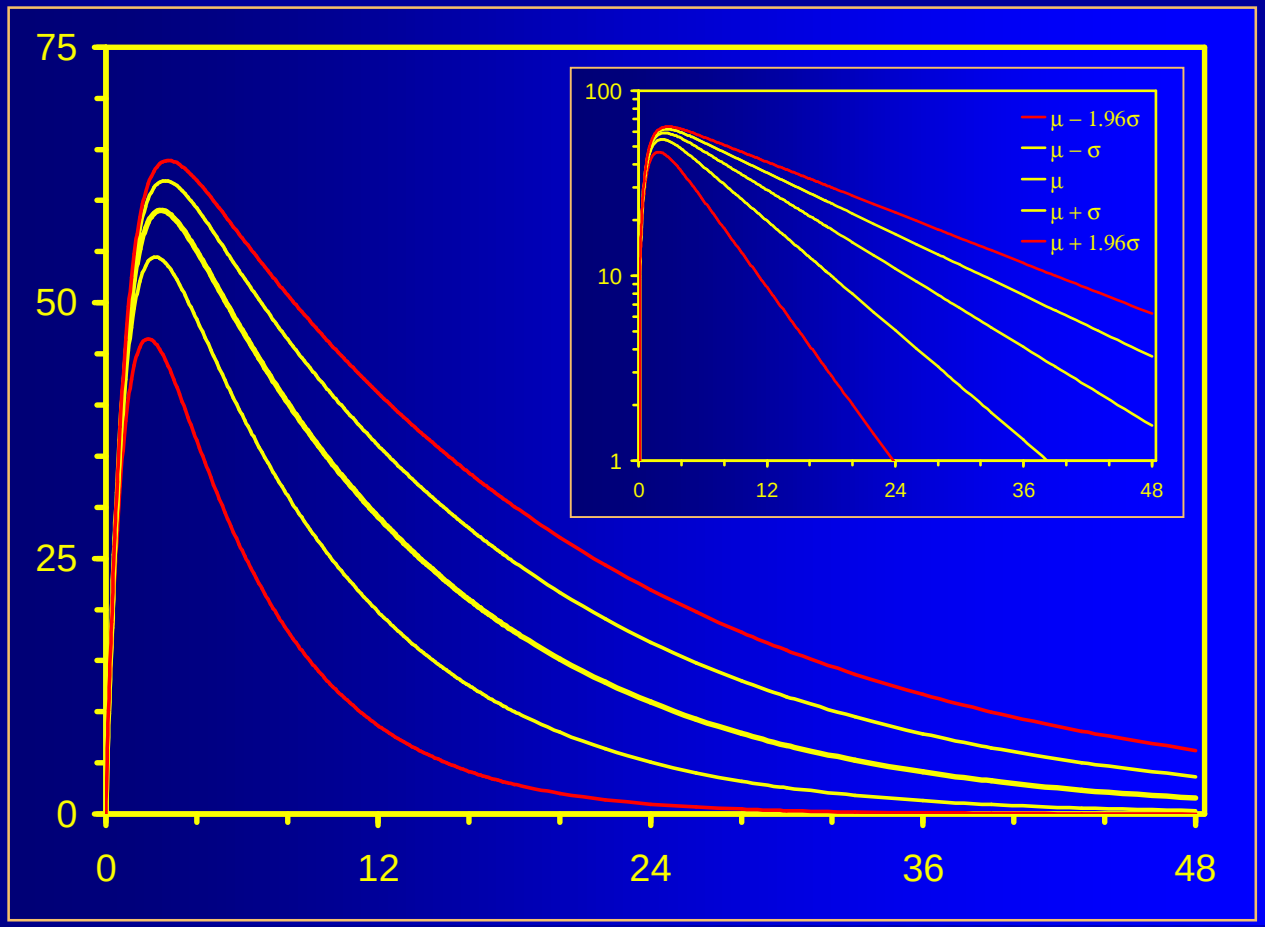
Half lives

- Drug specific, *but* …
  - The *apparent* elimination represents the *slowest* rate constant (controlled release, topicals, transdermals) – *not* necessarily elimination!
  - Avoid the term ‘terminal elimination’ – might not be true
  - Important in designing studies
    - To meet $AUC_t \geq 80\%\ AUC_\infty$ criterion
    - To plan sufficiently long wash-out (avoid carry-over)
    - To plan saturation phase for steady state

Half lives

- Dealing with literature data
  - What if only mean ±SD is given?
    - Assuming normal distribution:
      $\mu \pm \sigma$ covers 68.27% of values (15.87% of values are expected to lie outside of $\mu \pm \sigma$)
    - Example: 8.5 ± 2.4 hours, 36 subjects.
      0.1587 × 36 = 5.71 or in at least five subjects we may expect a half life of > 10.9 hours.
    - Plan for 95% coverage ($z_{0.95}$ = 1.96): $p_{0.95} = \mu \pm z_{0.95} \times \sigma$
      8.5 ± 1.96 × 2.4 = [3.80, 13.2] hours.
      We may expect a half life of >13.2 hours in ~one subject (0.05/2 × 36 = 0.90).

Half lives

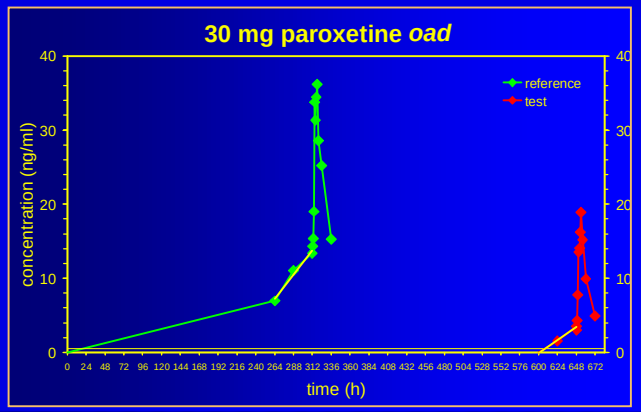


Single Dose / Multiple Dose

- Single Dose recommended in most GLs, but steady-state studies
  - may be required:
    - in the case of dose- or time-dependent pharmacokinetics
    - for most modified release products (*additionally* to single dose BE)
  - may be considered:
    - if problems of sensitivity preclude sufficiently precise plasma concentration measurements after SD administration. With current developments in bioanalytical methodology, you should have strong evidence of infeasibility if you claim the necessity of a MD study based on lacking methods.
      Regulators are concerned with efficacy/safety issues – not with the budget of pharmaceutical companies!

Single Dose / Multiple Dose

- Steady-state studies
  - No Wash-out between Periods (Switch-Over)!
  - In order to fulfil the superposition principle of linear pharmacokinetics ($AUC_\tau = AUC_\infty$), you should demonstrate achievement of steady-state
    - Linear regression of pre-dose values in saturation phase
      - slope (from at least the last three values) should not significantly (p>0.05, two-sided) differ from zero,
      - subjects not in steady-state at begin of the profile(s) should be excluded from the evaluation – if stated in protocol!



Washout in MD Studies

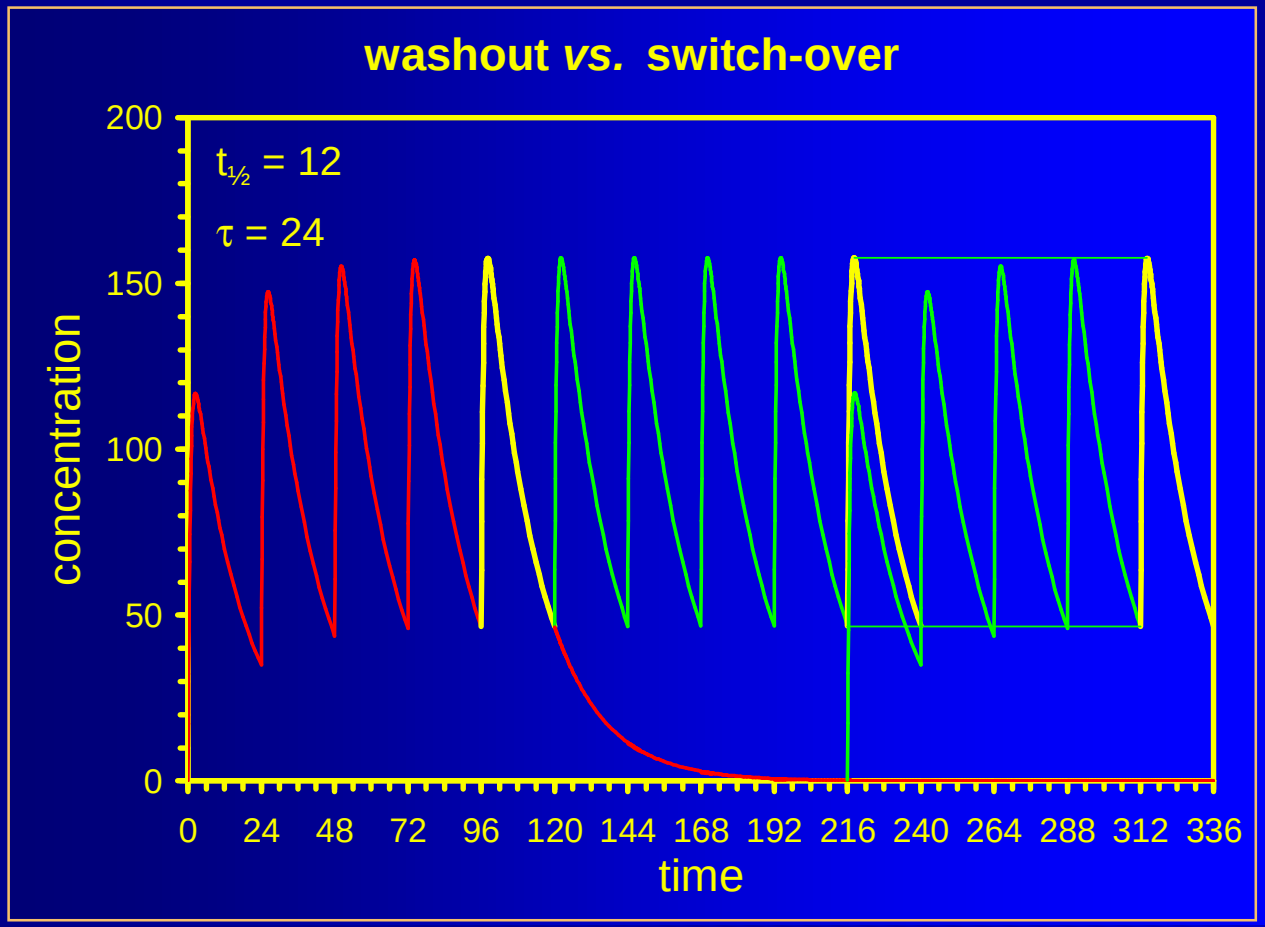
- EMA GL on BE (2010)

  The treatment periods should be separated by a wash out period sufficient to ensure that drug concentrations are below the lower limit of bioanalytical quantification in all subjects at the beginning of the second period. Normally at least 5 elimination half-lives are necessary to achieve this. In steady-state studies, the wash out period of the previous treatment last dose can overlap with the build-up of the second treatment, provided the build-up period is sufficiently long (at least 5 times the terminal half-life).

  - Justified by PK Superposition Principle
  - 'Switch-over Design'

2001 NfG:
3 half-lives

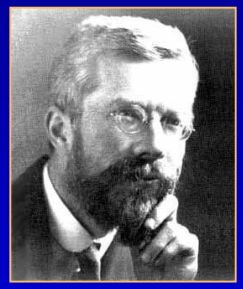
Washout in MD Studies

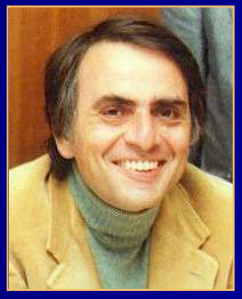


To bear in Remembrance...

To call the statistician after the experiment is done may be no more than asking him to perform a *post-mortem* examination: he may be able to say what the experiment died of.

Ronald A. Fisher



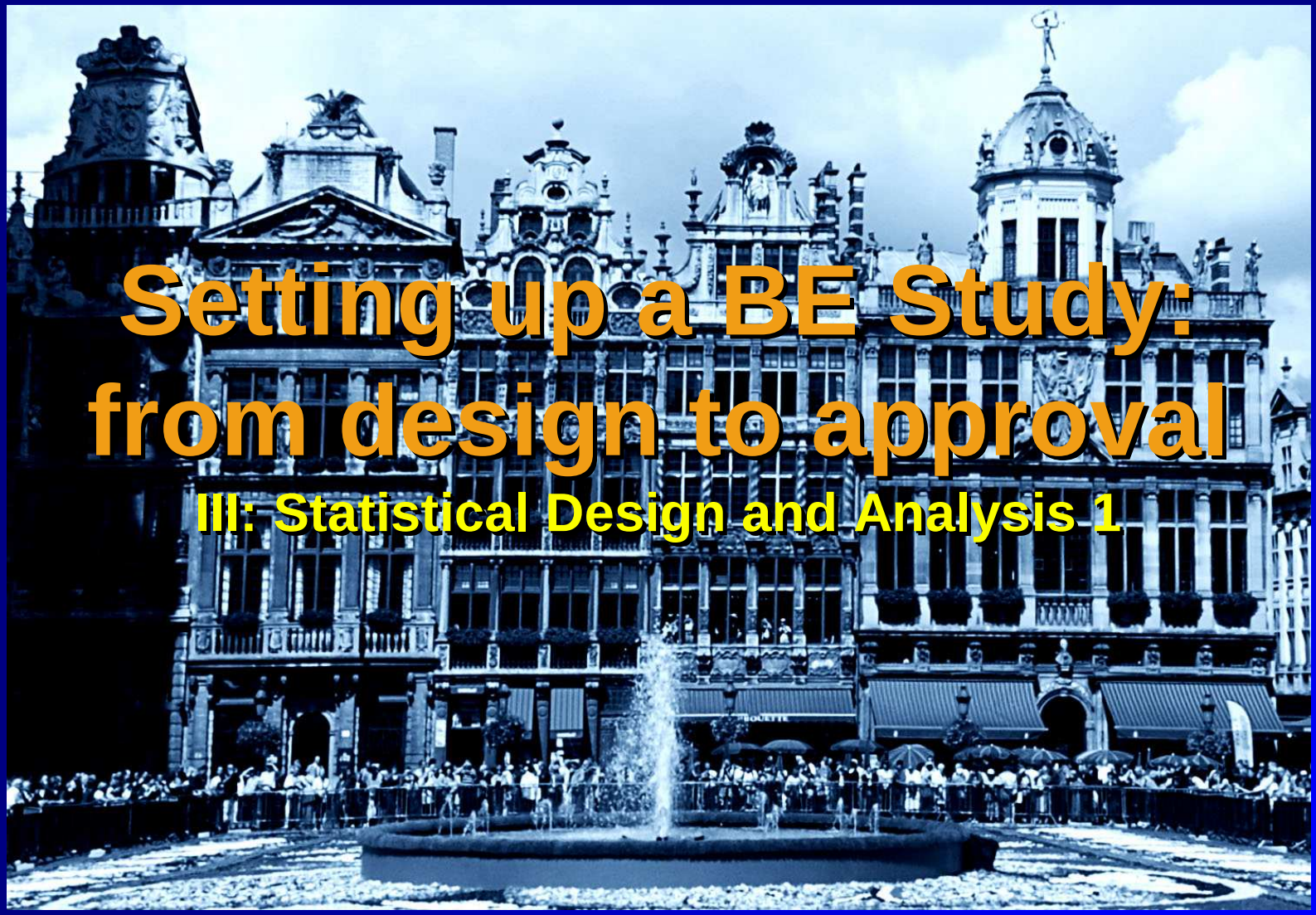


[The] impatience with ambiguity can be criticized in the phrase:
absence of evidence is not evidence of absence.

Carl Sagan

[…] our greatest mistake would be to forget that data is used for serious decisions in the very real world, and bad information causes suffering and death.

Ben Goldacre



Setting up a BE Study: from design to approval

III: Statistical Design and Analysis 1

Assumptions: Statistics

Multiplicative Model (X-over without carryover)

$$X_{ijk} = \mu \cdot \pi_k \cdot \Phi_l \cdot s_{ik} \cdot e_{ijk}$$

X_{ijk}: *ln*-transformed response of j-th subject ($j=1,\ldots,n_i$) in i-th sequence ($i=1,2$) and k-th period ($k=1,2$), μ: global mean, μ_l: expected formulation means ($l=1,2$: $\mu_1=\mu_{test}$, $\mu_2=\mu_{ref.}$), π_k: fixed period effects, Φ_l: fixed formulation effects ($l=1,2$: $\Phi_1=\Phi_{test}$, $\Phi_2=\Phi_{ref.}$)

Assumptions: Statistics

Multiplicative Model (X-over without carryover)

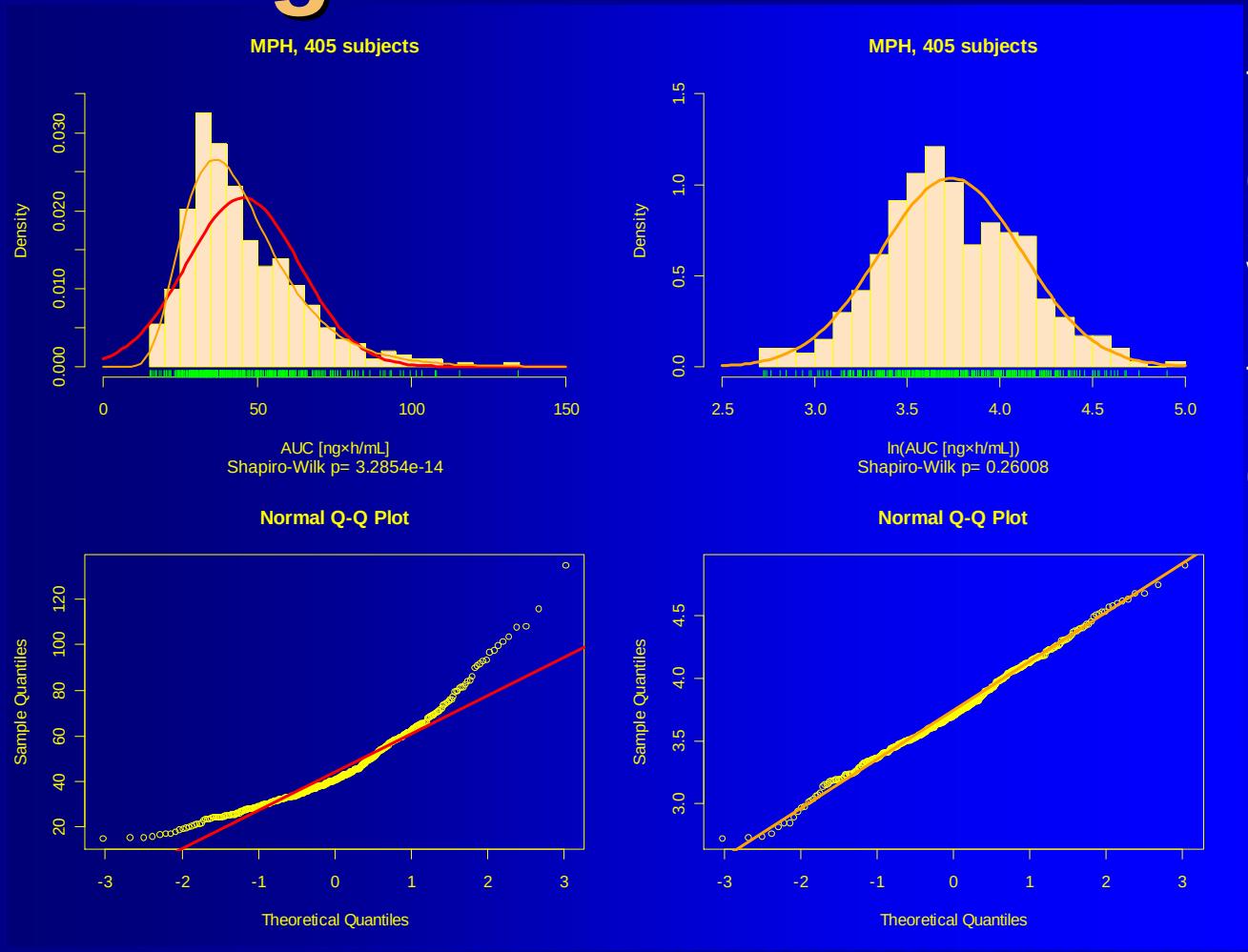
$$X_{ijk} = \mu \cdot \pi_k \cdot \Phi_l \cdot s_{ik} \cdot e_{ijk}$$

s_{ik}: random subject effect, e_{ijk}: random error

Main Assumptions:

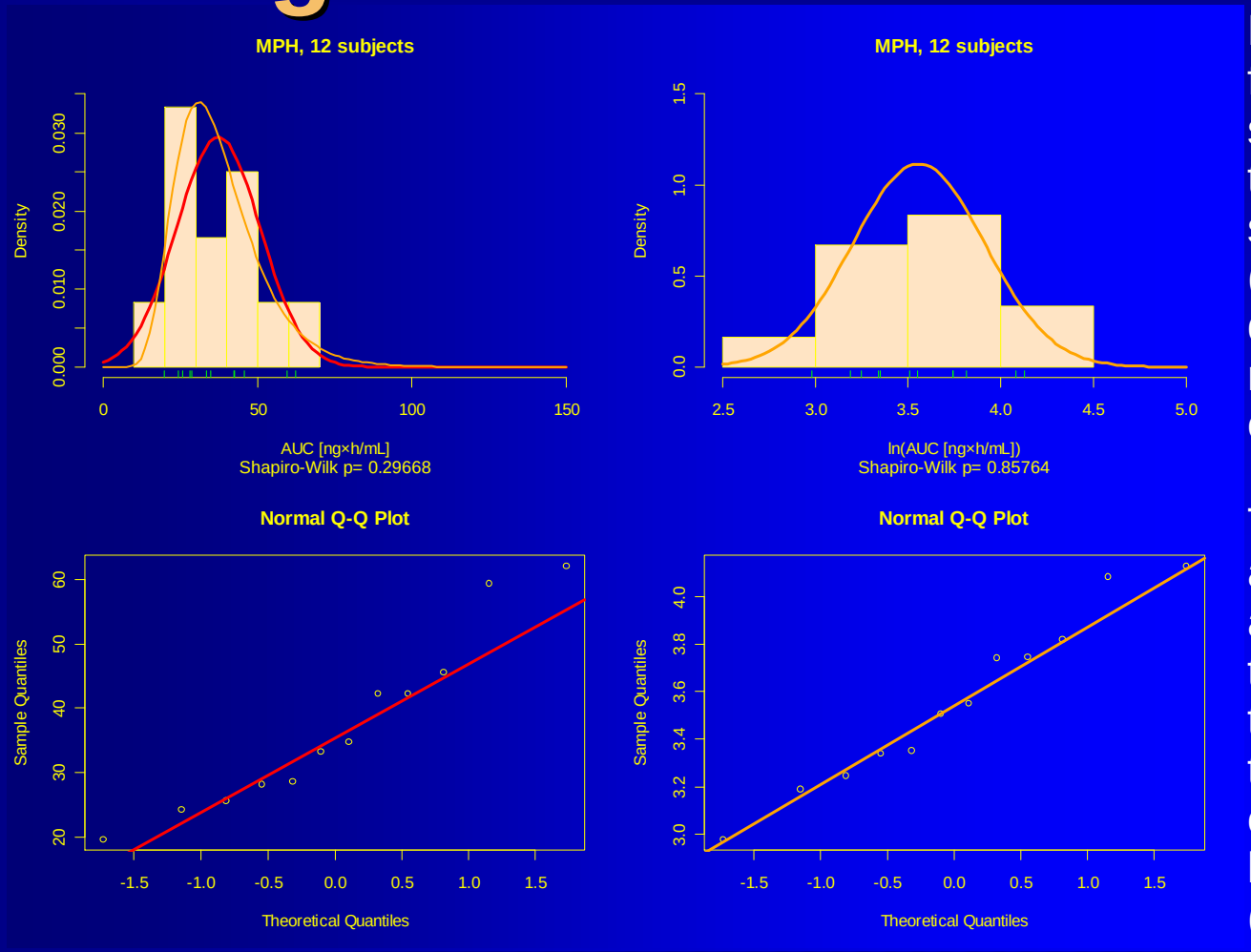
- All $ln\{s_{ik}\}$ and $ln\{e_{ijk}\}$ are independently and normally distributed about unity with variances σ^2_s and σ^2_e.
- All observations made on different subjects are independent.

Log-Transformation



Clearly in favor of a lognormal distribution. Shapiro-Wilk test highly significant for normal distribution (rejected).

Log-Transformation



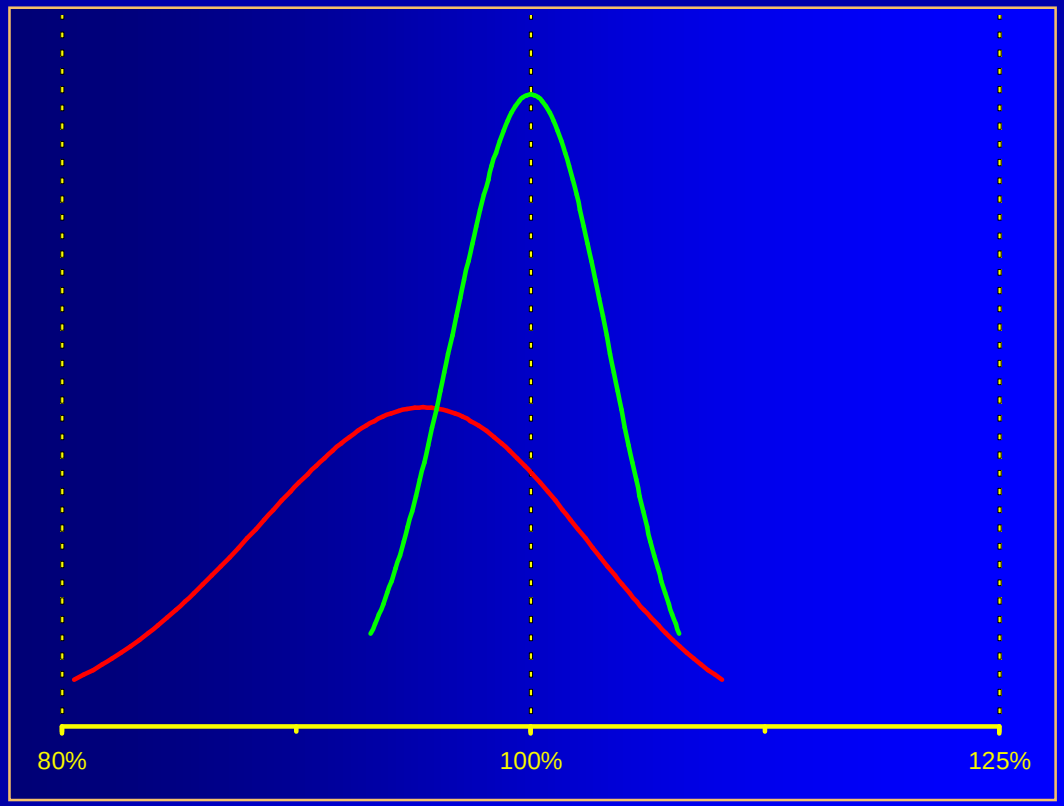
Data set from a real study. Both tests *not* significant (assumed distributions not reject-ed).

Tests not acceptable according to GLs; log-transforma-tion based on prior knowledge (PK)!

Science → Regulations

- Independent Identically Distributions (IID)
  What if …

$$\sigma^2_{WT} \neq \sigma^2_{WR}$$



Global Harmonization?

- Drugs with a narrow therapeutic range
  - USA, Japan — No difference to other drugs
  - WHO, EU, NZ, India — 90 % CI; Acceptance range may be tightened, *e.g.*, 0.9000–1.1111
  - RSA — 90 % CI within 0.80–1.25 (C_{max})
  - Brazil — 95 % CI within 0.80–1.25
  - Canada — No different procedure given in GL, but considered in current draft
    - AUC — 90 % CI within 0.90–1.12
    - C_{max} — 90 % CI within 0.80–1.25

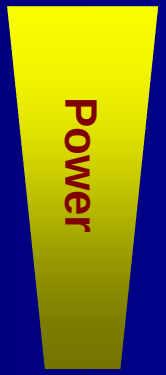
http://www.hc-sc.gc.ca/dhp-mps/alt_formats/pdf/consultation/drug-medic/draft_ebauche_cbs-eng.pdf (25 Jan 2010)

Basic Designs

- Single Dose / Multiple Dose
  - Cross-over
    - Standard 2×2
    - Higher Order Designs (for more than two treatments)
      - Incomplete Block Designs
      - Latin Squares
      - Variance Balanced Designs (Williams' Designs)
    - Replicate designs
  - Parallel Groups

Basic Designs

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



Full replicate (TRTR | RTRT)

Partial replicate (TRR | RTR | RRT)

Standard 2×2 cross-over (RT | TR)

Parallel (R | T)

Basic Designs

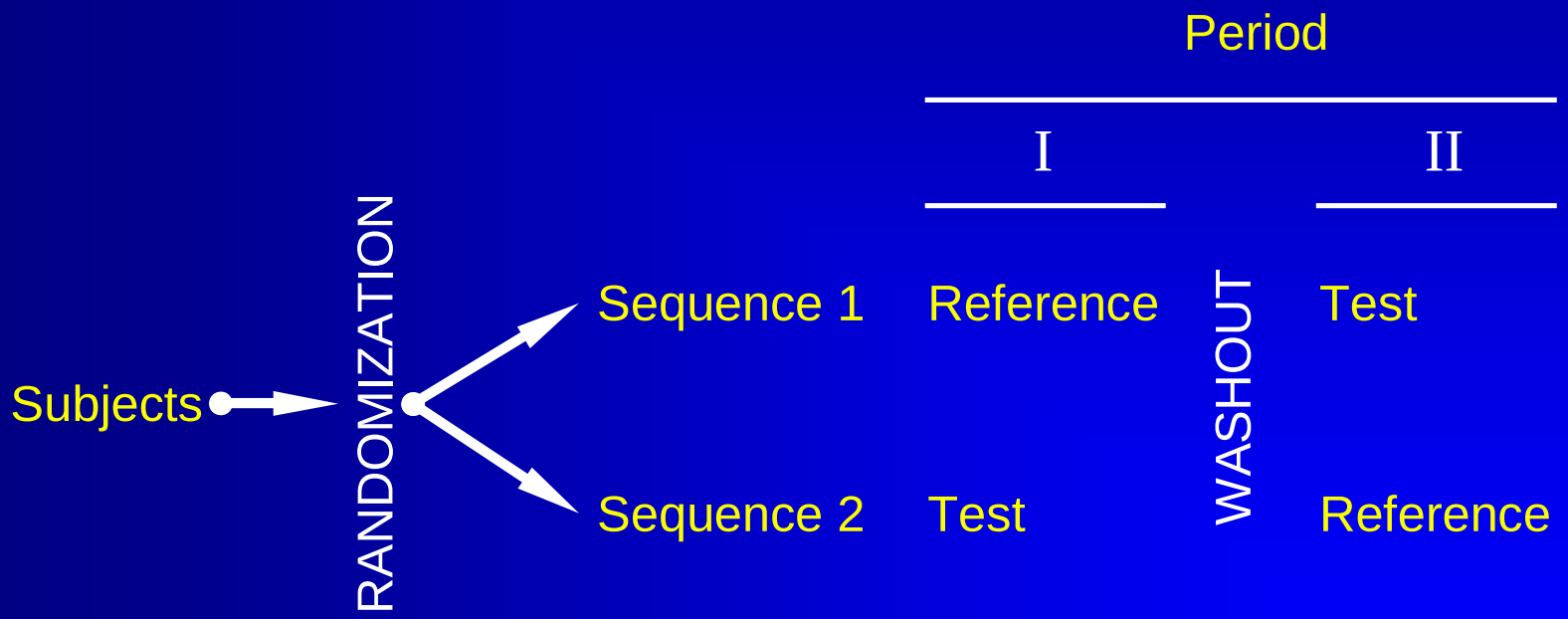
- Parallel Groups (patients, long half-life drugs)
- Cross-over (generally healthy subjects)
  - Higher Order Designs (more than two formulations)
    - Incomplete Block Designs
    - Latin Squares
    - Variance Balanced Designs (Williams' Designs)
  - Standard 2×2×2
  - Replicate designs

Cross-over designs

- Standard 2×2×2 (two-treatment two-sequence two-period) design
  - Each subject is randomly assigned to either sequence RT or sequence TR at two treatment periods
    - Dosing periods are separated by a washout period of sufficient length for the drug received in the first period to be completely metabolized or excreted from the circulation.
    - Smaller subject numbers compared to a parallel design, since the *within-subject* variability determines sample size (rather than *between-subject* variability).

Cross-over designs

- Standard 2×2×2 design



Cross-over designs:
Assumptions

Multiplicative Model (X-over without carryover)

$$X_{ijk} = \mu \cdot \pi_k \cdot \Phi_l \cdot s_{ik} \cdot e_{ijk}$$

- All $ln\{s_{ik}\}$ and $ln\{e_{ijk}\}$ are independently and normally distributed about unity with variances σ^2_s and σ^2_e.
  - ➔ This assumption may not hold true for all formulations; if the reference formulation shows higher variability than the test formulation, a *'good'* test will be penalized for the *'bad'* reference.
- All observations made on different subjects are independent.
  - ➔ This assumption should not be a problem, unless you plan to include twins or triplets in your study…

Cross-over designs

- Standard 2×2×2 design
  - Advantages
    - Globally applied standard protocol for BE
    - Straigthforward statistical analysis
  - Disadvantages
    - Not suitable for drugs with long half life (→ parallel groups)
    - Not optimal for studies in patients with instable diseases (→ parallel groups)
    - Not optimal if CV is uncertain (→ Two-Stage Sequential Designs)
    - Not optimal for HVDs/HVDPs (→ Replicate Designs)

Cross-over designs:
Evaluation

- Mainly by ANOVA and LMEM (linear mixed effects modeling). Results are identical for balanced datasets, and differ only slightly for imbalanced ones.
- Avoid M$-Excel! Almost impossible to validate; tricky for imbalanced datasets – a nightmare for higher-order X-overs. Replicates impossible.
- Software: SAS, Phoenix/WinNonlin, Kinetica*, EquivTest/PK*, S+, Package *bear* for *R*.

* 2×2 X-over only

Cross-over designs: Example

subject	T	R
1	28.39	35.44
2	39.86	49.42
3	32.75	36.78
4	33.36	33.40
5	34.97	34.81
6	24.29	24.65
7	28.61	31.77
8	45.44	45.54
9	59.49	65.29
10	27.87	28.23
11	24.26	25.71
12	42.30	37.01

	sequence RT			sequence TR	
subject	P I	P II	subject	P I	P II
2	39.86	49.42	1	28.39	35.44
3	32.75	36.78	4	33.36	33.40
5	34.97	34.81	6	24.29	24.65
8	45.44	45.54	7	28.61	31.77
10	27.87	28.23	9	59.49	65.29
11	24.26	25.71	12	42.30	37.01

Ordered by treatment sequences (RT | TR)

ANOVA on log-transformed data →

Cross-over designs: Example

Sequence	Period 1		Period 2		Sequence mean	
1	1R = $X_{.11}$	3.5103	1T = $X_{.21}$	3.5768	$X_{..1}$	3.5436
2	2T = $X_{.12}$	3.5380	2R = $X_{.22}$	3.5883	$X_{..2}$	3.5631
Period mean	$X_{.1.}$	3.5241	$X_{.2.}$	3.5826	$X_{...}$	3.5533

RT = n_1 = 6

TR = n_2 = 6 $1/n_1+1/n_2$ 0.3333

balanced n = 12 1/n 0.0833 n_1+n_2-2 10

Analysis of Variance

Source of variation	df	SS	MS	F	P-value	CV
Inter-subjects						
Carry-over	1	0.00230	0.00230	0.0144	0.90679	
Residuals	10	1.59435	0.15943	29.4312	4.32E-6	28.29%
Intra-subjects						
Direct drug	1	0.00040	0.00040	0.0733	0.79210	
Period	1	0.02050	0.02050	3.7844	0.08036	
Residuals	10	0.05417	0.00542			7.37%
Total	23	1.67172				

δ_{ML} **1.0082** *MLE (maximum likelihood estimator) of Delta-ML*

X_R 3.5493 *LS (least squares mean for the reference formulation)* $\exp(X_R)$ 34.79

X_T 3.5574 *LS (least squares mean for the test formulation)* $\exp(X_T)$ 35.07

Cross-over designs: Example

Classical (Shortest) Confidence Interval

± x rule: 20 [100 - x; 1 / (100 - x)]

θ_L	-0.2231	θ_U	+0.2231	α 0.0500	p=1-2·α 0.9000
δ_L	**80%**	δ_U	**125%**	$t_{2\cdot\alpha,df}$	1.8125
L_1	-0.0463	U_1	0.0626	*difference within Theta-L AND Theta-U; bioequivalent*	
L_2	**95.47%**	U_2	**106.46%**	***difference within Delta-L AND Delta-U; bioequivalent***	

δ_{ML}	**100.82%**	*MLE; maximum likelihood estimator*
δ_{MVUE}	100.77%	*MVUE; minimum variance unbiased estimator*
δ_{RM}	100.98%	*RM; ratio of formulation means*
δ_{MIR}	101.44%	*MIR; mean of individual subject ratios*

Cross-over designs: Example

- Calculation of 90% CI (2-way cross-over)
  - Sample size (n) 12, Point Estimate (PE) 100.82%, Residual Mean Squares Error (MSE) from ANOVA (*ln*-transformed values) 0.005417, $t_{\alpha,n-2}$ 1.8125
    - Standard Error (SE_Δ) of the mean difference

$$SE_\Delta = \sqrt{MSE}\sqrt{\frac{2}{n}} = \sqrt{0.005417}\sqrt{\frac{2}{12}} = 0.030047$$

    - Confidence Interval

$$CL_L = e^{\ln PE - t_{2\alpha,df}\cdot SE_\Delta} = e^{0.0081349 - 1.8125\times 0.030047} = 95.47\%$$

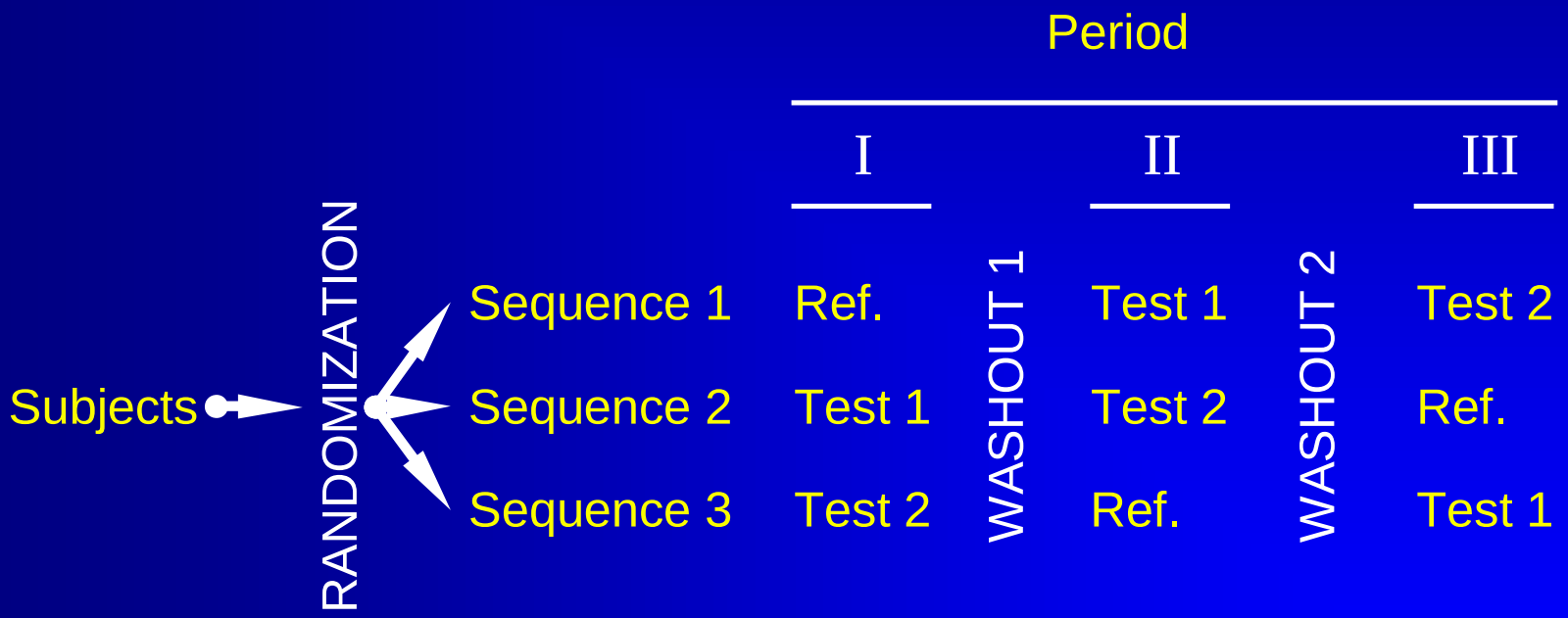
$$CL_H = e^{\ln PE + t_{2\alpha,df}\cdot SE_\Delta} = e^{0.0081349 + 1.8125\times 0.030047} = 106.46\%$$

Cross-over designs

- Higher Order Designs (for more than two treatments)
  - Latin Squares
    Each subject is randomly assigned to sequences, where number of treatments = number of sequences = number of periods.
  - Variance Balanced Designs

Cross-over designs

- 3×3×3 Latin Square design



Cross-over designs

- 3×3×3 Latin Square design
  - Advantages
    - Allows to choose between two candidate test formulations or comparison of a test formulation with two references.
    - Easy to adapt.
    - Number of subjects in the study is a multiplicative of three.
    - Design for establishment of Dose Proportionality.
  - Disadvantages
    - Statistical analysis more complicated (especially in the case of drop-outs and a small sample size) – not available in all software.
    - Extracted pairwise comparisons are imbalanced.
    - May need measures against multiplicity (increasing the sample size).
    - Not mentioned in any guideline.

Cross-over designs

- Higher Order Designs (for more than two treatments)
  - Variance Balanced Designs (Williams' Designs)
    - For *e.g.*, three formulations there are three possible pairwise differences among formulation means (*i.e.*, form. 1 *vs.* form. 2., form 2 *vs.* form. 3, and form. 1 *vs.* form. 3).
    - It is desirable to estimate these pairwise effects with the same degree of precision (there is a common variance for each pair).
      - Each formulation occurs only once with each subject.
      - Each formulation occurs the same number of times in each period.
      - The number of subjects who receive formulation *i* in some period followed by formulation *j* in the next period is the same for all *i # j*.
    - Such a design for three formulations is the three-treatment six-sequence three-period Williams' Design.

Cross-over designs

- Williams' Design for three treatments

Sequence	Period		
	I	II	III
1	R	T_2	T_1
2	T_1	R	T_2
3	T_2	T_1	R
4	T_1	T_2	R
5	T_2	R	T_1
6	R	T_1	T_2

Cross-over designs

- Williams' Design for four treatments

Sequence	Period			
	I	II	III	IV
1	R	T_3	T_1	T_2
2	T_1	R	T_2	T_3
3	T_2	T_1	T_3	R
4	T_3	T_2	R	T_1

Cross-over designs

- Williams' Designs
  - Advantages
    - Allows to choose between two candidate test formulations or comparison of a test formulation with two references.
    - Design for establishment of Dose Proportionality.
    - Paired comparisons (*e.g.*, for a nonparametric method) can be extracted, which are also balanced .
    - Mentioned in ANVISA GL and & hidden in EMA's.
  - Disadvantages
    - More sequences for an *odd* number of treatment needed than in a Latin Squares design (but equal for even number).
    - Statistical analysis more complicated (especially in the case of drop-outs) – not available in some softwares.
    - May need measures against multiplicity (increasing the sample size).

Cross-over designs

- Extraction of 2×2 comparisons (T_1/R, T_2/R)
  - Latin Squares

imbalanced

Seq.	P_1	P_2	P_3
1	T_1	T_2	R
2	T_2	R	T_1
3	R	T_1	T_2

Seq.	P_1'	P_2'
1	T_1	R
2	R	T_1
3	R	T_1

Seq.	P_1"	P_2"
1	T_2	R
2	T_2	R
3	R	T_2

  - Williams' design

balanced

Seq.	P_1	P_2	P_3
1	T_1	T_2	R
2	T_2	R	T_1
3	R	T_1	T_2
4	T_1	R	T_2
5	T_2	T_1	R
6	R	T_2	T_1

Seq.	P_1'	P_2'
1	T_1	R
2	R	T_1
3	R	T_1
4	T_1	R
5	T_1	R
6	R	T_1

Seq.	P_1"	P_2"
1	T_2	R
2	T_2	R
3	R	T_2
4	R	T_2
5	T_2	R
6	R	T_2

Cross-over designs

- Higher Order Designs (cont'd)
  - Bonferroni-correction needed (sample size!)
    - *If more than one formulation will be marketed* (for three simultaneous comparisons without correction patients' risk increases from 5 % to 14 %).
    - *Sometimes* requested by regulators in dose proportionality.

k	$P_{\alpha=0.05}$	$P_{\alpha=0.10}$	$\alpha_{adj.}$	$P_{\alpha adj.}$	$\alpha_{adj.}$	$P_{\alpha adj.}$
1	5.00%	10.00%	0.0500	5.00%	0.100	10.00%
2	9.75%	19.00%	0.0250	4.94%	0.050	9.75%
3	14.26%	27.10%	0.0167	4.92%	0.033	6.67%
4	18.55%	34.39%	0.0125	4.91%	0.025	9.63%
5	22.62%	40.95%	0.0100	4.90%	0.020	9.61%
6	26.49%	46.86%	0.0083	4.90%	0.017	9.59%

Cross-over designs

- Higher Order Designs (cont'd)
  - Effect of α-adjustment on sample size (expected T/R 95%, CV_{intra} 20%, power 80%)

CV%	2×2 α 0.05	6×3 $\alpha_{adj.}$ 0.025	comp. 2×2	4×4 $\alpha_{adj.}$ 0.0167	comp. 2×2
10.0	8	12	+50%	16	+100%
12.5	10	12	+20%	16	+60%
15.0	12	18	+50%	16	+33%
17.5	16	24	+50%	24	+50%
20.0	20	24	+20%	28	+40%
22.5	24	30	+25%	36	+50%
25.0	28	36	+29%	40	+49%
27.5	34	42	+24%	48	+41%
30.0	40	54	+35%	56	+40%

Cross-over designs

- Replicate designs
  - Each subject is randomly assigned to sequences, where at least one of the treatments is administered at least twice.
    - Not only the global within-subject variability, but also the within-subject variability per treatment may be estimated.
    - Smaller subject numbers compared to a standard 2×2×2 design – but outweighed by an increased number of periods.
    - Same overall number of individual treatments!
    - Mandatory in the EU if scaled acceptance range for C_{max} is aimed at (CV_{WR}>30% must be demonstrated within the study).

Cross-over designs

- Replicate designs
  - Advantages
    - Some experience from FDA's initiative on Population BE (PBE) and Individual BE (IBE).
    - Reference Scaled Average Bioequivalence (RSABE)
    - Handling of outliers (Subject-by-Formulation Interaction may be ruled out).
    - Mentioned in RSA GL; FDA's API GLs and EMA for C_{max}.
  - Disadvantages
    - Statistical analysis complicated (especially in the case of drop-outs and if RSABE is the target) – not available in standard software.
    - Many publications, but still no agreement on methodology (!)

Cross-over designs

- Replicate designs
  - Examples
    - Three-period two-sequence (3×2)
      T R T
      R T R
      Sample size to obtain the same power as a 2×2×2 study: 75%
    - Four-period two-sequence (4×2)
      T R T R
      R T R T
      Sample size to obtain the same power as a 2×2×2 study: 50%
    - *and many others…* (FDA for RSABE: TRR | RTR | RRT)
    - The statistical model is a little bit complicated – and dependent on the actual design

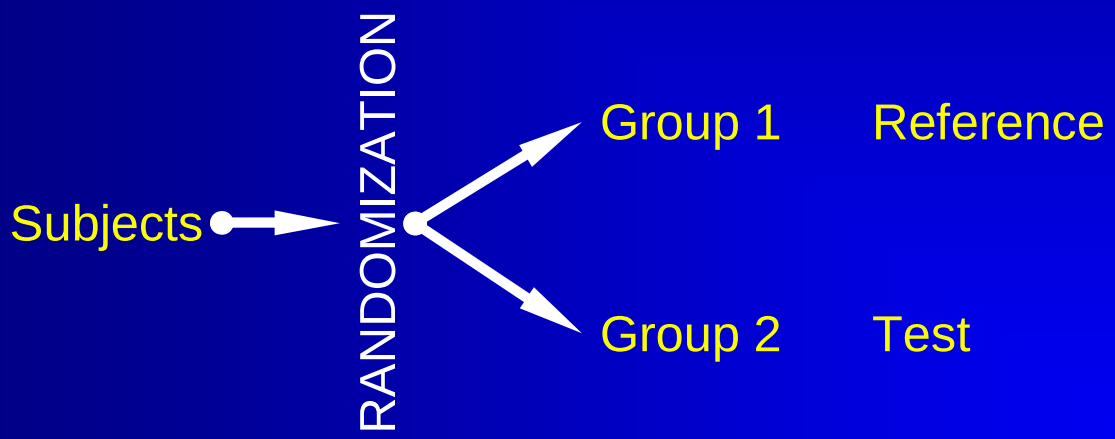
$$X_{ijkl} = \mu \cdot \pi_k \cdot \Phi_l \cdot s_{ij} \cdot e_{ijkl}$$

Parallel Groups

- Two-group parallel design
  - Each subject receives one – and only one – treatment in a random fashion
    - Usually each group contains the same number of subjects.
    - Higher subject numbers compared to a cross-over design, since the *total (between+within)-subject* variability determines sample size (rather than *within-subject* variability).

Parallel Groups

- Two-group parallel design



Parallel Groups

- Two-group parallel design
  - Advantages
    - Clinical part – *sometimes* – faster than X-over.
    - Straigthforward statistical analysis.
    - Drugs with long half life.
    - Potentially toxic drugs or effect and/or AEs unacceptable in healthy subjects.
    - Studies in patients, where the condition of the disease irreversibly changes.
  - Disadvantages
    - (Much) lower statistical power than X-over for the same sampe size.
    - Phenotyping mandatory for drugs showing polymorphism.

Parallel Groups: Example

- One group is treated with the test formulation and another group with reference.
- Quite common that the dataset is imbalanced, *i.e.*, $n_1 \neq n_2$.
- FDA guidance against the assumption of equal variance. Not implemented in PK soft-ware (Phoenix/WNL, Kinetica)!

Subj.	Group 1 (T)	Group 2 (R)
1-13	100	110
2-14	103	113
3-15	80	96
4-16	110	90
5-17	78	111
6-18	87	68
7-19	116	111
8-20	99	93
9-21	122	93
10-22	82	82
11-23	68	96
12-24	NA	137
n	11	12
mean	95	100
s^2	298	314
s	17.3	17.7

Parallel Groups: Example

- Pooled variance

$$s_0^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{10 \cdot 298 + 11 \cdot 314}{10 + 11 - 2} = 306.4$$

- Pooled standard deviation

$$s_0 = \sqrt{s_0^2} = \sqrt{306.4} = 17.50$$

- 90% Confidence interval

$$CI = \left|\overline{x}_1 - \overline{x}_2\right| \pm t_{2\alpha, n_1+n_2-2} s_0 \sqrt{\frac{n_1 + n_2}{n_1 n_2}} =$$

$$= 5 \pm 1.721 \cdot 17.50 \cdot 0.4174 = [-7.6, +17.6]$$

Parallel Groups: Example

- But we want a ratio, not a difference!
  Now we have only –7.6 ≤ [T-R = -5] ≤ +17.6…
- Maybe we can use (R–7.6)/R and (R+17.6)/R to get a CI of 92.4% – 117.6%?
- No. Let's repeat the analysis with logtransformed data.

Parallel Groups: Example

Subj.	Group 1 (T)	ln (T)	Group 2 (R)	ln (R)
1-13	100	4.605	110	4.700
2-14	103	4.635	113	4.727
3-15	80	4.382	96	4.564
4-16	110	4.700	90	4.500
5-17	78	4.357	111	4.710
6-18	87	4.466	68	4.220
7-19	116	4.754	111	4.710
8-20	99	4.595	93	4.533
9-21	122	4.804	93	4.533
10-22	82	4.407	82	4.407
11-23	68	4.220	96	4.564
12-24	NA	NA	137	4.920
n	11	11	12	12
mean	95	4.539	100	4.591
s^2	298	0.03418	314	0.03231
s	17.3	0.1849	17.7	0.1798

$$s_0^2 = \frac{10 \cdot 0.03418 + 11 \cdot 0.03231}{10 + 11 - 2} =$$

$$= 0.03320$$

$$s_0 = \sqrt{s_0^2} = \sqrt{0.03320} = 0.1812$$

$$CI_{\ln} = 0.05203 \pm 1.721 \cdot 0.1822 \cdot 0.4174 =$$

$$= [-0.1829, +0.07886]$$

$$CI = e^{[-0.1829, +0.07886]} = [83.28\%, 108.20\%]$$

Parallel Groups: Example

- Not finished yet …
- Analysis flawed* (assumes equal variances; against FDA's guidance)!
- Degrees of freedom for the *t*-value have to be modified, *e.g.*, by the Welch-Satterthwaite approximation:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1+1)} + \frac{s_2^4}{n_2^2(n_2+1)}}$$

* **Moser BK and GR Stevens**
Homogeneity of variance in the two-sample means test
Amer Statist 46:19-21 (1992)

Parallel Groups: Example

- Instead of the simple $\nu = n_1 + n_2 - 2 = 21$, we get

$$\nu = \frac{\left(\frac{0.03418}{11} + \frac{0.03231}{12}\right)^2}{\frac{0.001169}{121 \cdot 12} + \frac{0.001044}{144 \cdot 13}} = 20.705$$

- Maybe it's time to leave M$-Excel.
- Easy to calculate in R.

Parallel Groups: Example

```
T <- c(100,103,80,110,78,87,116,99,
       122,82,68)
R <- c(110,113,96,90,111,68,111,93,
       93,82,96,137)
par.equal1 <- t.test(log(R), log(T),
  alternative="two.sided", mu=0,
  paired=FALSE, var.equal=TRUE,
  conf.level=0.90)
par.equal1
 Two Sample t-test

data:  log(T) and log(R)
t = 0.684, df = 21, p-value = 0.5015
alternative hypothesis: true
difference in means is not equal to 0
90 percent confidence interval:
 -0.1829099  0.0788571
sample estimates:
mean of x mean of y
 4.538544  4.590570
round(100*exp(par.equal1$conf.int),
digits=2)
  83.28 108.20
```

liberal!

```
T <- c(100,103,80,110,78,87,116,99,
       122,82,68)
R <- c(110,113,96,90,111,68,111,93,
       93,82,96,137)
par.equal0 <- t.test(log(R), log(T),
alternative="two.sided", mu=0,
paired=FALSE, var.equal=FALSE,
conf.level=0.90)
par.equal0
  Welch Two Sample t-test

data:  log(T) and log(R)
t = 0.6831, df = 20.705, p-value = 0.5021
alternative hypothesis: true difference
in means is not equal to 0
90 percent confidence interval:
 -0.18316379  0.07911102
sample estimates:
mean of x mean of y
 4.538544  4.590570
round(100*exp(par.equal0$conf.int),
digits=2)
 83.26 108.23
```

Parallel Groups

- Design Issues
  - EMEA NfG on BA/BE (2001)
    - 3.2.4 Genetic phenotyping
      'Phenotyping and/or genotyping of subjects should be considered for […] all studies using parallel group design. If a drug is known to be subject to major genetic polymorphism, studies could be performed in panels of subjects of known phenotype or genotype for the polymorphism in question.'
  - Since the comparison is based on *intra-subject* effects
    - One study of the major phenotype/genotype.
    - Two studies of the respective phenotype/genotype – only if requested!

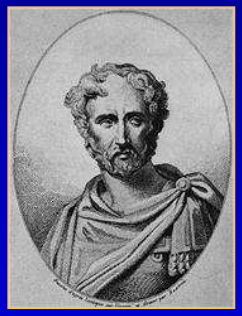
Parallel Groups

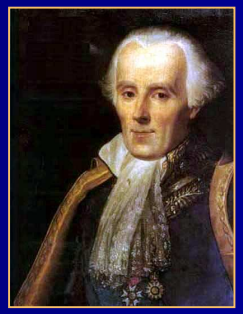
- Design Issues
  - EMA GL on BE (2010)
    - 4.1.3 Subjects / Selction of Subjects
      ‘Phenotyping and/or genotyping of subjects may be considered for safety or pharmacokinetic reasons.’
  - Wording has changed since old NfG, but content stayed the same!
  - Specifically not only for parallel designs!

To bear in Remembrance...

In these matters the only certainty is
that nothing is certain.

Gaius Plinius Secundus (Pliny the Elder)



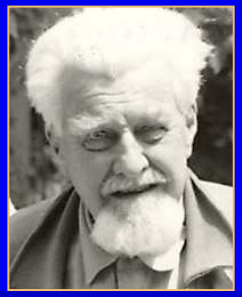


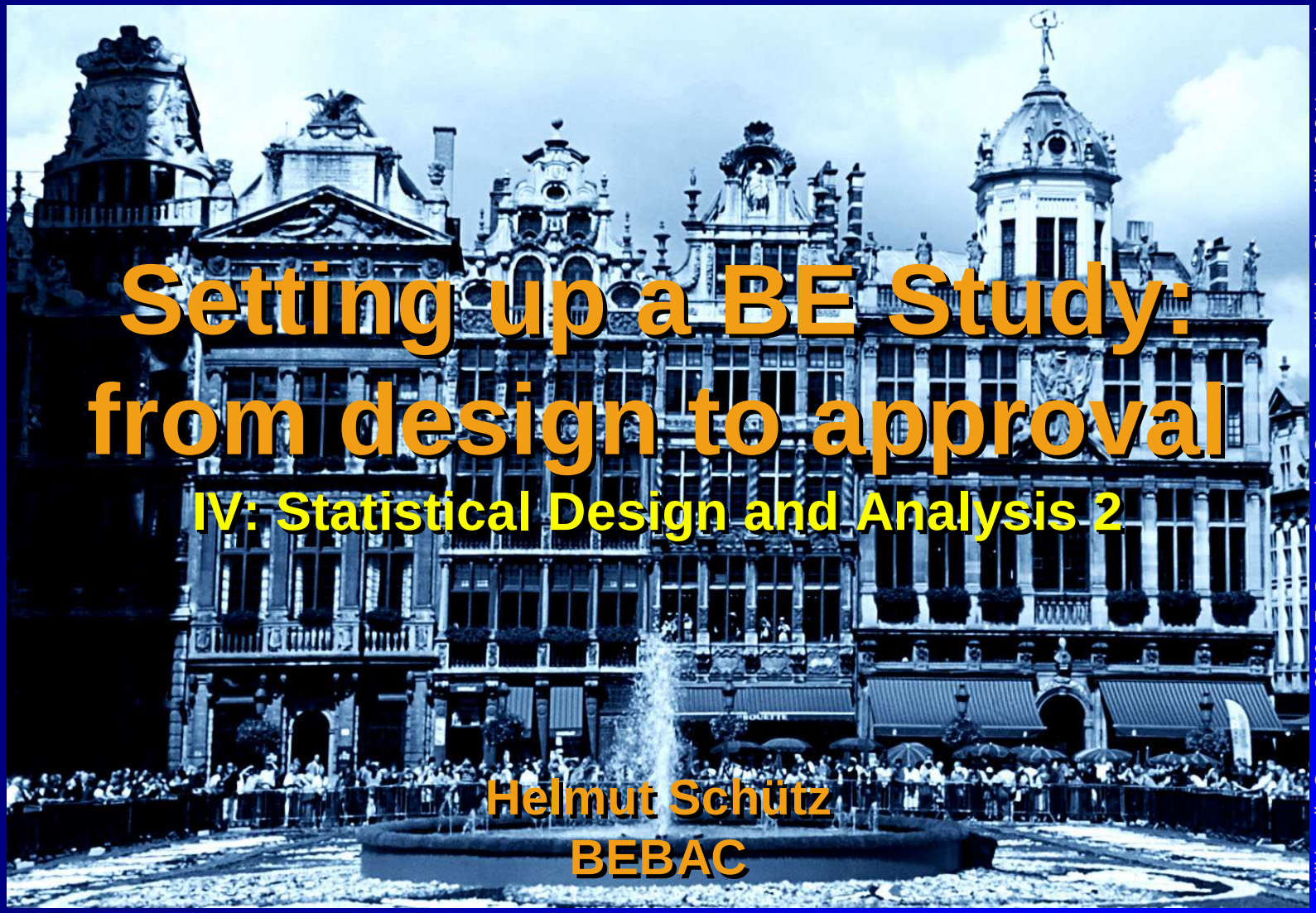
The theory of probabilities is at bottom
nothing but common sense reduced to calculus.

Pierre-Simon Laplace

It is a good morning exercise for a research scientist
to discard a pet hypothesis every day before
breakfast.
It keeps him young.

Konrad Lorenz





Setting up a BE Study: from design to approval

IV: Statistical Design and Analysis 2

Helmut Schütz
BEBAC

Pitfalls

- Pilot studies
- Sample size estimation
- Low variability
- Metrics of early exposure
- Highy variable drugs / drugs products
- Two-stage sequential designs

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

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

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

Pilot Studies: Sample Size

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

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

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.

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!

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)

Published data

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

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

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

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

    - 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)

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

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

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

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

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

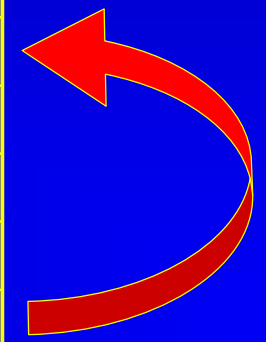
Sensitivity to Imbalance

- 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

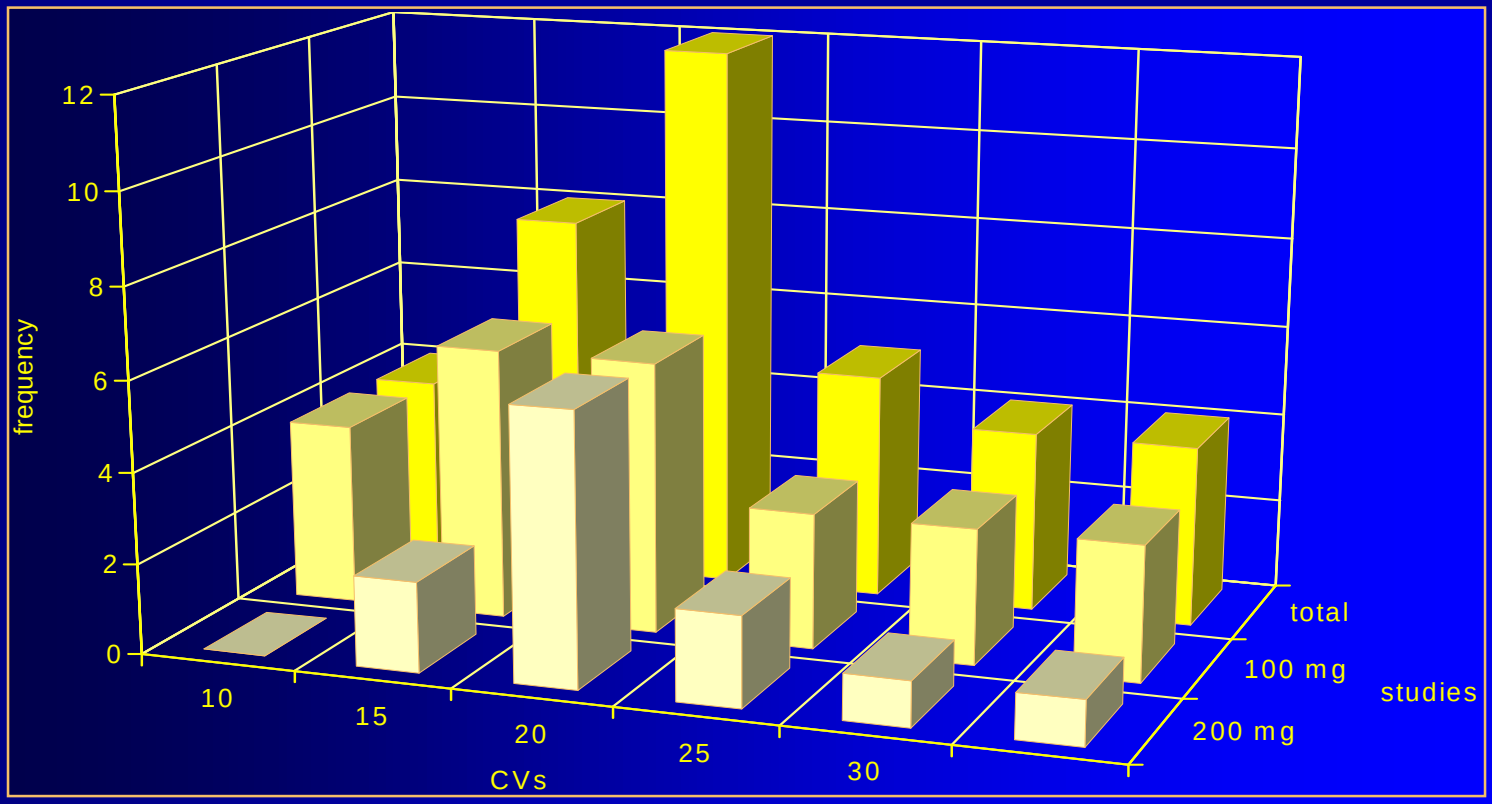


No Algebra...

- 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
```

Literature data



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

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’ Japan

Sample Size (Limits)

- Maximum
  - NZ: If the calculated number of subjects appears to be higher than is ethically justifiable, it may be necessary to accept a statistical power which is less than desirable. Normally it is not practical to use more than about 40 subjects in a bioavailability study.
  - All others: Not specified (judged by IEC/IRB or local Authorities).
    ICH E9, Section 3.5 applies: 'The number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed.'

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.

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

Cookbook?

Hierarchy of Designs

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

Full replicate (TRTR | RTRT)

Partial replicate (TRR | RTR | RRT)

Standard 2×2 cross-over (RT | RT)

Parallel (R | T)

  - Variances which can be estimated:

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

Coefficient(s) of Variation

- From any design one gets variances of *lower* design levels (only!)
  - Example: Total CV% from a 2×2 cross-over used in planning a parallel design study
    - Intra-subject CV% (within) → $CV_{intra}\% = 100 \cdot \sqrt{e^{MSE_W} - 1}$
    - Inter-subject CV% (between) → $CV_{inter}\% = 100 \cdot \sqrt{e^{\frac{MSE_B - MSE_W}{2}} - 1}$
    - Total CV% (pooled) → $CV_{total}\% = 100 \cdot \sqrt{e^{\frac{MSE_B + MSE_W}{2}} - 1}$

Hauschke D, Steinijans VW and E Diletti
Presentation of the intrasubject coefficient of variation for sample size planning in bioequivalence studies
Int J Clin Pharmacol Ther 32/7, 376-378 (1994)

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

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 $(1-\alpha)$ % 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	$1-\alpha$	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	$\sigma^2_w \times df$	$CV_{intra\ /\ pooled}$	$>CL_{upper}$
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	$1-\alpha$	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
```

α- vs. β-Error

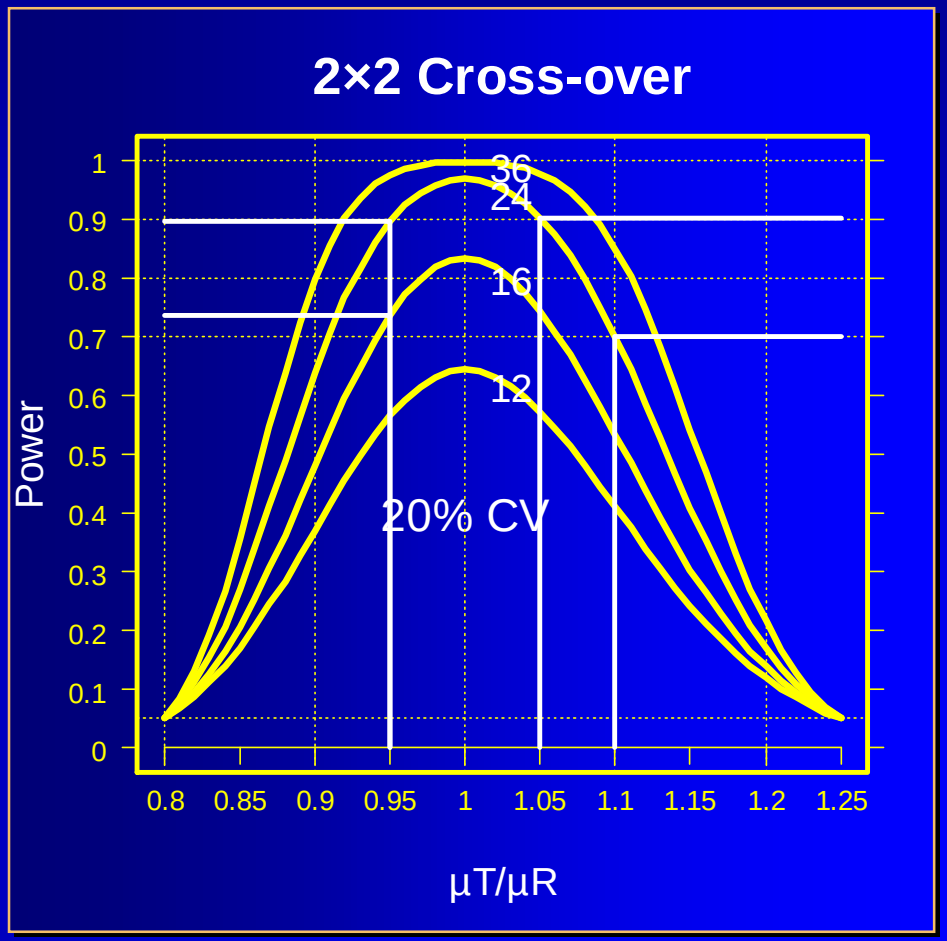
- α-Error: Patient's risk to be treated with a bioinequivalent formulation.
  - Although α is generally set to 0.05, sometimes $<$0.05 (*e.g.*, NTDIs in Brazil, multiplicity, interim analyses).
- β-Error: Producer's risk to get no approval for a bioequivalent formulation.
  - Generally *set* in study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

Either **a study demonstrated BE** ***or*** **not.**

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



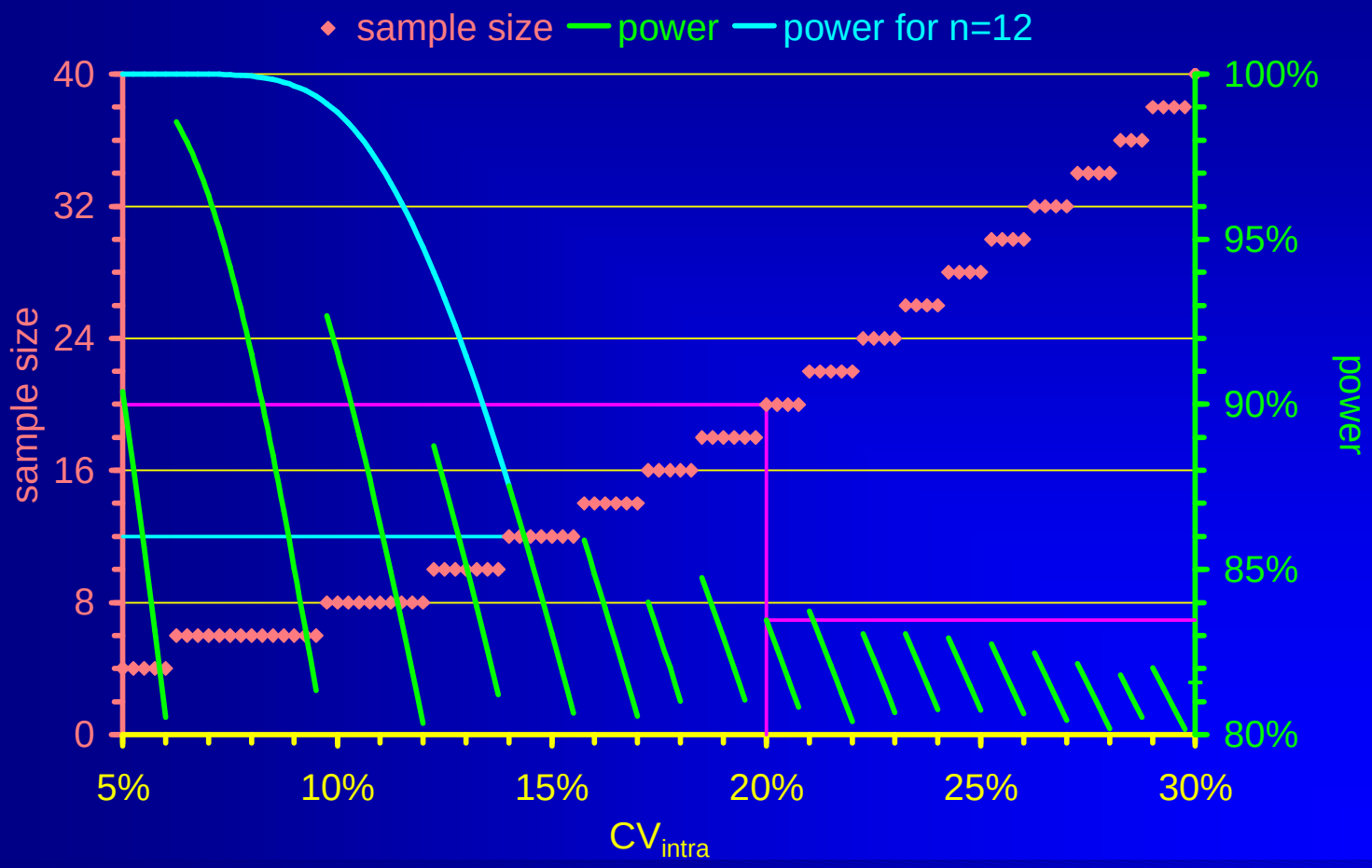
Power *vs.* Sample Size

- 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% $\rightarrow$ 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, 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)

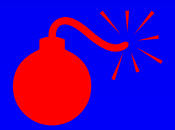
Patient's risk α 0.05, Power 80% (Producer's risk β 0.2), AR [0.80 – 1.25], CV 0.2 (20%), T/R 0.95

1. $\Delta = \ln(0.8)-\ln(T/R) = -0.1719$
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

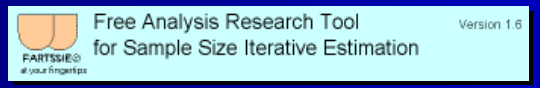
1. $\Delta = \ln(T/R) - \ln(1.25) = 0.1719$
2. Start with e.g. n=8/sequence
   1. $df_\alpha = \text{roundup}(2 \cdot n-2) \cdot 2-2 = (2 \times 8-2) \times 2-2 = 26$
   2. $df_\beta = \text{roundup}(4 \cdot n-2) = 4 \times 8-2 = 30$
   3. $t_{\alpha,df} = 1.7056$
   4. $t_{\beta/2,df} = 0.8538$
   5.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

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

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

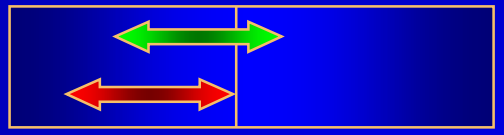
There is simple intuition behind results like these: If my car made it to the top of the hill, then it is powerful enough to climb that hill; if it didn't, then it obviously isn't powerful enough. Retrospective power is an obvious answer to a rather uninteresting question. A more meaningful question 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

Low Variability

- Drugs / Drug Products with CV_{intra} <10%
  - No specific statements in any guideline.
  - Problems may arise according to significant treatment effects in ANOVA (*i.e.*, although the 90% CI is within the acceptance range – 100% is not included) – even for the minimum sample size of 12.



  - Denmark
    - DKMA considers that the 90% CI for the ratio test versus reference should include 100% […].
    - Deviations are usually accepted if it can be adequately proved that the deviation has no clinically relevant impact on the efficacy and safety of the medicinal product.
      Danish Medicines Agency (DKMA)
      Bioequivalence and labelling of medicinal products with regard to generic substitution (13 Jul 2011)
      http://www.dkma.dk/1024/visUKLSArtikel.asp?artikelID=6437

Early Exposure

- Partial AUCs for Rapid Onset Drugs
  - US-FDA 2003 (III.A.8.a.)
    - […] that the partial area be truncated at the *population median of T_{max} values for the reference formulation*. We also recommend that at least two quantifiable samples be collected before the expected peak time to allow adequate estimation of the partial area.
  - Canada-TGD 2005
    - […] $AUC_{Reftmax}$ for a test product is defined as the area under the curve to the time of the maximum concentration of the reference product, *calculated for each study subject*.

Early Exposure (HVDP?)

- Partial AUCs for Rapid Onset Drugs (cont'd)

Example	median t_{maxref}	PE	nonparametric CI		BE	FDA	parametric CI		BE	TGD	BE
1	1.5 h	±0.00 h	-0.25 h (85%)	+0.25 h (115%)	yes	90.1%	75.0%	110.1%	no (CV 26.4%)	85.7%	yes
2	1.5 h	+0.26 h	±0.00 h (100%)	+0.50 h (130%)	no	66.1%	53.1%	82.0%	no (CV 29.7%)	62.4%	no

  - Even for formulations with *low* intra-subject variability…
    - Example 1: AUC_t 13.3% C_{max} 17.0%
    - Example 2: AUC_t 6.33% C_{max} 9.43%
  - …it is unlikely to be able to demonstrate BE due to high variability of this metric.

Early Exposure

- EU GL 2010 (Section 4.1.8)
  - A statistical evaluation of t_{max} is not required. However, if rapid release is claimed to be clinically relevant and of importance for onset of action or is related to adverse events, there should be no apparent difference in median t_{max} and its variability between test and reference product.

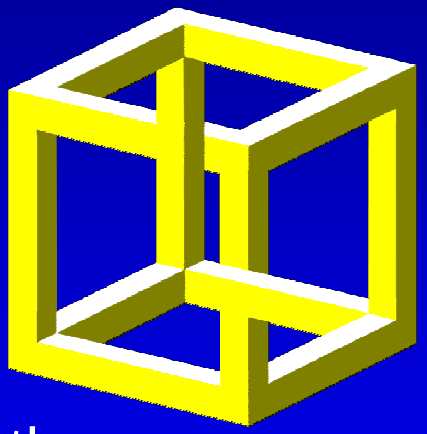
How to assess that?
'A non-parametric analysis is not acceptable.'

Highly Variable Drugs / Drug Products

- HVDs / HVDPs (intra-subject variability >30 %)
  - ✓USA Replicate Design recommended in product specific guidances: Minimum number of subjects (24?), restriction on GMR (0.80 – 1.25).
  - ± EU Widening of acceptance range (for C_{max} only: to maximum 69.84% – 143.19%), if CV_{WR} in the study >30%. Restriction on GMR (0.80 – 1.25).

HVDs/HVDPs

- Ways out?
  - Nonparametric methods
    'A non-parametric analysis is not acceptable.' (BE GL, Section 4.1.8)
  - Compartmental methods (Population PK)
    'The use of compartmental methods for the estimation of parameters is not acceptable.' (BE GL, Section 4.1.5)
  - Replicate designs could be considered e.g. for substances with highly variable pharmacokinetic characteristics. (EU BE GL, Section 4.1.1, 4.1.10)



HVDs/HVDPs

- All (!) ANDAs submitted to FDA/OGD 2003–2005 (1010 studies, 180 drugs)
  - 31% (57/180) highly variable (CV ≥30%)
  - of these HVDs/HVDPs,
    - 60% due to PK (*e.g.*, first pass metabol.)
    - 20% formulation performance
    - 20% unclear

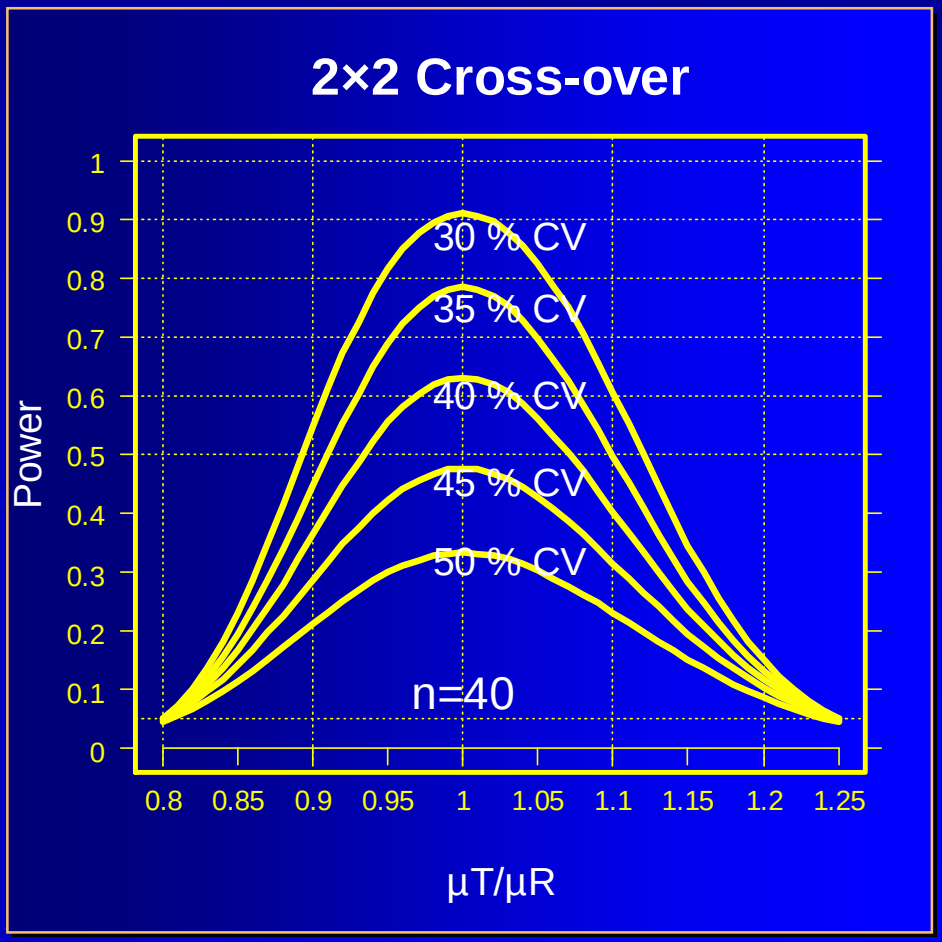
Davit BM, Conner DP, Fabian-Fritsch B, Haidar SH, Jiang X, Patel DT, Seo PR, Suh K, Thompson CL, and LX Yu
Highly variable drugs: observations from bioequivalence data submitted to the FDA for new generic drug applications
AAPS J 10(1): 148-56 (2008)

HVDs/HVDPs

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

$\mu T/\mu R$ 0.95, CV_{intra} 30%
→ power 0.816
$\mu T/\mu R$ 1.00, CV_{intra} 45%
→ power 0.476 <
Roulette 0.486 (!)

$\mu T/\mu R$ 0.95, CV_{intra} 45%
→ n=82 (power 0.807)



HVDs/HVDPs (US/EU)

- Advisory Committee for Pharmaceutical Sciences (ACPS) to FDA (10/2006) on HVDs
- Follow-up papers in 2008 (ref. in API-GLs)
  - Partial replicate study design [TRR | RTR | RRT]
  - Reference Scaled Average Bioequivalence (RSABE)
  - Minimum sample size 36 (?) subjects
  - Point estimate restricted to [0.80,1.25]

Haidar SH, Davit B, Chen M-L, Conner D, Lee LM, Li QH, Lionberger R, Makhlouf F, Patel D, Schuirmann DJ, and LX Yu
Bioequivalence Approaches for Highly Variable Drugs and Drug Products
Pharmaceutical Research 25/1, 237-241 (2008)
http://www.springerlink.com/content/u503p62056413677/fulltext.pdf
Haidar SH, Makhlouf F, Schuirmann DJ, Hyslop T, Davit B, Conner D, and LX Yu
Evaluation of a Scaling Approach for the Bioequivalence of Highly Variable Drugs
The AAPS Journal, 10/3, (2008) DOI: 10.1208/s12248-008-9053-4

HVDs/HVDPs

- Is suggested EU-method of any good?
  - Replicate designs *without scaling* (AUC)
    - reduce the number of subjects (to 75% for a 3-period design and to 50% for a 4-period design as compared to a conventional 2×2),
    - while keeping the *theoretical* number of treatments constant:
      - The potentional drop-out rate increases.
      - Practically *more* treatments must be administered in order to maintain the desired power!

HVDs/HVDPs

- Example
  - AR [0.80,1.25], CV_{intra} 49.5%, T/R 0.95%, power 80% ($n_{2\times2}$ 96, $n_{4\times2}$ 48)
  - Expected dropout rate of 5% / washout
    - 2×2 study: 96+6=102 subjects (199 treatments)
    - 4×2 study: 48+10=58 subjects (214 treatments)

58 → 55 → 52 → 49
(5.2%, 5.5%, 5.8%)

56 → 53 → 50 → 48
(5.4%, 5.7%, 4.0%)

Ethical?

  - Proposed FDA Scaling-Method: AR [0.7006,1.4273], PE [0.80,1.25], n 34 (!)

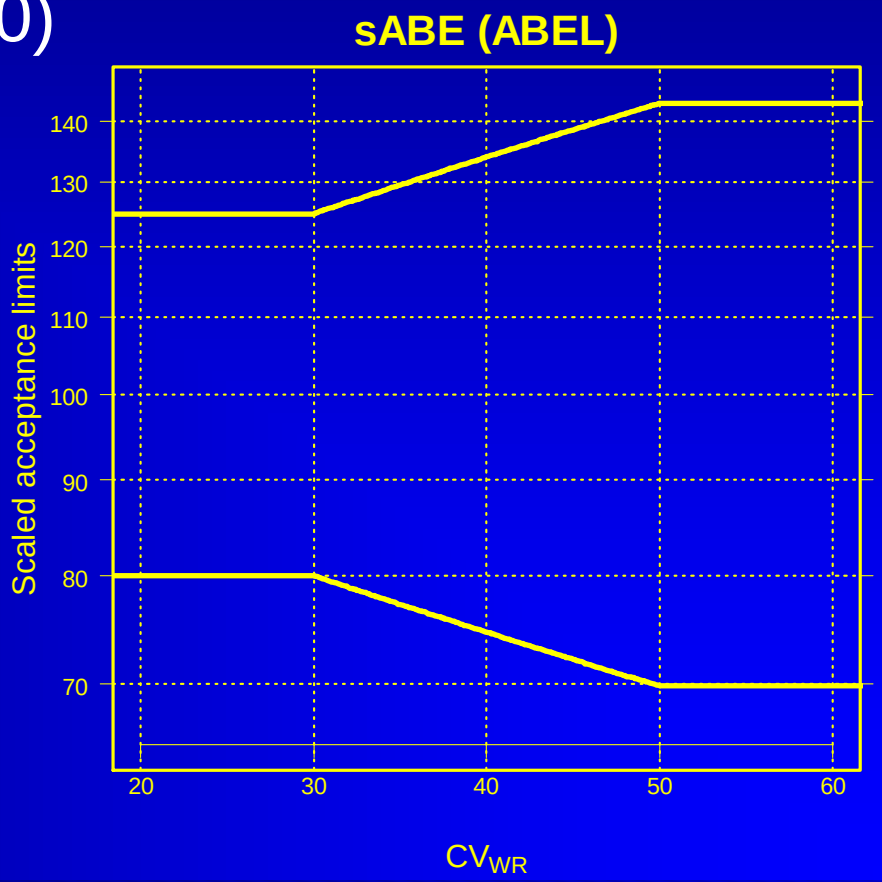
Highly Variable Drugs / Drug Products

- EU GL on BE (2010)
  - Scaling allowed for C_{max} only (not AUC!) based on CV_{WR} >30% in the study.
  - Limited to a maximum of CV_{WR} 50% (*i.e.*, higher CVs are treated *as if* CV = 50%).
  - PE restricted with 80% – 125% in any case.
  - No commercial software for sample size estimation can handle the PE restriction.
  - Monte Carlo simulations necessary.

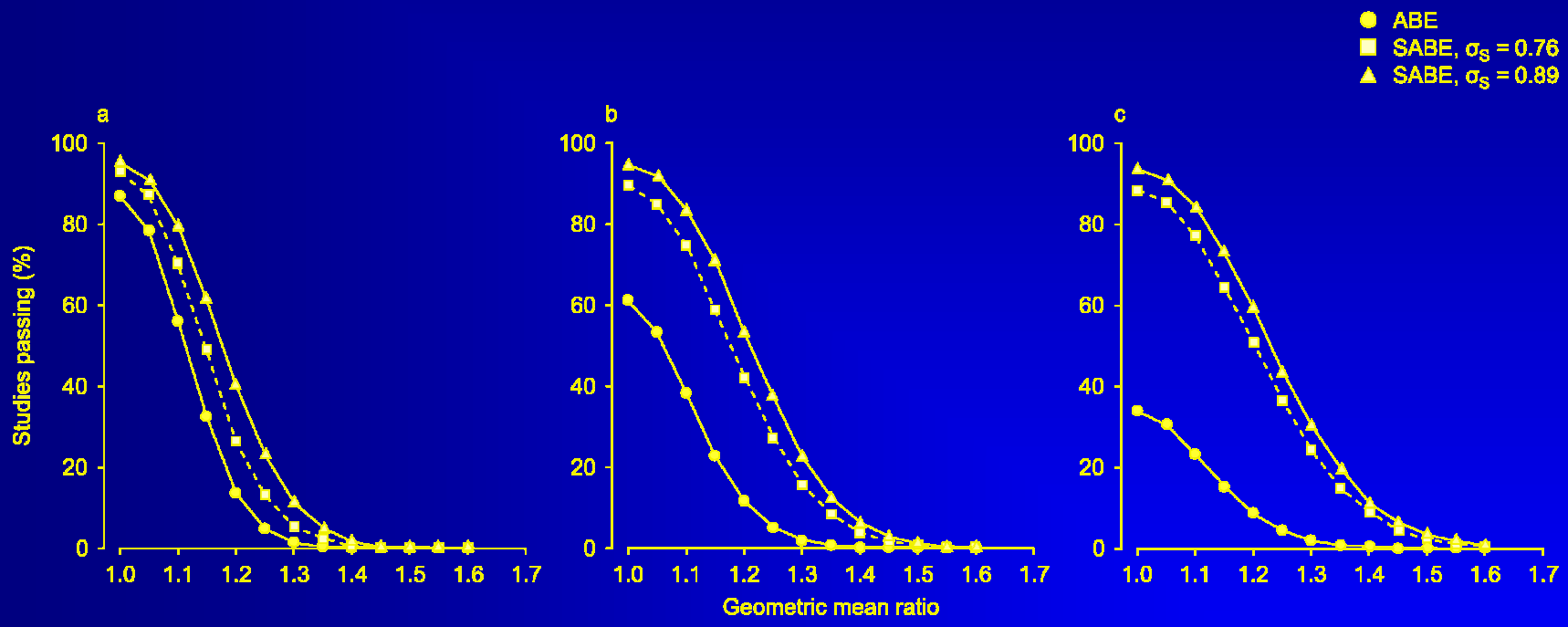
HVDs/HVDPs

- 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



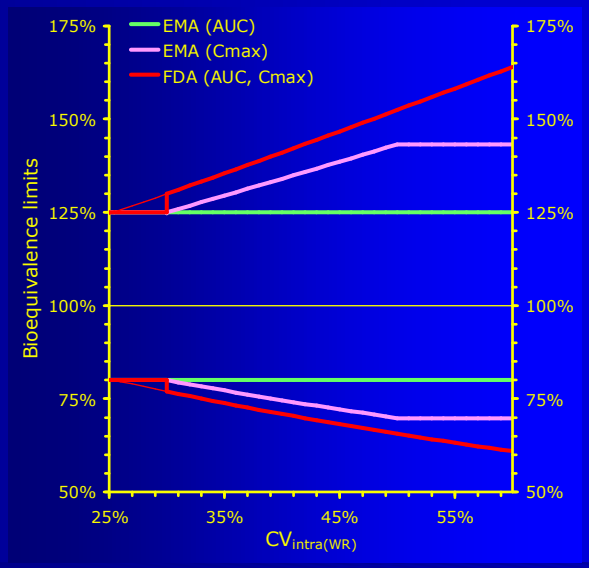
HVDs/HVDPs

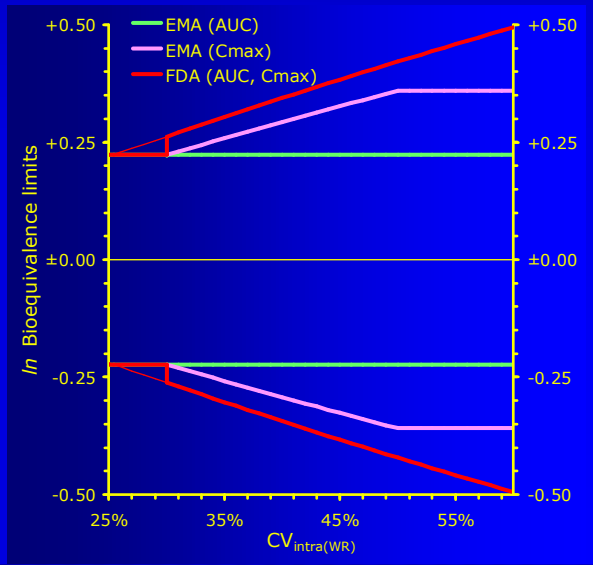


Totfalushi *et al.* (2009), Fig. 3
Simulated (n=10000) three-period replicate design studies (TRT-RTR) in 36 subjects; GMR restriction 0.80–1.25. (a) CV=35%, (b) CV=45%, (c) CV=55%.
ABE: Conventional Average Bioequivalence, SABE: Scaled Average Bioequivalence, 0.76: EU criterion, 0.89: FDA criterion.

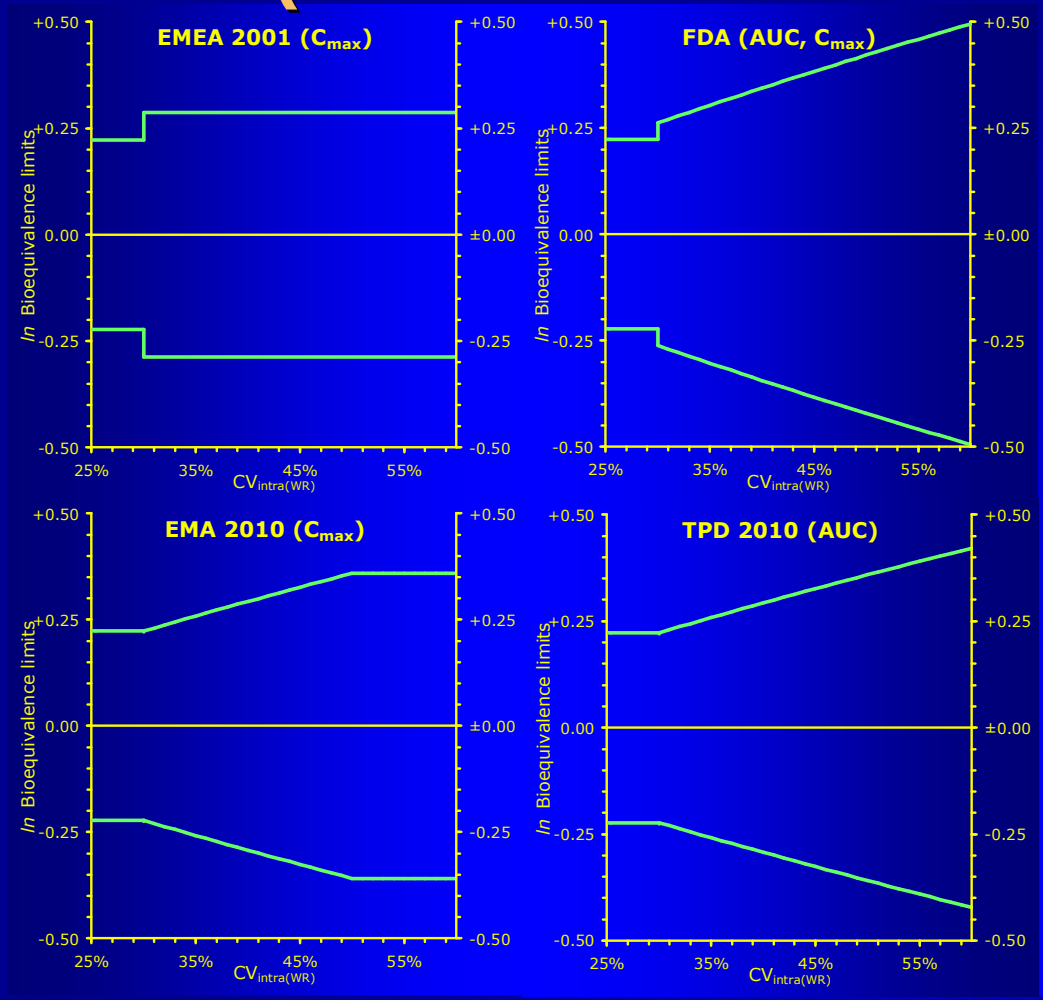
HVDPs (US/EU)

- FDA's and EMA's approaches differ; FDA's leads to a discontinuity of the acceptance range at CV=30% because FDA's scaling CV is 25.396% (σ_{WR} 0.25) – but *applied* at CV >30%.





HVDPs (Global Harmonization?)



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) is identical to the concentional 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.

HVDs/HVDPs

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

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

and

$$\theta_s = \frac{\ln(1.25)}{\sigma_0} = -\frac{\ln(0.80)}{\sigma_0} \cong 0.760$$

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)

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 (0.760):

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

EMA Example (ABEL)

- Data set I: 2-Sequence Full Replicate Design (RTRT | TRTR), *imbalanced*
  (n=77: 4 periods, n=69: 3 periods, n=6: 2 periods)
  Method B

```
proc mixed data=replicate;
class formulation subject period sequence;
model logDATA= sequence period formulation;
random subject(sequence);
estimate "test-ref" formulation -1 1 / CL alpha=0.10;
run;
```

EMA, Committe Human Medicinal Products (CHMP),
CHMP Pharmacokinetics Working Party (PKWP)
Questions & Answers: Positions on specific questions addressed to the Pharmacokinetics Working Party; Clarification on the recommended statistical method for the analysis of a bioequivalence study
EMA/618604/2008 Rev. 3, London, 26 January 2011
http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002963.pdf

EMA Example

- Data set I

Test data discarded for calculation of CV_{WR}

```
data var;
set replicate;
if formulation='R';
run;

proc glm data=var;
class subject period sequence;
model logDATA= sequence subject(sequence) period;
run;
```

EMA Example

- Evaluation with Phoenix/WinNonlin 6.2
  - Calculation of the scaled acceptance range [L,U] based on the limiting CV_{WR} and the regulatory constant k (0.760).

$$CV_{WR} = 100\sqrt{e^{\sigma^2_{WR}} - 1} \qquad [L,U] = e^{\mp k \cdot \sigma_{WR}}$$

Dependent	Parameter	Estimate	CVWR	L	U	Diff_to_detect
logData	Var(Residual)	0.1993136	46.96	71.23	140.40	28.77

σ^2_{WR}	0.1993136
CV_{WR}	46.96
L	71.23
U	140.40

Scaling applicable since 30% < CV_{WR} ≤ 50%

Helmut Schütz
Evaluation of Replicate Designs for Average Bioequivalence according to EMA's Guideline with Phoenix™ WinNonlin® (2011 Pharsight, A Certara Company, Tripos L.P.)
Vienna, April 2011
http://bebac.at/downloads/Replicate%20Designs%20for%20ABE%20according%20to%20EMA%20with%20Phoenix%20v2.3.pdf

EMA Example

```
Bioequivalence Statistics

User-Specified Confidence Level for CI's = 90.0000
Percent of Reference to Detect for 2-1 Tests = 20.0%
A.H.Lower =  0.800    A.H.Upper =  1.250

Formulation variable: Formulation
Reference: R    LSMean=    7.670014   SE=    0.101295   GeoLSM=    2143.110761
-----------------------------------------------------------------------------
Test:      T    LSMean=    7.816102   SE=    0.101395   GeoLSM=    2480.218425

    Difference =       0.1461,  Diff_SE=  0.0465,    df= 216.9
    Ratio(%Ref) =    115.7298

    CI  90% = (  107.1689,  124.9746)

    Average bioequivalence shown for confidence=90.00 and percent=20.0.
```

ABE

107.17 – 124.97

passed 80 – 125

passed 75 – 133

EMA Example

```
Bioequivalence Statistics

User-Specified Confidence Level for CI's = 90.0000
Percent of Reference to Detect for 2-1 Tests = 28.77%
A.H.Lower =  0.712    A.H.Upper =  1.404

Formulation variable: Formulation
Reference: R    LSMean=    7.670014   SE=    0.101295   GeoLSM=    2143.110761
-------------------------------------------------------------------------------
Test:      T    LSMean=    7.816102   SE=    0.101395   GeoLSM=    2480.218425

    Difference =      0.1461,  Diff_SE=  0.0465,   df= 216.9
    Ratio(%Ref) =   115.7298

    CI  90% = (  107.1689,  124.9746)

    Average bioequivalence shown for confidence=90.00 and percent=28.77.
```

ABEL

107.17 – 124.97

passed 71.23 – 140.40

PE 115.73

within 80.00 – 125.00

EMA Example

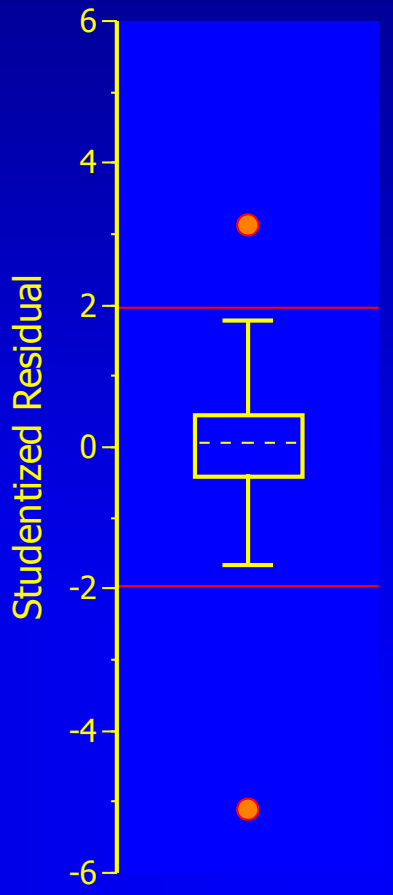
- Outliers?
  - GL 2010, Section 4.1.10: *'The applicant should justify that the calculated intra-subject variability is a reliable estimate and that it is not the result of outliers.'*
  - Boxplots were discussed at the EGA-workshop 2010: *'The outlier cannot be removed from evaluation but should not be taken into account for calculation of within-subject variabi-lity and extension of the acceptance range. An outlier test is not an expectation of the medicines agencies but outliers could be shown by a box plot. This would allow the medicines agencies to compare the data between them.'*

European Generic Medicines Association (EGA)
Revised EMA Bioequivalence Guideline, Questions & Answers
London, June 2010
http://www.egagenerics.com/doc/EGA_BEQ_Q&A_WEB_QA_1_32.pdf

EMA Example

- Outliers
  - Data set I: Based on studentized intra-subject residuals two severe outliers (outside ±3×IQR) are detected
  - If these two outliers are excluded from the calculation of CV_{WR}, scaling almost useless!

	n=77	n=75
σ^2_{WR}	0.1993136	0.0984319
CV_{WR}	46.96	32.16
L	71.23	78.79
U	140.40	126.93



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.

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 & 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)

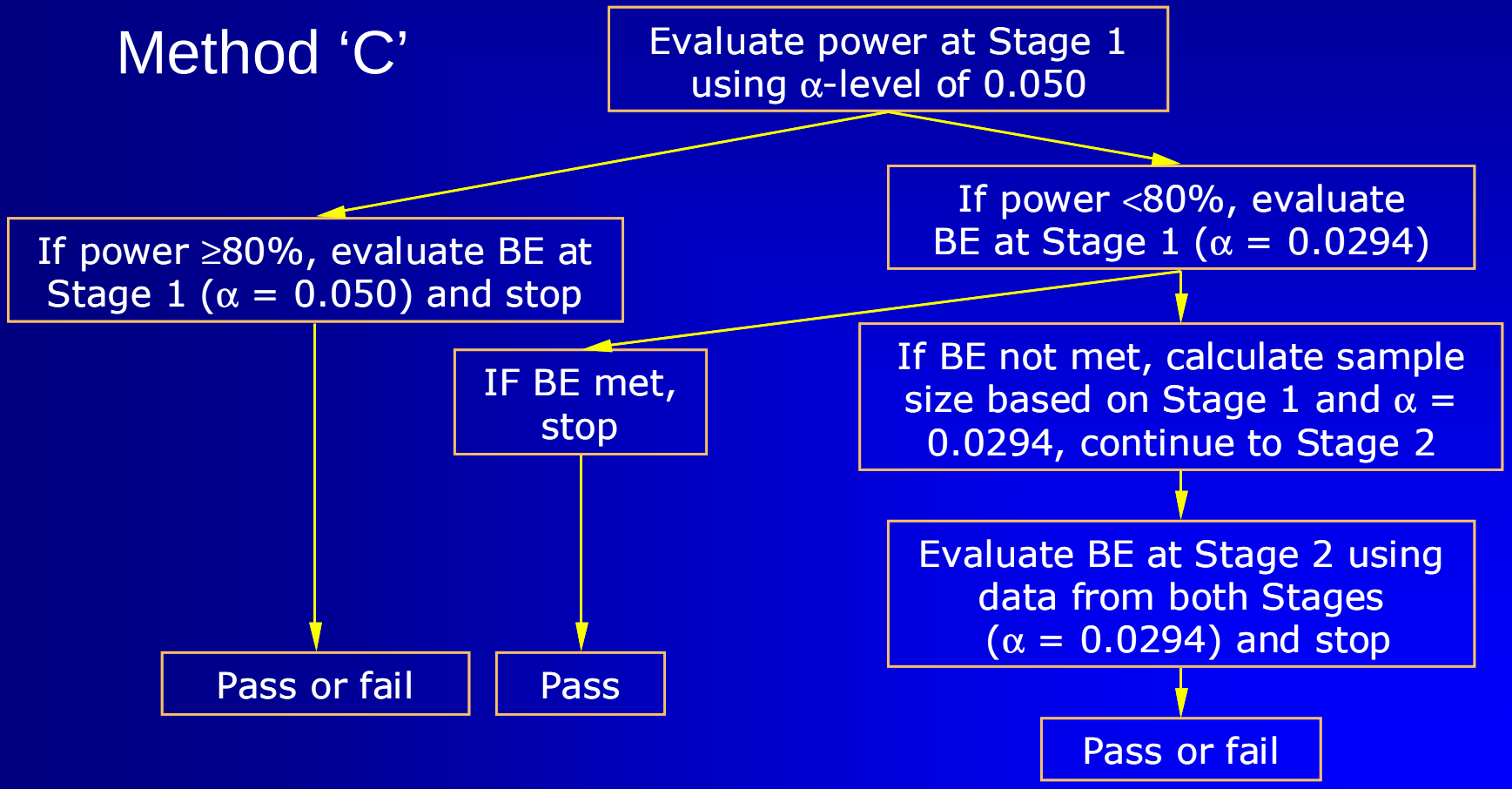
Sequential Designs

- Methods by Potvin *et al.* (2008) promising
  - Supported by 'The Product Quality Research Institute' (members: FDA/CDER, Health Canada, USP, AAPS, PhRMA, …)
    - Acceptable 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'



Sequential Designs

- Methods by Potvin *et al.* (2008) limited to point estimate of 0.95 and 80% power
  - Follow-up paper in 2011
    - Slight inflation of patient's risk (α 0.0547) observed in Methods B/C if PE 0.90 instead of 0.95 was used
    - New Method D (α 0.0280)
    - 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

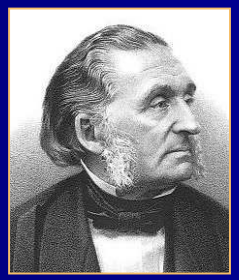
To bear in Remembrance...

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

Congratulations!
Setting up a BE Study: from design to approval
Open Questions?

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