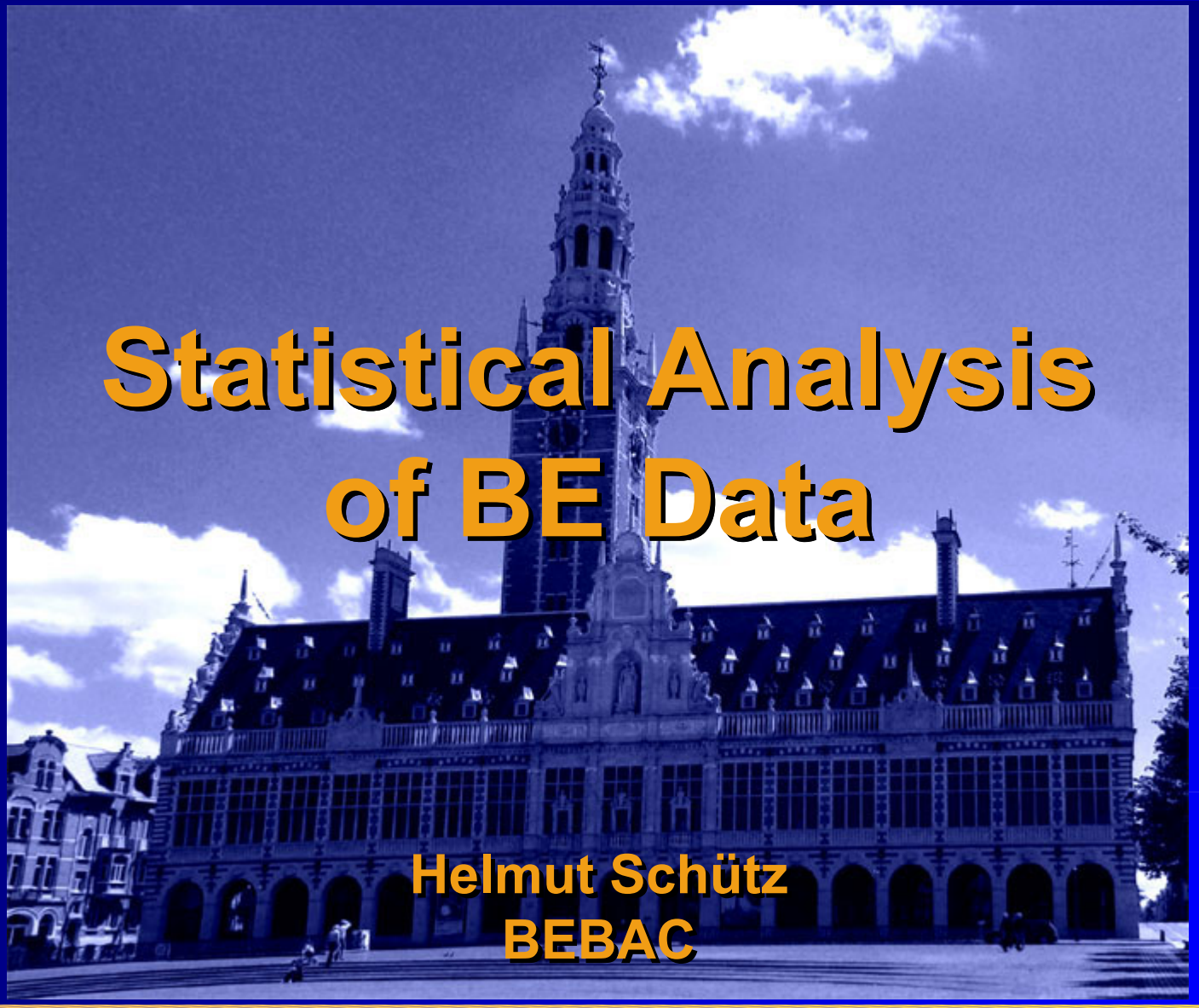
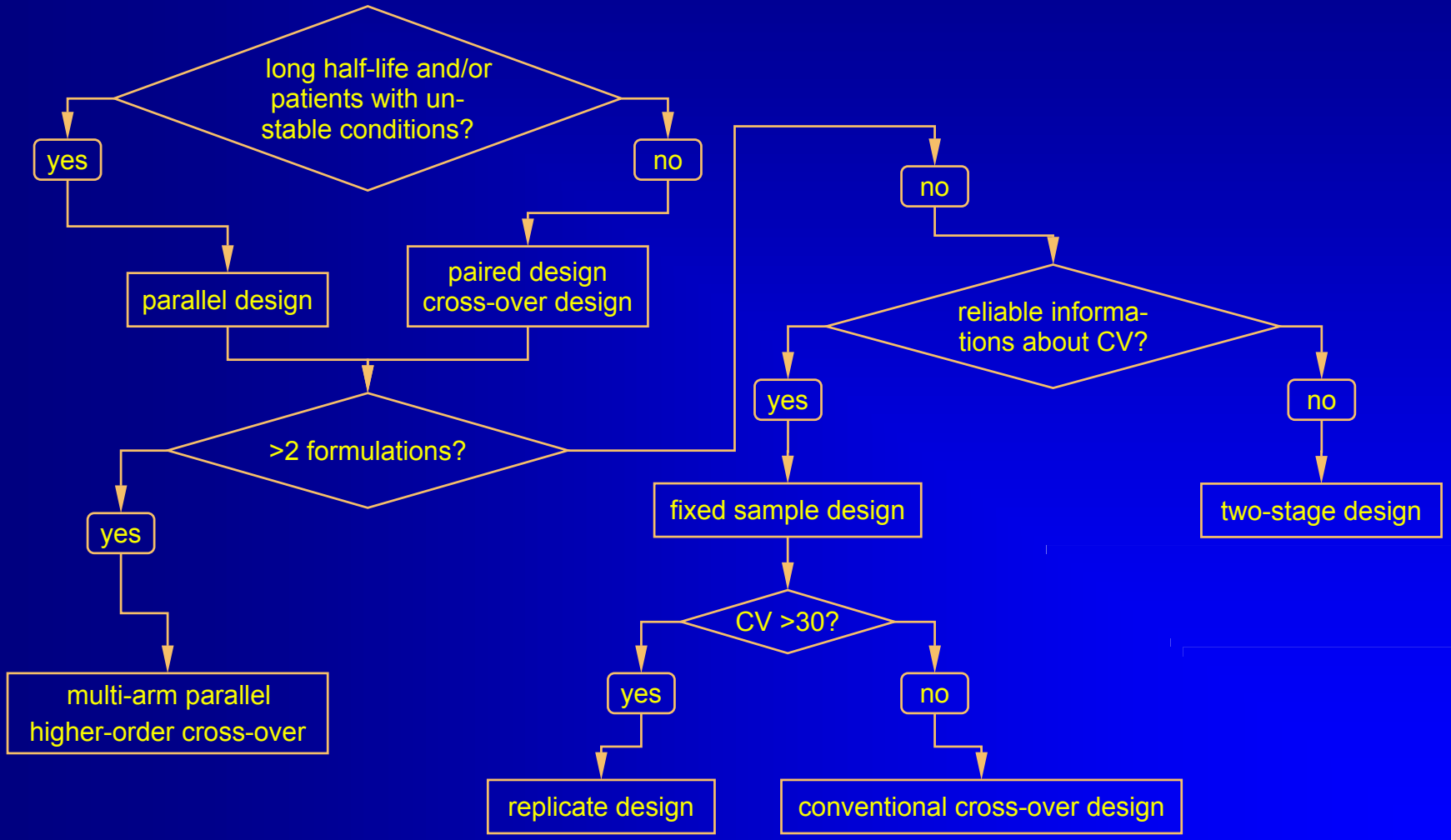


Statistical Analysis of BE Data

Helmut Schütz
BEBAC

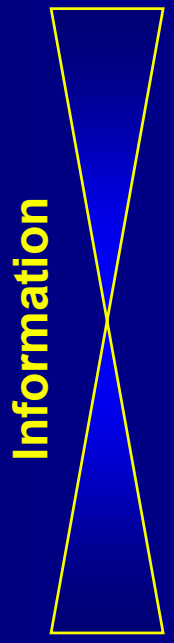
Designs


long half-life and/or patients with un-stable conditions?
yes
no
parallel design
paired design cross-over design
>2 formulations?
yes
multi-arm parallel higher-order cross-over
no
reliable informa-tions about CV?
yes
no
fixed sample design
two-stage design
CV >30?
yes
no
replicate design
conventional cross-over design

Designs

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

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



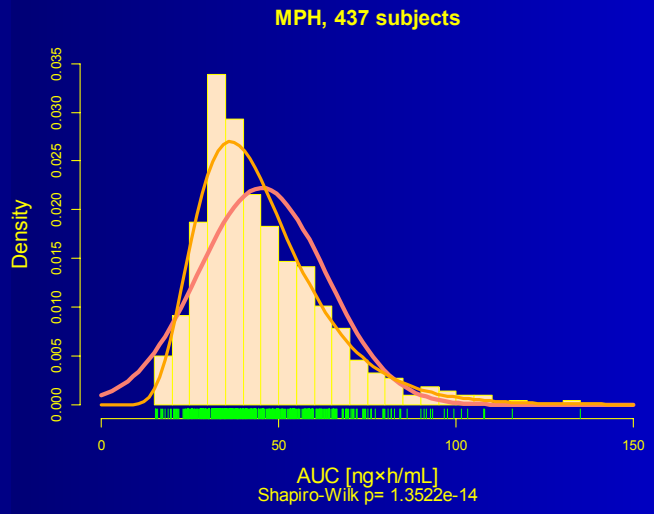
  - Variances which can be estimated:

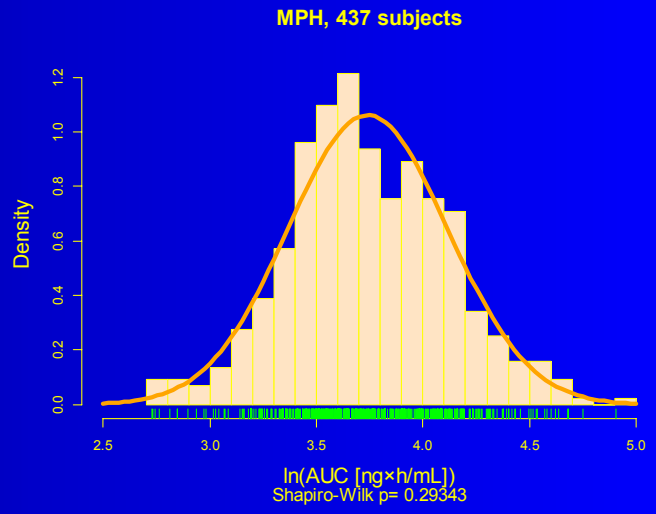
    Parallel: total variance (between + within)
    2×2 Xover: + between, within subjects
    Partial replicate: + within subjects (reference)
    Full replicate: + within subjects (reference, test)

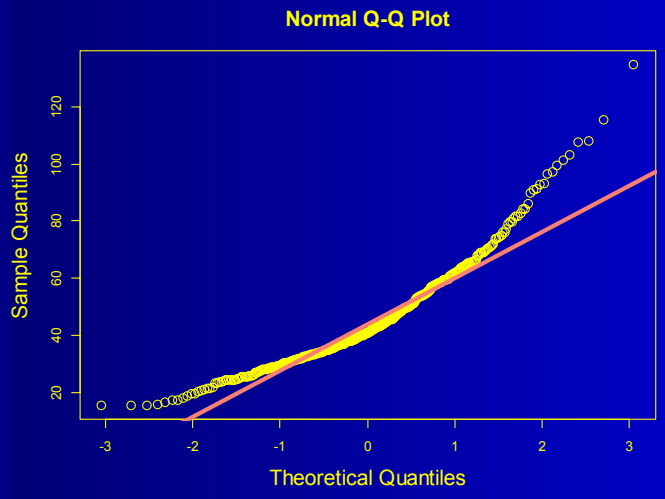
Data Transformation?

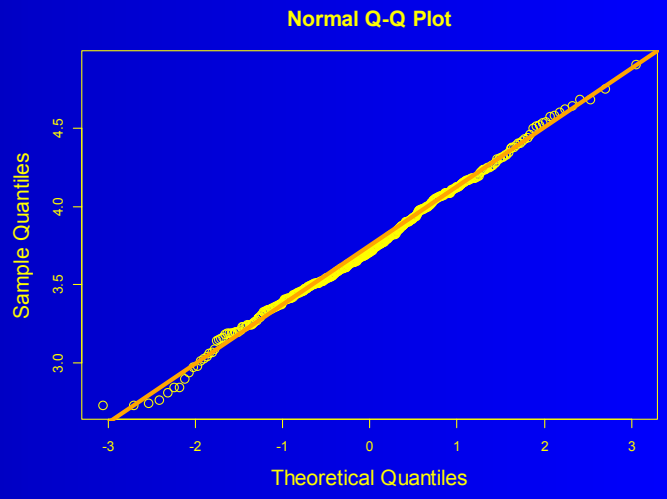
- BE testing started in the early 1980s with an acceptance range of 80% – 120% of the reference based on the *normal* distribution
- Was questioned in the mid 1980s
  - Like many biological variables AUC and C_{max} *do not* follow a normal distribution
    - Negative values are impossible
    - The distribution is skewed to the right
    - Might follow a *lognormal* distribution
  - Serial dilutions in bioanalytics lead to multiplicative errors

Data Transformation?







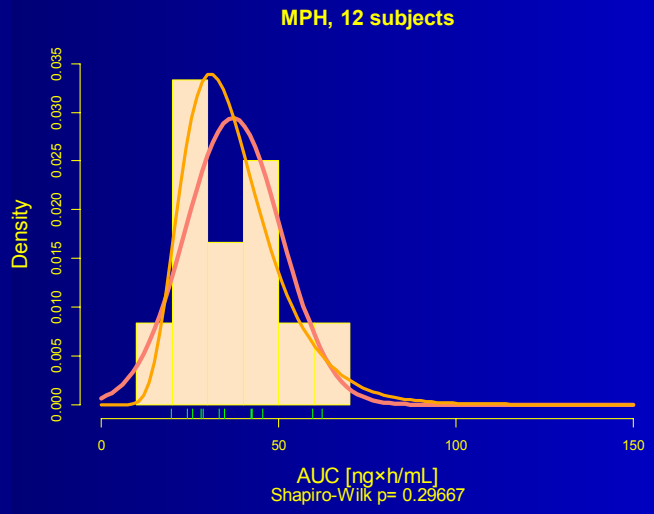


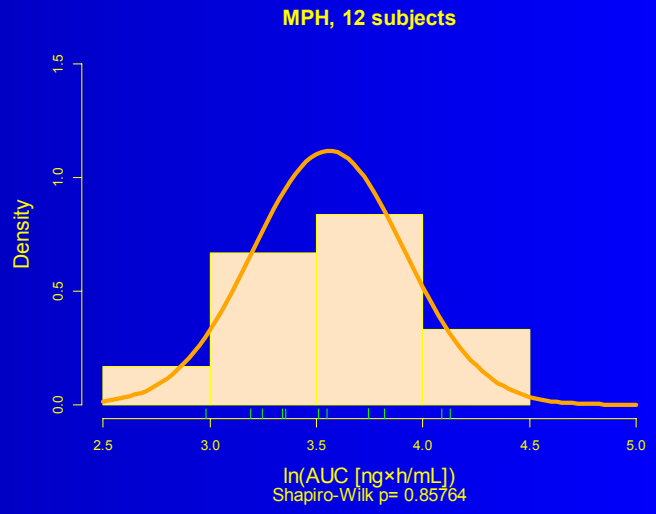
Pooled data from real studies.

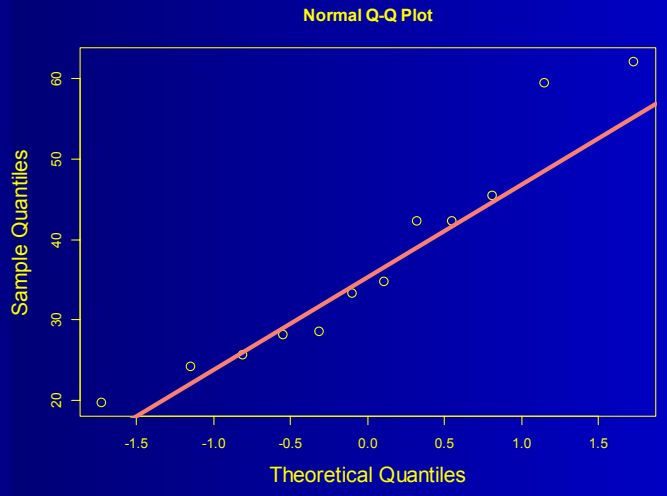
Clearly in favor of a lognormal distribution.

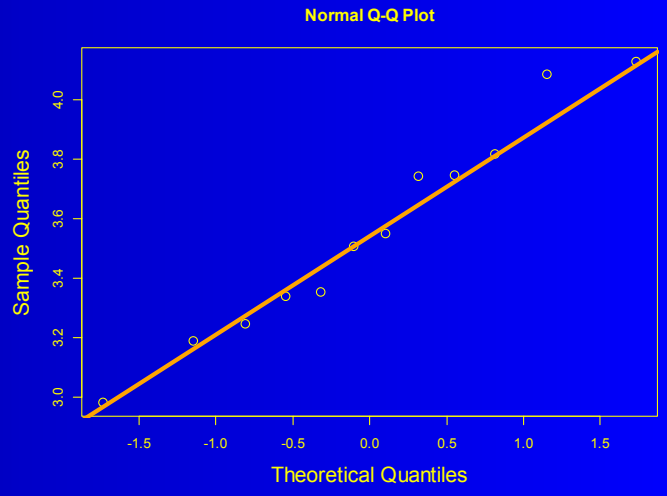
Shapiro-Wilk test highly significant for normal distribution (assumption rejected).

Data Transformation!









Data of a real study.

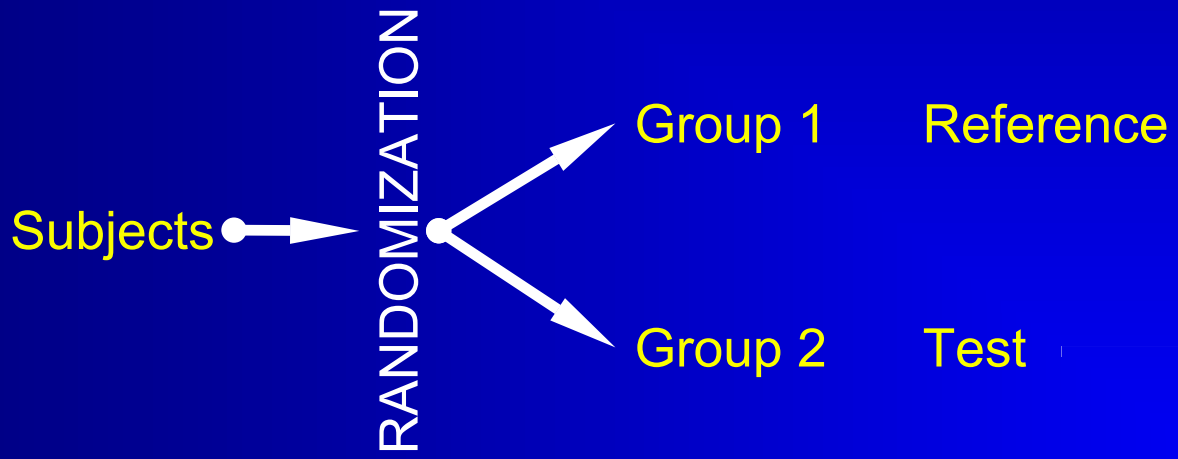
Both tests *not* significant (assumptions accepted).

Tests not acceptable according to GLs.

Transformation based on prior knowledge (PK)!

Parallel design

- Two-Group Parallel Design



Parallel design
(independent 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
    - Lower statistical power than X-over (*rule of thumb:* sample size should at least be doubled).
    - Phenotyping mandatory for drugs showing polymorphism.

Parallel design

- One group is treated with the test formulation and another group with reference
- Quite common that the dataset is imbalanced, *i.e.*, $n_1 \neq n_2$
- Guidelines against assumption of equal variances.
  Not implemented in PK soft-ware (Phoenix/WinNonlin, 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 design

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

$$s_0^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} =$$

$$= \frac{10\times 0.03418 + 11\times 0.03231}{10+11-2} =$$

$$= 0.03320$$

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

$$CI_{\ln} = |\bar{x}_1 - \bar{x}_2| \pm t_{1-\alpha, n_1+n_2-2} s_0 \sqrt{\frac{n_1+n_2}{n_1 n_2}}$$

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

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

$$CI = e^{[-0.1829, +0.07886]} =$$

$$= [83.28\%, 108.20\%]$$

Parallel design

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

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

Parallel design

- Instead of the simple $\nu = n_1 + n_2 - 2 = 21$ (t 1.7207) we get

$$\nu = \frac{\left(\frac{0.03418}{11} + \frac{0.03231}{12}\right)^2}{\frac{0.001169}{121 \cdot 10} + \frac{0.001044}{144 \cdot 11}} = 20.705$$

  and t 1.7219…

- It's time to leave M$-Excel
- Easy to calculate in R

Parallel design

```
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 design

- There is just a minor difference in CIs (83.26–108.23% *vs.* 83.28–108.20%), but there was also only little imbalance in the dataset (n_1 11, n_2 12) and variances were quite similiar (s_1^2 0.03418, s_2^2 0.03231).
- If a dataset is more imbalanced and the variances are 'truely' different, the outcome may be *substantially* different. Generally the simple *t*-test is liberal, *i.e.*, the patients' risk is increased!

Parallel design

- One million simulated BE studies
  - Lognormal distribution
  - $Mean_{Test}$ 95, $Mean_{Reference}$ 100 (target ratio 95%)
  - $CV\%_{Test}$ 25%, $CV\%_{Reference}$ 40% ('bad' reference or inhomogenous groups)
  - n_{Test} 24, $n_{Reference}$ 20
  - If width of CI (*t*-test) < CI (Welch-test) the outcome was considered 'liberal'
  - Result: *t*-test for homogenous variances was liberal in 97.62% of cases…

Parallel design

```
set.seed(1234567) # Use this line only to reproduce a run
sims    <- 1E6    # Number of simulations (1 mio simulations will take a couple of minutes)
nT      <- 24     # Subjects in test group
nR      <- 20     # Subjects in reference group
MeanT   <- 95     # Mean test (original scale)
MeanR   <- 100    # Mean reference (original scale)
CVT     <- 0.25   # CV test 25%
CVR     <- 0.40   # CV (bad) reference 40%
MeanlogT<- log(MeanT)-0.5*log(1+CVT^2) # Centered means log scale
MeanlogR<- log(MeanR)-0.5*log(1+CVR^2)
SDlogT  <- sqrt(log(1+CVT^2))          # Standard dev. log scale
SDlogR  <- sqrt(log(1+CVR^2))
Conserv <- 0      # Counters
Liberal <- 0
for (iter in 1:sims){
  PKT     <- rlnorm(n=nT, mean=MeanlogT, sd=SDlogT) # simulated T
  PKR     <- rlnorm(n=nR, mean=MeanlogR, sd=SDlogR) # simulated R
  TtestRes<- t.test(log(PKR), log(PKT), var.equal=TRUE,  conf.level=0.90)
  WelchRes<- t.test(log(PKR), log(PKT), var.equal=FALSE, conf.level=0.90)
  WidthT  <- abs(TtestRes$conf.int[1] - TtestRes$conf.int[2])
  WidthW  <- abs(WelchRes$conf.int[1] - WelchRes$conf.int[2])
  if (WidthT<WidthW){
    Liberal <- Liberal + 1
    }else{
    Conserv <- Conserv + 1
  }
}
result  <- paste(paste("t-test compared to Welch-test\n"),
           paste("Conservative =", 100*Conserv/sims, "%\n"),
           paste("Liberal =", 100*Liberal/sims, "%\n"),
           paste("Number of simulations =",sims,"\n"))
cat(result)
```

Paired design
(dependent groups)

- Every subject is treated both with test and reference.
- Generally more powerful than parallel design, because every subject acts as their own reference.
- CI is based on within- (aka *intra*-) subject variance rather than on between- (aka *inter*-) subject variance.

Subj.	Test	Ref.	s^2_{within}
1	100	110	50
2	103	113	50
3	80	96	128
4	110	90	200
5	78	111	545
6	87	68	181
7	116	111	13
8	99	93	18
9	122	93	421
10	82	82	0
11	68	96	392
12	95	137	882
n	12	12	12
mean	95	100	240
$s^2_{between}$	271	314	
$s_{between}$	16.4	17.7	

Paired design

Subj.	ln (Test)	ln (Ref.)	Δ (T–R)	(Δ-mean)²
1	4.605	4.700	-0.095	0.00199
2	4.635	4.727	-0.093	0.00176
3	4.382	4.564	-0.182	0.01731
4	4.700	4.500	+0.201	0.06321
5	4.357	4.710	-0.353	0.09125
6	4.466	4.220	+0.246	0.08830
7	4.754	4.710	+0.044	0.00899
8	4.595	4.533	+0.063	0.01283
9	4.804	4.533	+0.271	0.10379
10	4.407	4.407	±0.000	0.00258
11	4.220	4.564	-0.345	0.08649
12	4.554	4.920	-0.366	0.09945
n	12	12	Σ -0.609	Σ 0.57794
mean	4.540	4.591	-0.0507	
$s^2_{between}$	0.03110	0.03231	0.0525	s^2_{within}
$s_{between}$	0.1763	0.1798	0.2292	s_{within}

$$\bar{\Delta} = \frac{1}{n}\sum_{i=1}^{i=n}(T_i - R_i) = -\frac{0.609}{12} = -0.05075$$

$$s^2_\Delta = \frac{1}{n-1}\sum_{i=1}^{i=n}(T_i - R_i - \bar{\Delta})^2 = \frac{0.57794}{11} = 0.05254$$

$$s_\Delta = \sqrt{s^2_\Delta} = \sqrt{0.05254} = 0.2292$$

$$CI_{\ln} = \bar{\Delta} \pm t_{1-\alpha,n-1} s_\Delta \sqrt{\frac{1}{n}} =$$

$$= -0.05075 \pm 1.796 \cdot 0.2292 \sqrt{\frac{1}{12}} =$$

$$= [-0.16958, +0.06808]$$

$$CI = e^{[-0.16958, +0.06808]} = [84.40\%, 107.05\%]$$

Parallel:
83.28%,108.20%

Paired *vs.* parallel design

- Only small difference (84.40–107.50% *vs.* parallel 83.28–108.20%) since based on simulated data not accounting for different CVs (*intra vs. inter*-subject).
- Let's have a look at real data; subsets of the MPH dataset of 437 subjects.
  - 48 subjects parallel: 95.86% [75.89 –121.10%]
  - First 12 subjects paired: 100.82% [94.91 –107.09%]
  - Second 12 subjects paired: 91.15% [86.81 – 95.71%]
  - Width of CI of the paired design is only ~¼ of the parallel! Reason: CV_{intra} ~7%, CV_{total} ~28%.

R code

```
#Example MPH 20mg MR AUCinf
T <- c(28.39,49.42,36.78,33.36,34.81,24.29,
       28.61,45.54,59.49,28.23,25.71,42.30,
       62.14,19.69,42.36,97.43,48.57,75.97,
       67.93,79.22,61.68,90.80,60.64,89.91)
R <- c(35.44,39.86,32.75,33.40,34.97,24.65,
       31.77,45.44,65.29,27.87,24.26,37.01,
       63.94,20.65,43.03,115.63,57.40,69.02,
       73.98,91.47,79.65,92.86,70.46,101.40)

#Parallel log-scale (n=48)
par <- t.test(log(T), log(R),
        alternative="two.sided", mu=0,
        paired=FALSE, var.equal=FALSE,
        conf.level=0.90)
result <- paste(paste(
         " Back transformed (raw data scale)",
         "\n Point estimate:",
         round(100*exp(par$estimate[1]-
           par$estimate[2]),
           digits=2),"%\n"),
         paste("90 % confidence interval:"),
         paste(round(100*exp(par$conf.int[1]),
           digits=2), "-"),
         paste(round(100*exp(par$conf.int[2]),
           digits=2),"%\n"))
par
cat(result)
```

```
#Paired first 12 subjects (using first dataset)
T1     <- T[1:12]; R1 <- R[1:12]
pair1  <- t.test(log(T1), log(R1),alternative="two.sided",
            mu=0, paired=TRUE, conf.level=0.90)
result <- paste(paste(" Back transformed (raw data scale)",
                "\n Point estimate:",
                round(100*exp(pair1$estimate),
                digits=2),"%\n"),
           paste("90 % confidence interval:"),
           paste(round(100*exp(pair1$conf.int[1]),
             digits=2), "-"),
           paste(round(100*exp(pair1$conf.int[2]),
             digits=2),"%\n"))
pair1
cat(result)

#Paired second 12 subjects (using first dataset)
T2     <- T[13:24]; R2 <- R[13:24]
pair2  <- t.test(log(T2), log(R2),alternative="two.sided",
            mu=0, paired=TRUE, conf.level=0.90)
result <- paste(paste(" Back transformed (raw data scale)",
                "\n Point estimate:",
                round(100*exp(pair2$estimate),
                digits=2),"%\n"),
           paste("90 % confidence interval:"),
           paste(round(100*exp(pair2$conf.int[1]),
             digits=2), "-"),
           paste(round(100*exp(pair2$conf.int[2]),
             digits=2),"%\n"))
pair2
cat(result)
```

R's results

```
        Welch Two Sample t-test

data:  log(T) and log(R)
t = -0.3036, df = 45.69, p-value = 0.7628
alternative hypothesis: true difference in
means is not equal to 0
90 percent confidence interval:
 -0.2759187  0.1914053
sample estimates:
mean of x mean of y
 3.840090  3.882346

Back transformed (raw data scale)
Point estimate: 95.86 %
90 % confidence interval: 75.89 - 121.1 %
```

```
        Paired t-test

data:  log(T1) and log(R1)
t = 0.2418, df = 11, p-value = 0.8133
alternative hypothesis: true difference in means
is not equal to 0
90 percent confidence interval:
 -0.05227222  0.06854199
sample estimates:
mean of the differences
            0.008134884

Back transformed (raw data scale)
Point estimate: 100.82 %
90 % confidence interval: 94.91 - 107.09 %
```

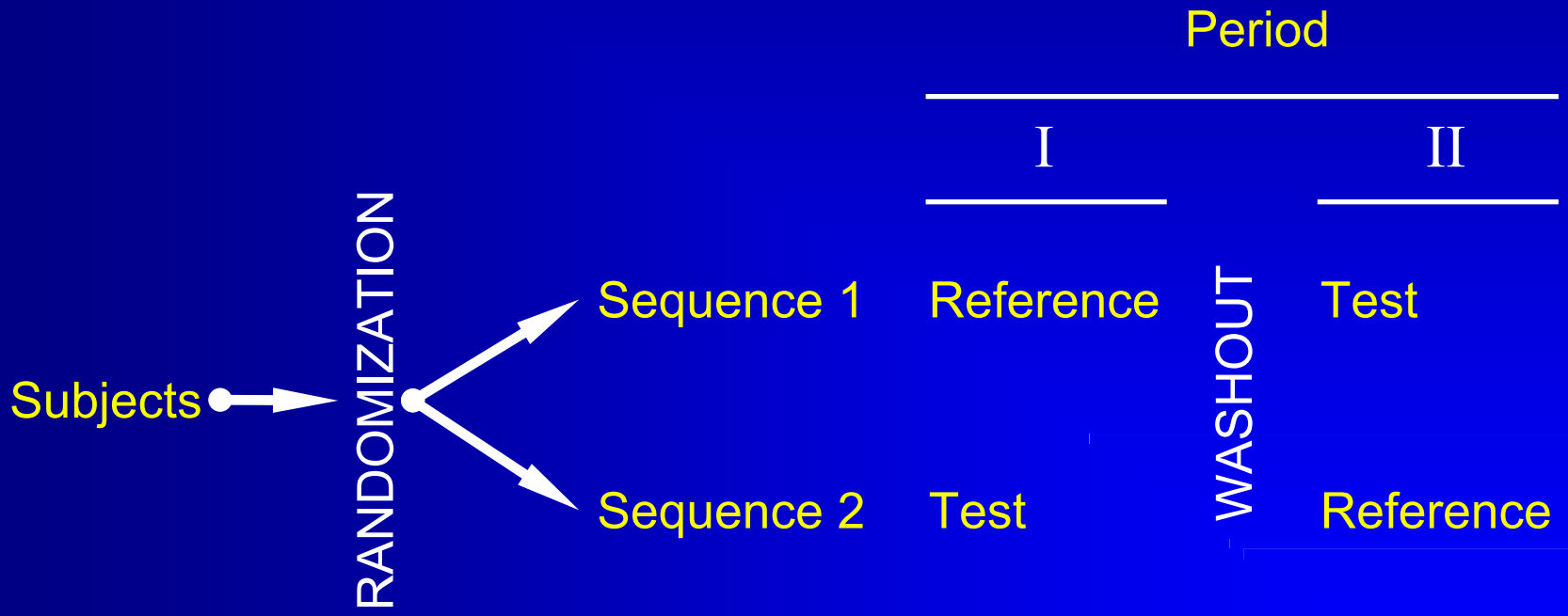
```
        Paired t-test

data:  log(T2) and log(R2)
t = -3.4076, df = 11, p-value = 0.00585
alternative hypothesis: true difference in means
is not equal to 0
90 percent confidence interval:
 -0.14147665 -0.04381995
sample estimates:
mean of the differences
             -0.0926483

Back transformed (raw data scale)
Point estimate: 91.15 %
90 % confidence interval: 86.81 - 95.71 %
```

Cross-over designs

- Standard 2×2×2 Design



Cross-over designs (cont'd)

- Every subject is treated both with test and reference
- Subjects are randomized into two groups; one is receiving the formulations in the order RT and the other one in the order TR. These two orders are called sequences
- Whilst in a paired design we must rely on the assumption that no external influences affect the periods, a cross-over design will account for that

Cross-over design: Model

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 ($l=1,2$: $\Phi_1=\Phi_{test}$, $\Phi_2=\Phi_{ref.}$)

Cross-over design:
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 (cont'd)

- Standard 2×2×2 design
  - Advantages
    - Globally applied standard protocol for bioequivalence, PK interaction, food studies
    - 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 for HVDs/HVDPs (→ Replicate Designs)

Cross-over design: 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.
- Suitable software: SAS, Phoenix/WinNonlin, Kinetica, and EquivTest/PK (both only 2×2 Xover), S+, Package *bear* for R (freeware).

Cross-over design: 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 design: Example

Sequence	Period 1		Period 2		Sequence mean	
1	1R = $X_{\cdot 11}$	3.5103	1T = $X_{\cdot 21}$	3.5768	$X_{\cdot\cdot 1}$	3.5436
2	2T = $X_{\cdot 12}$	3.5380	2R = $X_{\cdot 22}$	3.5883	$X_{\cdot\cdot 2}$	3.5631
Period mean		$X_{\cdot 1\cdot}$ 3.5241		$X_{\cdot 2\cdot}$ 3.5826	$X_{\cdots}$	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 design: 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 design: 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, $t_{1-\alpha,n-2}$ 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

$$CL_L = e^{\ln PE - t_{1-\alpha,df}\cdot SE_\Delta} = e^{0.0081349 - 1.8125\times 0.030047} = 95.47\%$$

$$CL_H = e^{\ln PE + t_{1-\alpha,df}\cdot SE_\Delta} = e^{0.0081349 + 1.8125\times 0.030047} = 106.46\%$$

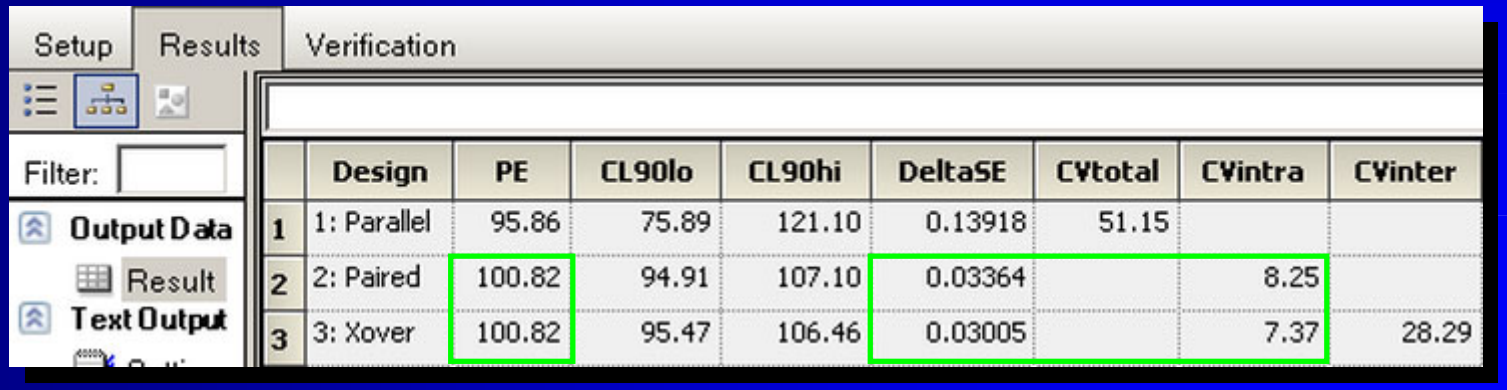
R code / result

```
#Cross-over 12 subjects
T1     <- c(28.39,33.36,24.29,28.61,59.49,42.30)
T2     <- c(49.42,36.78,34.81,45.54,28.23,25.71)
R1     <- c(39.86,32.75,34.97,45.44,27.87,24.26)
R2     <- c(35.44,33.40,24.65,31.77,65.29,37.01)
RT     <- log(R1) - log(T2)
TR     <- log(R2) - log(T1)
n1     <- length(RT)
mRT    <- mean(RT)
vRT    <- var(RT)
n2     <- length(TR)
mTR    <- mean(TR)
vTR    <- var(TR)
mD     <- mean(log(c(T1,T2))) - mean(log(c(R1,R2)))
MSE    <- (((n1-1)*vRT + (n2-1)*vTR)/(n1+n2-2))/2
alpha  <- 0.05
lo     <- mD - qt(1-alpha,n1+n2-2)*sqrt(MSE)*
            sqrt((1/(2*n1) + 1/(2*n2)))
hi     <- mD + qt(1-alpha,n1+n2-2)*sqrt(MSE)*
            sqrt((1/(2*n1) + 1/(2*n2)))
result <- paste(
          paste(" Back transformed (raw data scale)",
                "\n Point estimate☺",
                round(100*exp(mD), digits=2),"%\n"),
          paste("90 % confidence interval:"),
          paste(round(100*exp(lo), digits=2), "-"),
          paste(round(100*exp(hi), digits=2),"%\n",
          paste("CVintra:",round(100*sqrt(exp(MSE)-1),
            digits=2),"%\n")))
cat(result)
```

```
Back transformed (raw data scale)
Point estimate: 100.82 %
90 % confidence interval: 95.47 - 106.46 %
CVintra: 7.37 %
```

Comparison of designs

- Further reduction in variability since the influence of periods is accounted for
  - Paired design: 100.82% [94.91–107.10%]
  - Cross-over design: 100.82% [95.47–106.46%]
  - Point estimates are identical; narrower CI – variability caused by period- and/or sequence-effects is reduced.



	Design	PE	CL90lo	CL90hi	DeltaSE	CVtotal	CVintra	CVinter
1	1: Parallel	95.86	75.89	121.10	0.13918	51.15		
2	2: Paired	100.82	94.91	107.10	0.03364		8.25	
3	3: Xover	100.82	95.47	106.46	0.03005		7.37	28.29

Comparison of designs

- Most important in an ANOVA table: residual mean error ($\rightarrow$ CI, CV_{intra} for future studies)
  - Carry-over can not be handled! Has to be excluded by design (sufficiently long washout)
  - Period effects are accounted for. Example: P2 ×10…

Hypothesis	F_stat	P_value	Data	DF	SS	MS
sequence	0.0144	0.9068	original	1	0.002300	0.002300
sequence*subject	29.4312	4.321E-06	original	10	1.594347	0.159435
treatment	0.0733	0.7921	original	1	0.000397	0.000397
period	3.7844	0.08036	original	1	0.020501	0.020501
Error			original	10	0.054172	0.005417
sequence	0.0144	0.9068	P2 ×10	1	0.002300	0.002300
sequence*subject	29.4312	4.321E-06	P2 ×10	10	1.594347	0.159435
treatment	0.0733	0.7921	P2 ×10	1	0.000397	0.000397
period	6174.2345	3E-15	P2 ×10	1	33.447023	33.447023
Error			P2 ×10	10	0.054172	0.005417

CI_90_Lower	CI_90_Upper
95.471828	106.46102

CI_90_Lower	CI_90_Upper
95.471828	106.46102

Reading ANOVA tables

Analysis of Variance

Source of variation	df	SS	MS	F	P-value	CV
Between subjects						
Carry-over	1	0.00230	0.002300	0.0144	0.90679	
Residuals	10	1.59435	0.159435	29.4312	4.32E-6	28.29%
Within subjects						
Direct drug	1	0.00040	0.000397	0.0733	0.79210	
Period	1	0.02050	0.020501	3.7844	0.08036	
Residuals	10	0.05417	0.005417			7.37%
Total	23	1.67172				

Should not be tested: Design – washout!

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

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

Not surprising: different subjects!

Not important: Significant value would only mean that 100% is not included in the CI.

Not important: Both formulations would be affected in the same way.

balanced: $n_1 = n_2;\ n = n_1 + n_2$

$$CI = e^{\ln PE \pm t_{\alpha, n-2} \cdot \sqrt{MSE_W} \sqrt{\frac{2}{n}}}$$

imbalanced: $n_1 \neq n_2$

$$CI = e^{\ln PE \pm t_{\alpha, n_1+n_2-2} \cdot \sqrt{MSE_W} \sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}}$$

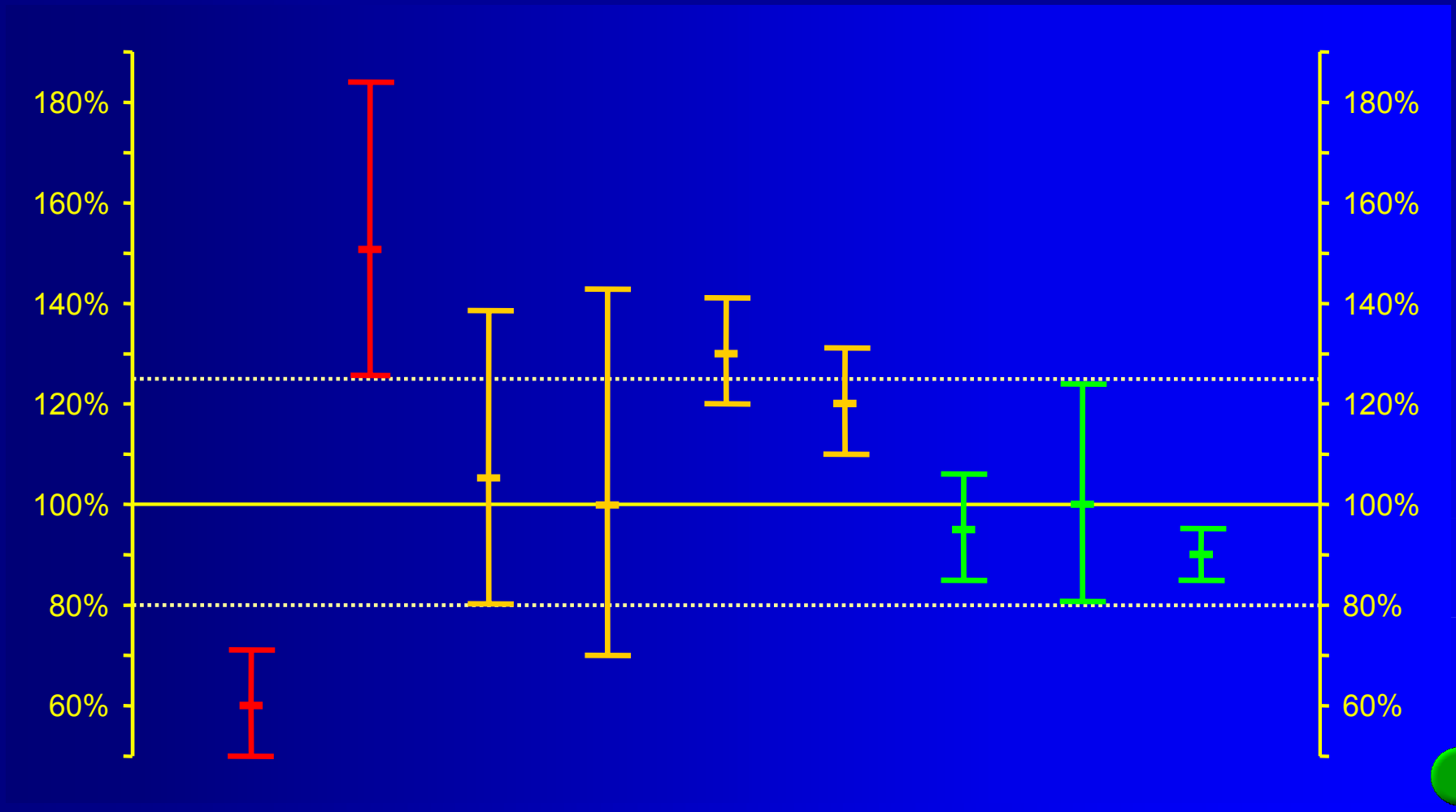
BE Evaluation

- Based on the design set up a statistical model.
- Calculate the test/reference ratio.
- Calculate the 90% confidence interval (CI) around the ratio.
- The *width* of the CI depends on the variability observed in the study.
- The *location* of the CI depends on the observed test/reference-ratio.

BE Assessment

- Decision rules based on the CI and the Acceptance Range (AR)
  - CI *entirely outside* the AR:
    Bioinequivalence proven
  - CI *overlaps* the AR (lies *not entirely within* the AR):
    Bioequivalence not proven
  - CI lies *entirely within* the AR:
    Bioequivalence proven

BE Assessment



Cross-over designs (cont'd)

- Special case: Evaluation of t_{max}
  - Since t_{max} is sampled from discrete values, a nonparametric method must be applied
  - Estimation of differences (linear model)
  - Wilcoxon Two-Sample Test (available in SAS 9.2 Proc NPAR1way, Phoenix/WinNonlin, EquivTest/PK, R package *coin*)
  - Since based on a discrete distribution, generally $\alpha<0.05$ (*e.g.*, n=12: 0.0465, 24: 0.0444, 32: 0.0469, 36: 0.0485, 48: 0.0486,…)

Hauschke D, Steinijans VW and E Diletti
A distribution-free procedure for the statistical analysis of bioequivalence studies
Int J Clin Pharm Ther Toxicol 28(2), 72–8 (1990)

Cross-over designs (cont'd)

Sequence 1 (RT)				Sequence 2 (TR)			
Subject	Period I	Period II	P.D.	Subject	Period I	Period II	P.D.
2	3.0	1.5	-1.5	1	2.0	2.0	±0.0
4	2.0	2.0	±0.0	3	2.0	2.0	±0.0
6	2.0	3.0	+1.0	5	2.0	3.0	+1.0
8	2.0	3.0	+1.0	7	2.0	1.5	-0.5
10	1.5	2.0	+0.5	9	3.0	2.0	-1.0
12	3.0	2.0	-1.0	11	2.0	1.5	-0.5
14	3.0	3.0	±0.0	13	3.0	1.5	-1.5

Cross-over designs (cont'd)

ADDITIVE (raw data) MODEL

metric: t_{max}

Sequence	Period 1		Period 2	
1	R_{L1} =	65	R_{U1} =	46
2	R_{L2} =	36	R_{U2} =	55
RT =	n_1 =	7		
TR =	n_2 =	7		
balanced	n =	14	$n_1 \cdot n_2$	49

$d_{\cdot 1}$ 0.0000 $d_{\cdot 2}$ -0.1786 *(mean period difference in sequence 1 / 2)*

$\tilde{Y}_R$ 2.000 *median of the reference formulation*

$\tilde{Y}_T$ 2.000 *median of the test formulation*

Distribution-Free Confidence Interval (Moses)

± x rule : 20

θ_L -0.429 θ_U +0.429 α 0.0487 $p = 1-2\cdot\alpha$ 0.9026

δ_L 80% δ_U 120%

L_W -0.250 U_W +0.750 ***difference outside Theta-L AND/OR Theta-U; not bioequivalent***

$\tilde{\theta}$ **+0.250** ***Hodges-Lehmann estimate (median of paired differences)***

Wilcoxon-Mann-Whitney Two One-Sided Tests Procedure (Hauschke)

W_L 37 W_U 18

$W_{0.95,n1,n2}$ 38 $W_{0.05,n1,n2}$ 12 *H0(1): diff. <= Theta-L AND H0(2): diff. => Theta-U; not bioequivalent*

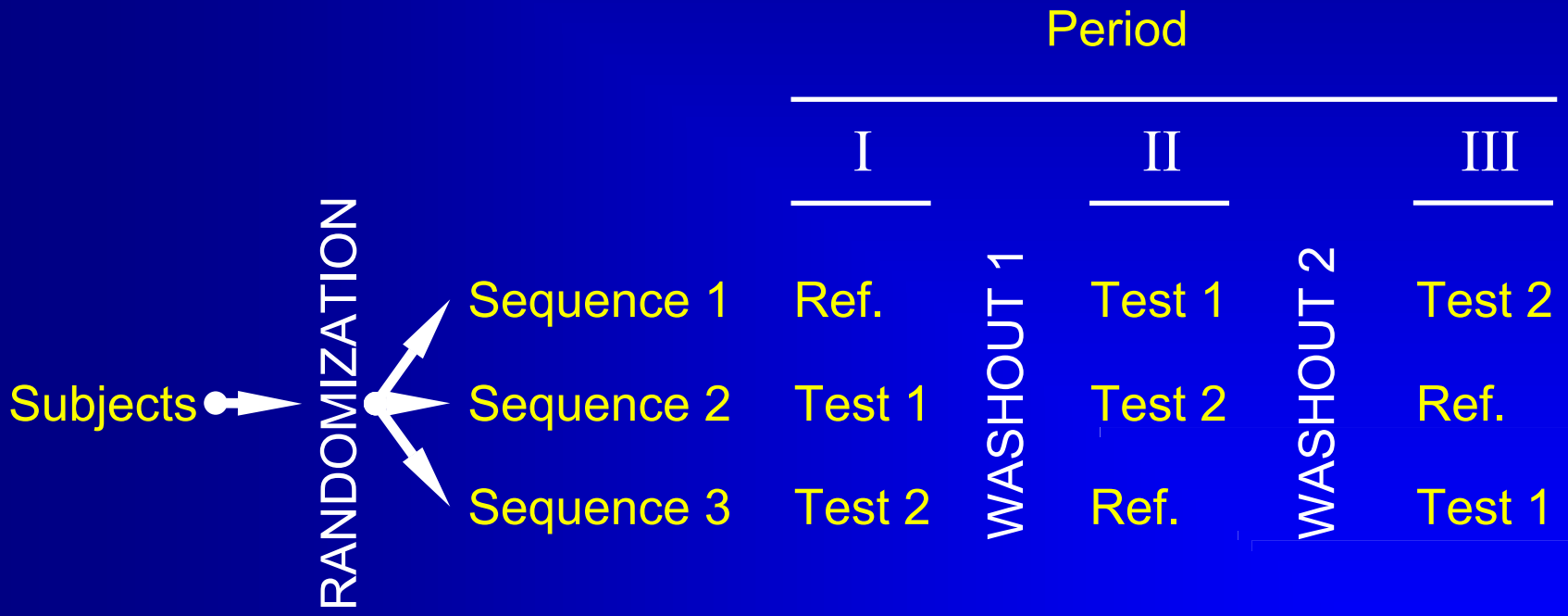
p_1 >0.0487 and p_2 >0.0487

Cross-over designs (cont'd)

- 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 (cont'd)

- 3×3×3 Latin Square Design



Cross-over designs (cont'd)

- 3×3×3 Latin Square design
  - Advantages
    - Allows to choose between two candidate test formulations or comparison of one 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 some pieces of software.
    - Extracted pairwise comparisons are imbalanced.
    - May need measures against multiplicity (increasing the sample size).
    - Not mentioned in any guideline.

Cross-over designs (cont'd)

- 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 (cont'd)

- 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 (cont'd)

- 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 (cont'd)

- Williams' Designs
  - Advantages
    - Allows to choose between two candidate test formulations or comparison of one 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 Brazil's (ANVISA) and EU's (EMA) guidelines.
  - Disadvantages
    - Mores 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 (cont'd)

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

k	$p_{\alpha=0.05}$	$p_{\alpha=0.10}$	α_{adj}	p_{corr}	α_{adj}	p_{corr}
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%

$$\alpha_{adj} = \alpha^{1/k}$$

$$p_{corr} = 1 - \left(1 - \alpha_{adj}\right)^{k}$$

Add-on / Two-Stage Designs

- Sometimes properly designed and executed studies fail due to
  - 'true' bioinequivalence,
  - poor study conduct (increasing variability),
  - pure chance (producer's risk hit),
  - false (over-optimistic) assumptions about variability and/or T/R-ratio.
- The patient's risk must be preserved
  - Already noticed at Bio-International Conferences (1989, 1992) and guidelines from the 1990s.

Sequential Designs

- Have a long and accepted tradition in clinical research (mainly phase III)
  - Based on work by Armitage *et al.* (1969), McPherson (1974), Pocock (1977), O'Brien and Fleming (1979), Lan & DeMets (1983), …
    - First proposal by Gould (1995) in the area of BE did not get regulatory acceptance in Europe, but
    - new methods stated in recent guidelines.

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

Sequential Designs

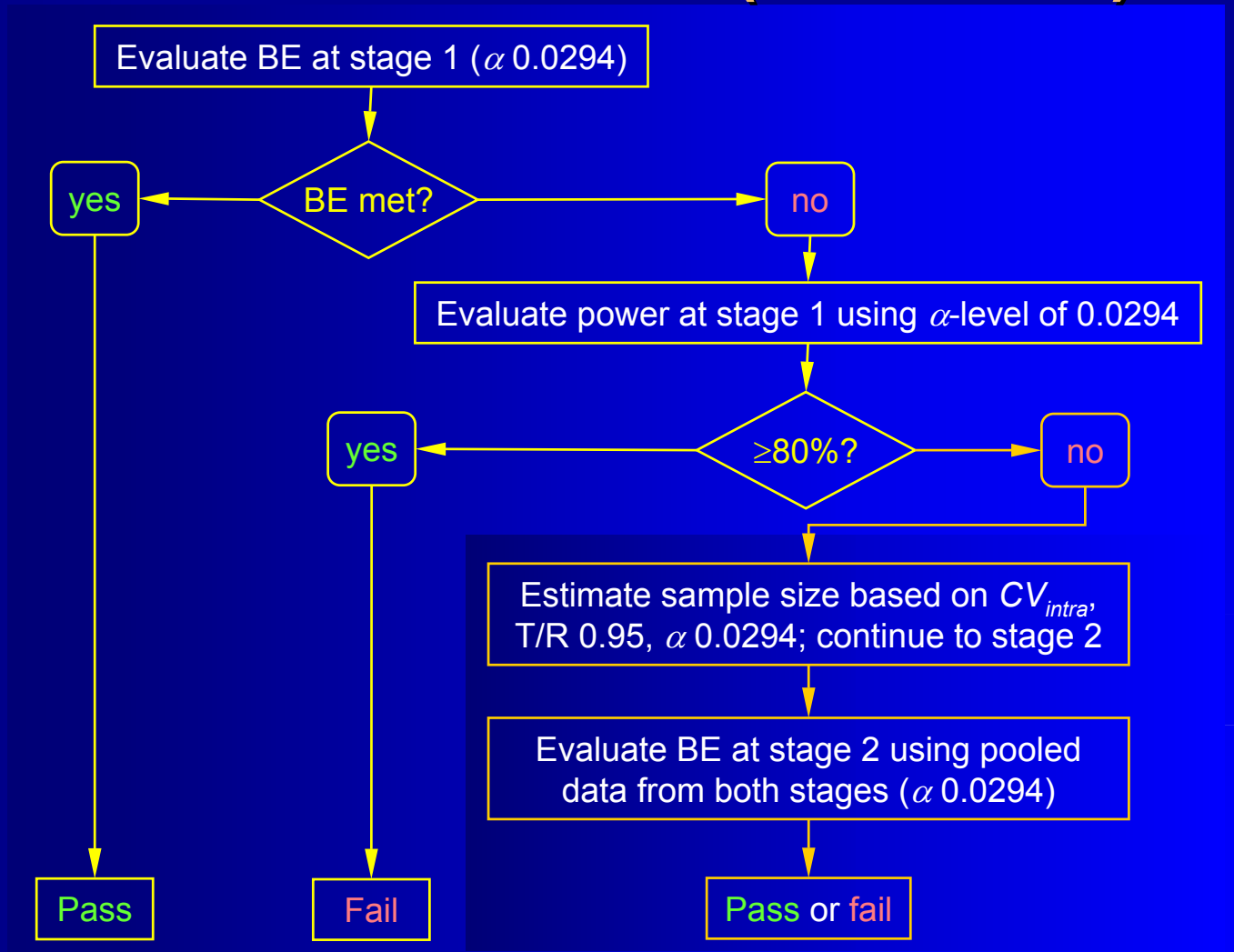
- Methods by Potvin *et al.* (2008) first validated framework in the context of BE
  - Supported by the 'Product Quality Research Institute' (members: FDA/CDER, Health Canada, USP, AAPS, PhRMA…)
    - Three of BEBAC's protocols accepted by German BfArM, one product approved in 06/2011.

Potvin D, Diliberti CE, Hauck WW, Parr AF, Schuirmann DJ, and RA Smith
Sequential design approaches for bioequivalence studies with crossover designs
Pharmaceut Statist 7(4), 245–62 (2008) DOI: 10.1002/pst.294

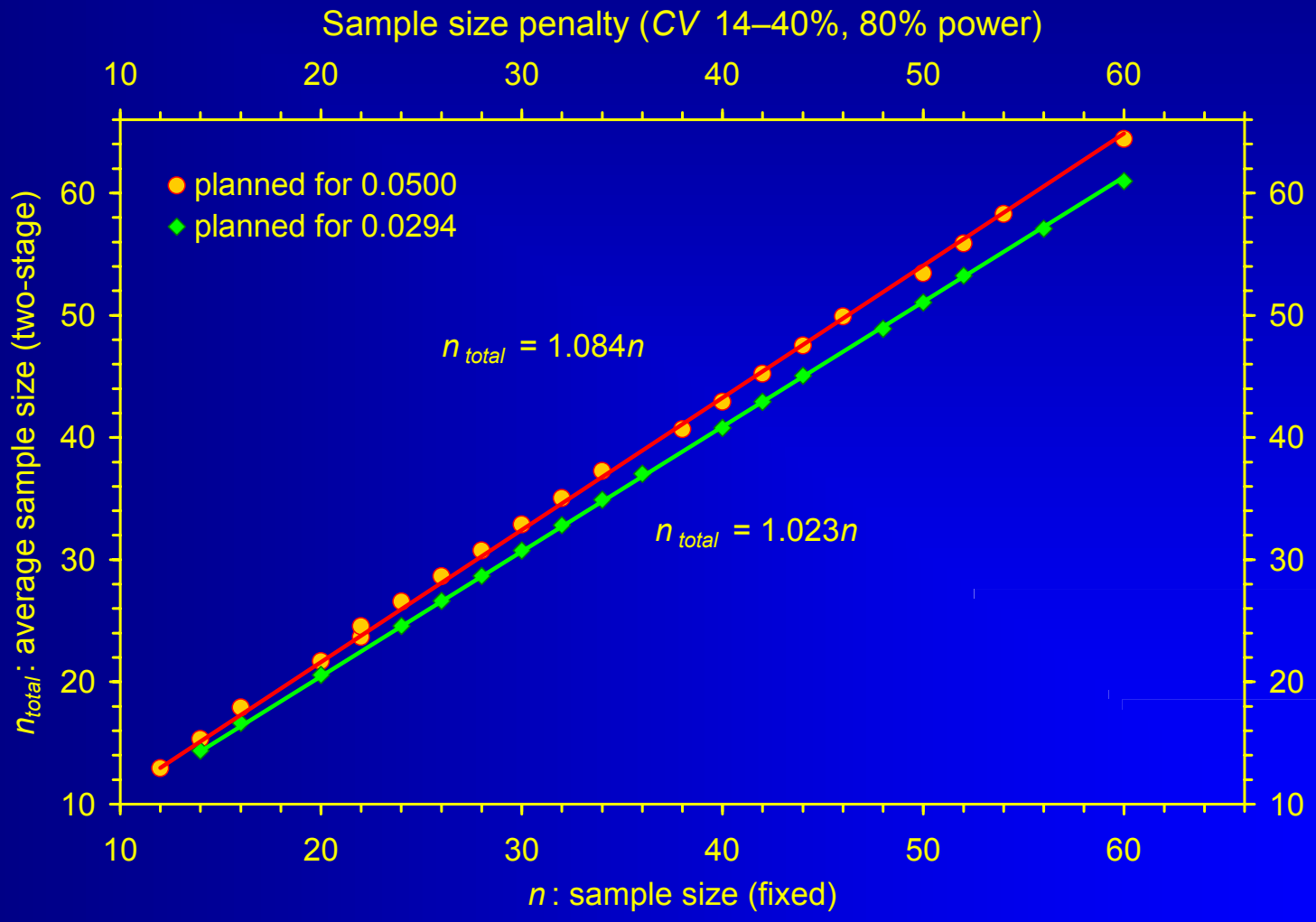
Review of Guidelines

- EMA (Jan 2010)
  Acceptable; Potvin *et al.* Method B preferred (?)
- Russia (Draft 2011)
  Acceptable (Methods B and C)
- Canada (May 2012)
  Potvin *et al.* Method C recommended
- FDA (Jun 2012)
  Potvin *et al.* Method C recommended
  API specific guidances: Loteprednol, Dexamethasone / Tobramycin

Potvin *et al.* (Method B)



Potvin *et al.* (Method B)



Potvin *et al.* (Method B)

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

Potvin *et al.* (Method B)

- Technical Aspects (cont'd)
  - No futility rule preventing to go into stage 2 with a very high sample size! Must be clearly stated in the protocol (unfamiliar to the IEC because common in Phase III).
  - Pocock's α 0.0294 is used in stage 1 and in the pooled analysis (data from stages 1 + 2), *i.e.*, the $1 - 2\times\alpha = 94.12\%$ CI is calculated.
  - Overall patient's risk preserved at ≤ 0.05.

Potvin *et al.* (Method B)

- Technical Aspects (cont'd) + EMA modification
  - If the study is stopped after stage 1, the statistical model is:

```
fixed:  sequence + period + treatment +
        subject(sequence)
```

  - If the study continues to stage 2, the model for the combined analysis is:

```
fixed:  stage + sequence + sequence(stage) +
        subject(sequence × stage) + period(stage) +
        treatment
```

  - No poolability criterion! Combining is *always allowed* – even if a significant difference between stages is observed. No need to test this effect.

Potvin *et al.* (Method B)

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

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

Example (Potvin Method B)

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

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

Bioequivalence Statistics
User-Specified Confidence Level for CI's = 94.1200
Percent of Reference to Detect for 2-1 Tests = 20.0%
A.H.Lower =  0.800   A.H.Upper =  1.250
Reference: Reference   LSMean = 0.954668  SE = 0.191772  GeoLSM = 2.597808
----------------------------------------------------------------------
Test:      Test        LSMean = 1.038626  SE = 0.191772  GeoLSM = 2.825331

    Difference  =   0.0840,  Diff_SE = 0.0737,  df = 10.0
    Ratio(%Ref) = 108.7583

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

12 subjects in stage 1, conventional BE model

CV_{intra} 18.2%

α 0.0294

Failed with 94.12% Confidence Interval

Example (Potvin Method B)

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

[1] 0.5251476
```

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

Power 52.5% – initiate stage 2

```
sampleN.TOST(alpha=0.0294, targetpower=0.80, logscale=TRUE,
             theta1=0.8, theta2=1.25, theta0=0.95,
             CV=0.182132, design='2x2', method='exact',
             print=TRUE)

++++++++++++ Equivalence test - TOST ++++++++++++
             Sample size estimation
-----------------------------------------------
Study design:  2x2 crossover
log-transformed data (multiplicative model)

alpha = 0.0294, target power = 0.8
BE margins        = 0.8 ... 1.25
Null (true) ratio = 0.95,  CV = 0.182132

Sample size
 n     power
20   0.829160
```

Estimate total sample size:
α 0.0294, T/R 95%, CV_{intra} 18.2%, 80% power

Total sample size 20: include another 8 in stage 2

Example (Potvin Method B / EMA)

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

Final variance parameter estimates:
Var(Sequence*Stage*Subject)   0.549653
          Var(Residual)       0.0458956
       Intrasubject CV        0.216714

Bioequivalence Statistics
User-Specified Confidence Level for CI's = 94.1200
Percent of Reference to Detect for 2-1 Tests = 20.0%
A.H.Lower =  0.800   A.H.Upper =  1.250
Formulation variable: Treatment
Reference: Reference   LSMean = 1.133431  SE = 0.171385  GeoLSM = 3.106297
-------------------------------------------------------------------------
Test:      Test        LSMean = 1.147870  SE = 0.171385  GeoLSM = 3.151473

    Difference  =   0.0144,  Diff_SE = 0.0677,  df = 17.0
    Ratio(%Ref) = 101.4544

                      Classical
    CI  90% = (   90.1729,  114.1472)
    CI User = (   88.4422,  116.3810)
    Average bioequivalence shown for confidence=94.12 and percent=20.0.
```

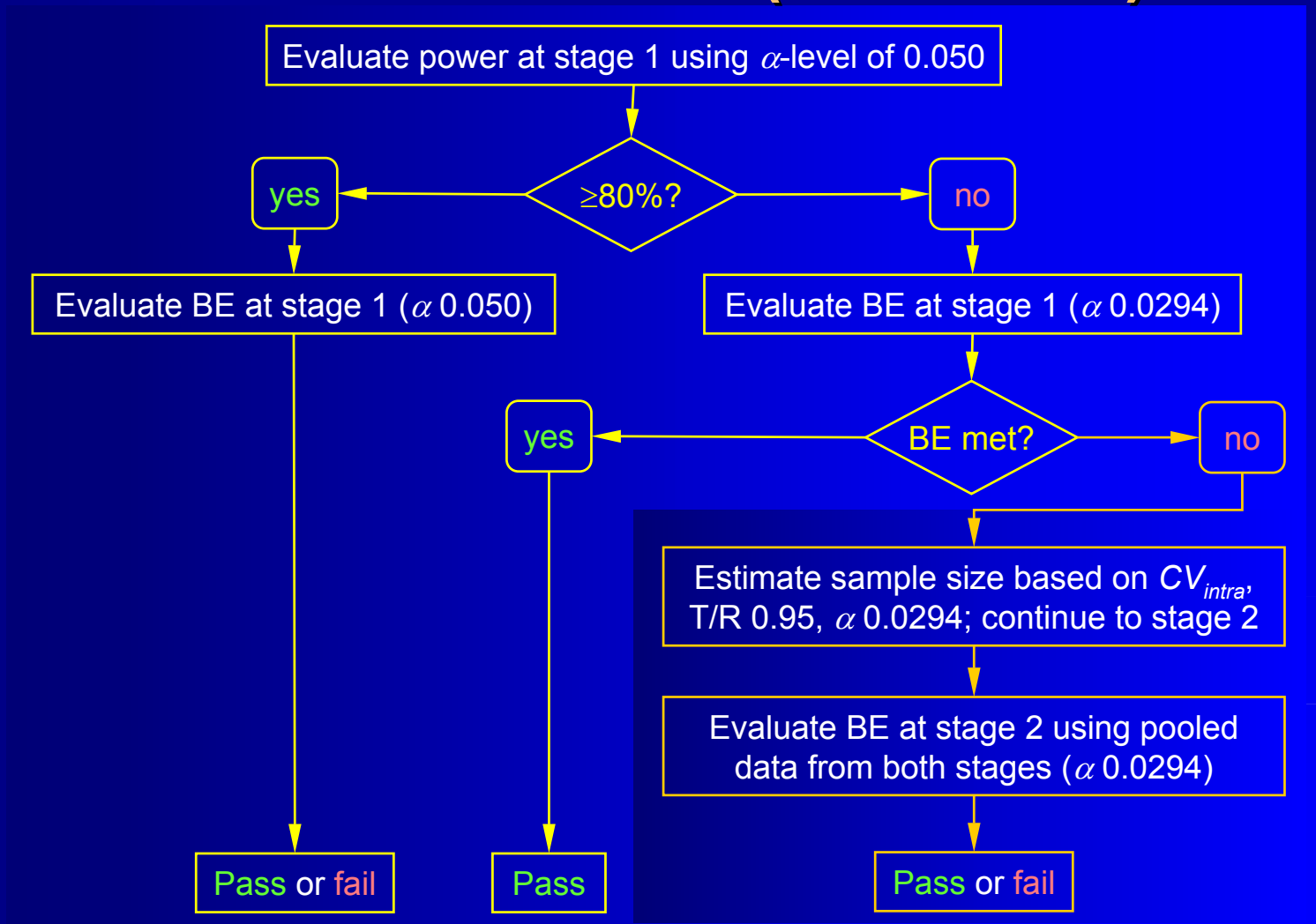
8 subjects in stage 2 (20 total), modified model in pooled analysis

Q&A Rev. 7 (March 2013)

α 0.0294 in pooled analysis

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

Potvin *et al.* (Method C)



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

- Pros & cons
  - Method C (*if power ≥80%!*) is a conventional BE study; no penality in terms of α needs to be applied.
  - Method C proceeds to stage 2 less often and has smaller average total sample sizes than Method B for cases where the initial sample size is reasonable for the CV.
  - If the size of stage 1 is low for the actual CV both methods go to stage 2 almost all the time; total sizes are similar.
  - Method B slightly more conservative than C.

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

- Recommendations
  - Method C preferred due to slightly higher power than method B (FDA, HPB). Method B for EMA (?)
  - Plan the study *as if* the CV is known
    - If assumptions turn out to be true = no penalty
    - If lower power (CV_{intra} higher than expected), BE still possible in first stage (penalty; 94.12% CI) or continue to stage 2 as a ‘safety net’.
  - Don’t jeopardize! Smaller sample sizes in the first stage than in a fixed design don’t pay off. Total sample sizes are ~10–20% higher.

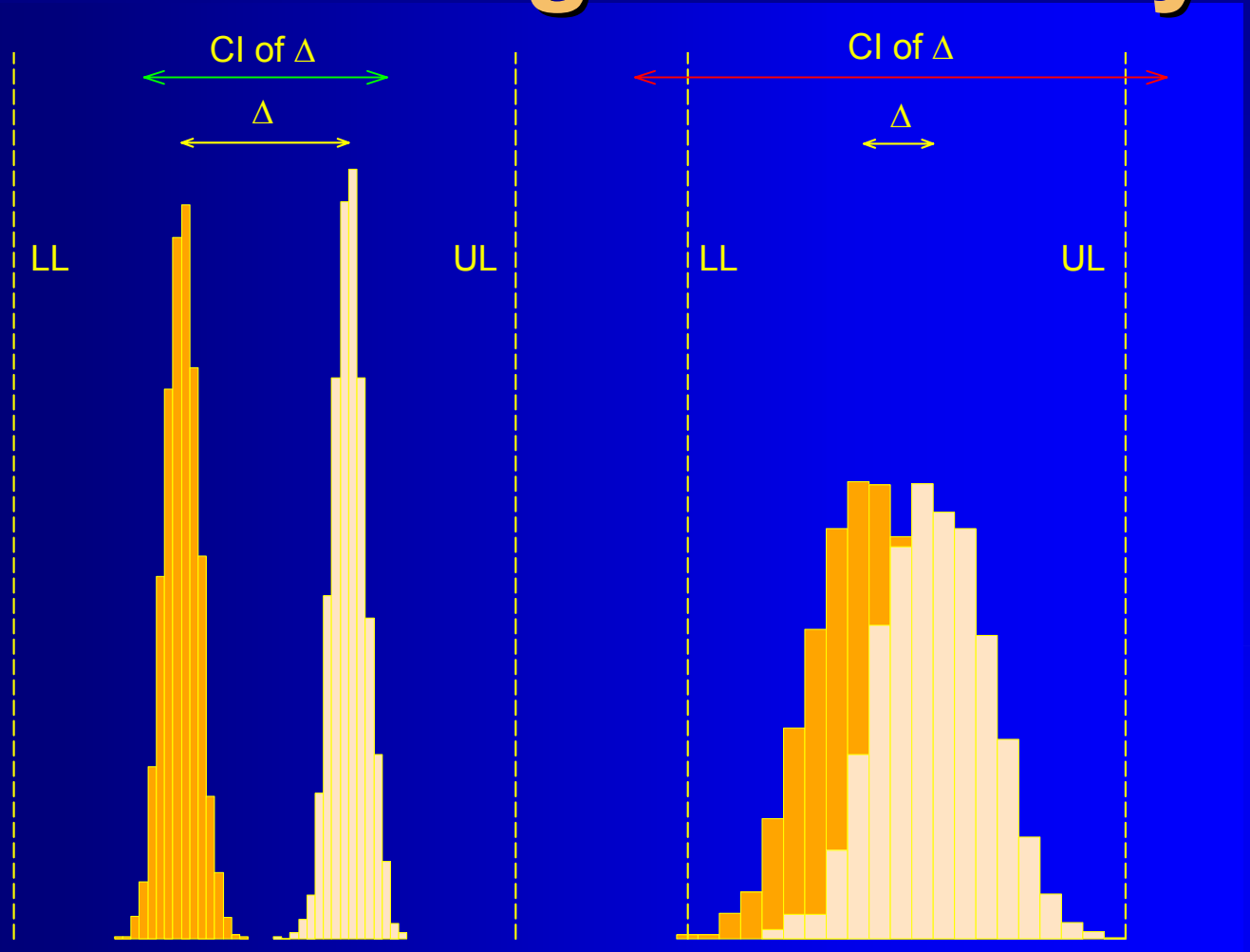
TSDs: Alternatives

- Methods by Potvin *et al.* (2008) limited to T/R of 0.95 and 80% power
  - Follow-up papers (T/R 0.95…0.90, 80…90% power)

reference	method	T/R	target power	*CV*	$\alpha_{adj.}$	max.$\alpha_{emp.}$
Potvin *et al.*	B	0.95	80%	10–100%	0.0294	0.0485
	C	0.95				0.0510
Montague *et al.*	D	0.90			0.0280	0.0518
Fuglsang	B	0.95	90%	10–80%	0.0284	0.0501
	D				0.0274	0.0503
	D	0.90			0.0269	0.0501

Montague TH, Potvin D, DiLiberti CE, Hauck WW, Parr AF, and DJ Schuirmann
Additional results for 'Sequential design approaches for bioequivalence studies with crossover designs'
Pharmaceut Statist 11(1), 8–13 (2011) DOI: 10.1002/pst.483
A Fuglsang
Sequential Bioequivalence Trial Designs with Increased Power and Controlled Type I Error Rates
AAPS J 15, pre-print online (2013) DOI: 10.1208/s12248-013-9475-5

High variability



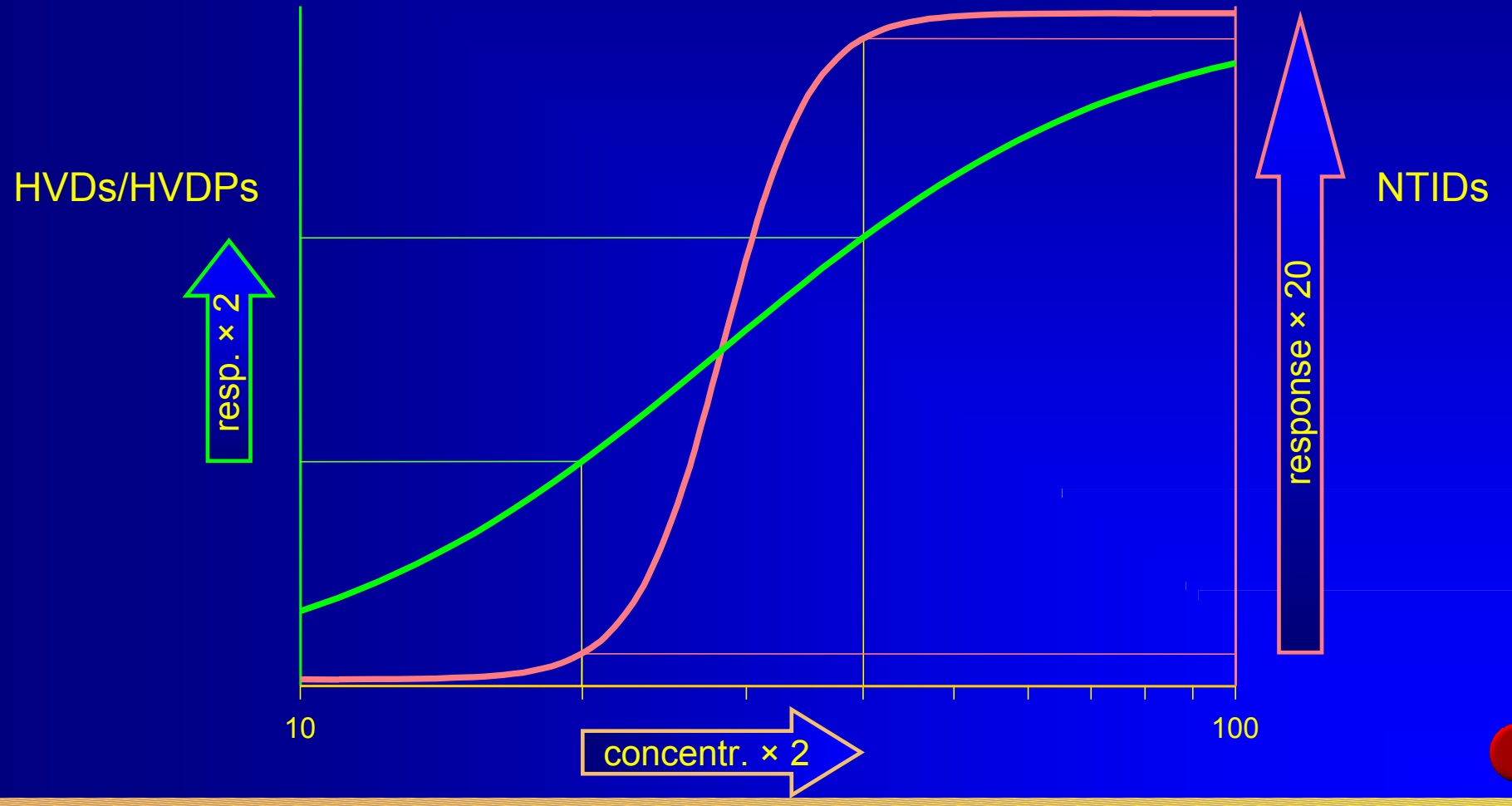
Modified from Fig. 1
Tothfálusi *et al.* (2009)

Counterintuitive concept of BE:

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

HVDs/HVDPs are safe

flat & steep PK/PD-curves



High variability

- For Highly Variable Drugs / Drug Products (HVDs/HVDPs) it may be almost impossible to show BE with a reasonable sample size.
- The common 2×2 cross-over design over assumes Independent Identically Distributions (IID), which may not hold. If *e.g.*, the variability of the reference is higher than the one of the test, one obtains a high common (pooled) variance and the test will be penalized for the 'bad' reference.

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. Note: Same overall number of individual treatments!

Replicate designs

- Any replicate design can be evaluated according to 'classical' (unscaled) Average Bioequivalence (ABE)
- ABE mandatory if scaling not allowed
  - FDA: s_{WR} <0.294 (CV_{WR} <30%); different models depend on design (*e.g.*, SAS **`Proc MIXED`** for full replicate and SAS **`Proc GLM`** for partial replicate).
  - EMA: CV_{WR} ≤30%; all fixed effects model according to 2011's Q&A-document preferred (*e.g.*, SAS **`Proc GLM`**).
  - Even if scaling is not intended, replicate design give more informations about formulation(s)

Application: HVDs/HVDPs

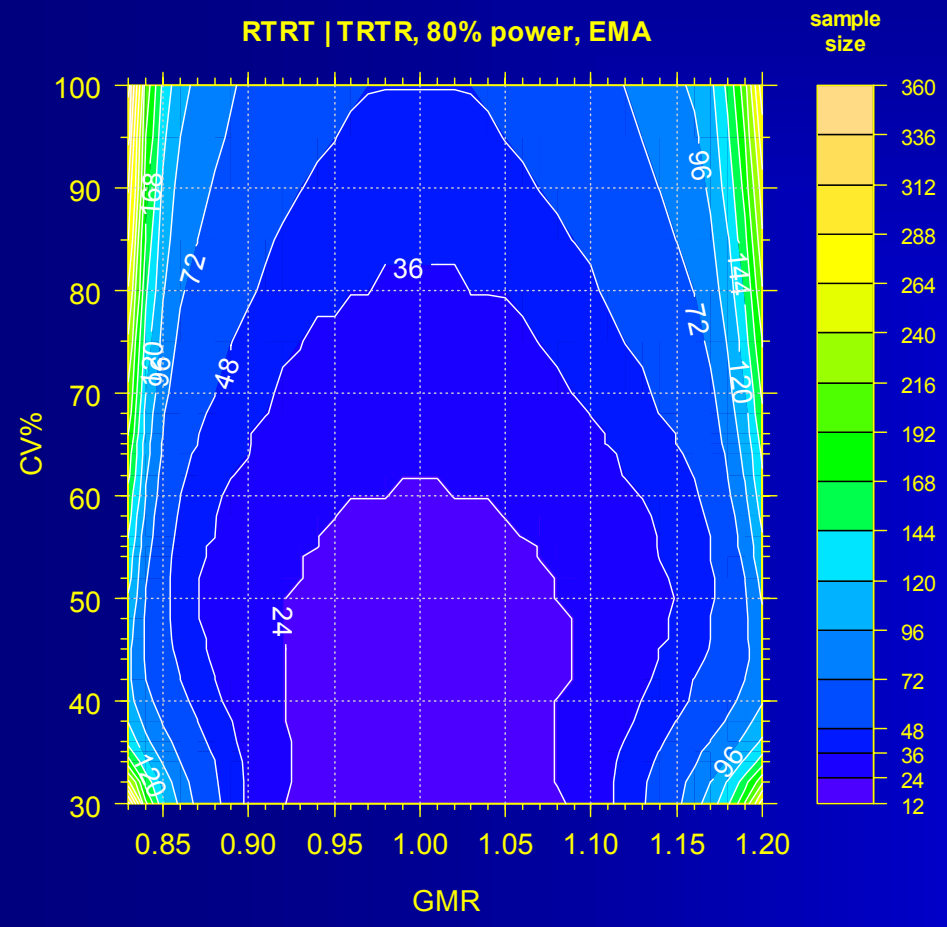
- CV_{WR} >30 %
  - ✓USA Recommended in API specific guidances. Scaling for AUC and/or C_{max} acceptable, GMR 0.80 – 1.25; ≥24 subjects.
  - ± EU Widening of acceptance range (only C_{max}) to maximum of 69.84% – 143.19%), GMR 0.80 – 1.25.
    Demonstration that CV_{WR} >30% is not caused by outliers.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

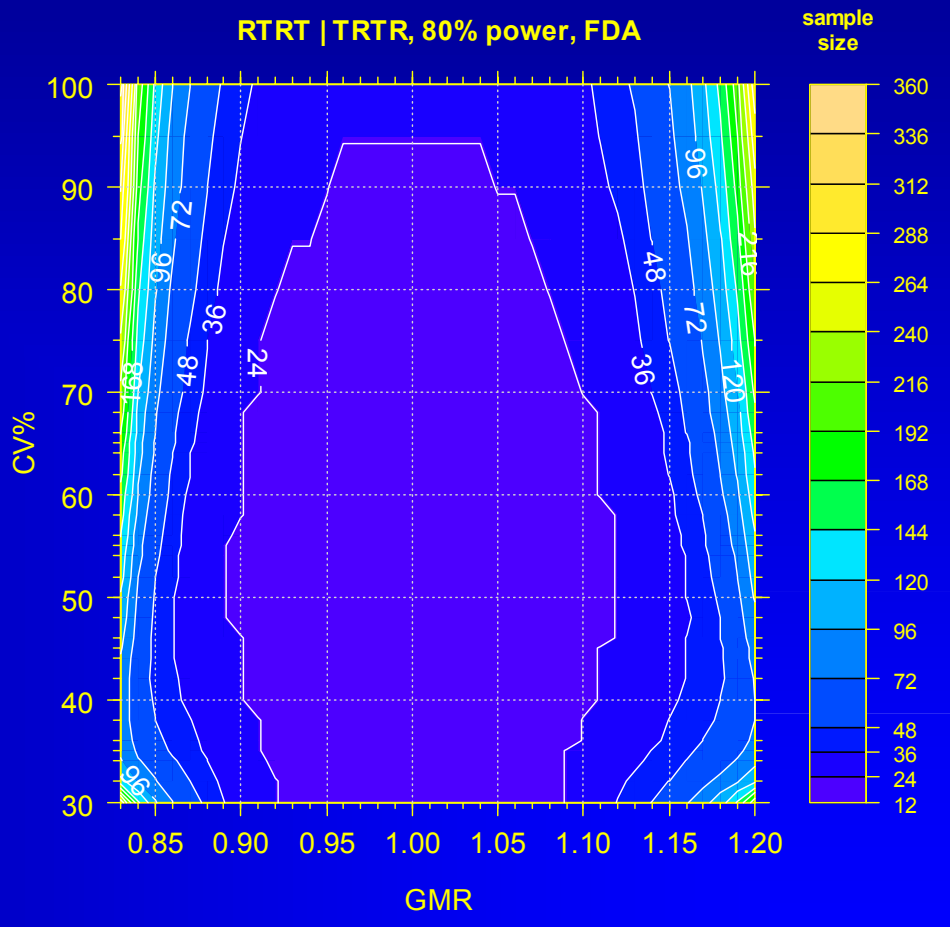
Replicate designs

- Two-sequence three-period
  T R T
  R T R
- Two-sequence four-period
  T R T R
  R T R T
- and many others…
  (FDA: TRR | RTR | RRT, aka ‘partial replicate’)
- The statistical model is complicated and depends on the actual design!

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

HVDPs (EMA/FDA; sample sizes)





HVDPs (EMA)

- EU GL on BE (2010)
  - Average Bioequivalence (ABE) with Expanding Limits (ABEL)
    - Based on σ_{WR} (the *intra*-subject standard deviation of the reference formulation) calculate the scaled acceptance range based on the regulatory constant k (θ_s=0.760); limited at CV_{WR} 50%.

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

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

HVDPs (EMA)

- Q&A document (March 2011)
  - Two methods proposed (Method A preferred)
    - Method A: All effects fixed; assumes equal variances of test and reference, and no subject-by-formulation interaction; only a common within (*intra*-) subject variance is estimated.
    - Method B: Similar to A, but random effects for subjects. Common within (*intra*-) subject variance and between (*inter*-) subject variance are estimated.
  - Outliers: Boxplots (of model residuals?) suggested.

Questions & Answers on the Revised EMA Bioequivalence Guideline
Summary of the discussions held at the 3rd EGA Symposium on Bioequivalence
June 2010, London
http://www.egagenerics.com/doc/EGA_BEQ_Q&A_WEB_QA_1_32.pdf

Example datasets (EMA)

- Q&A document (March 2011)
  - Data set I
    RTRT | TRTR full replicate, 77 subjects, imbalanced, incomplete
    - FDA
      s_{WR} 0.446 ≥0.294 → apply RSABE (CV_{WR} 46.96%)
      a. critbound –0.0921 ≤0 and
      b. PE 115.46% ⊂ 80.00–125.00% ✓
    - EMA
      - CV_{WR} 46.96% → apply ABEL (> 30%)
      - Scaled Acceptance Range: 71.23–140.40%
      - Method A: 90% CI 107.11–124.89% ⊂ AR; PE 115.66% ✓
      - Method B: 90% CI 107.17–124.97% ⊂ AR; PE 115.73% ✓

Example datasets (EMA)

- Q&A document (March 2011)
  - Data set II
    TRR | RTR | RRT partial replicate, 24 subjects, balanced, complete
    - FDA
      s_{WR} 0.114 <0.294 → apply ABE (CV_{WR} 11.43%)
      90% CI 97.05–107.76 ⊂ AR (CV_{intra} 11.55%) ✓
    - EMA
      - CV_{WR} 11.17% → apply ABE (≤30%)
      - Method A: 90% CI 97.32–107.46% ⊂ AR; PE 102.26% ✓
      - Method B: 90% CI 97.32–107.46% ⊂ AR; PE 102.26% ✓
      - A/B: CV_{intra} 11.86%

Outliers (EMA)

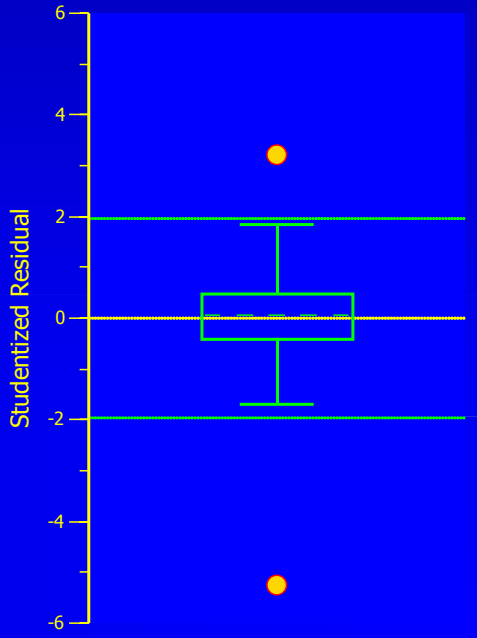
- EMA GL on BE (2010), Section 4.1.10
  - The applicant should justify that the calculated intra-subject variability is a reliable estimate and that it is not the result of outliers.
- EGA/EMA Q&A (2010)
  - Question:
    How should a company proceed if outlier values are observed for the reference product in a replicate design study for a Highly Variable Drug Product (HVDP)?

Outliers (EMA)

- EGA/EMA Q&A (2010)
  - Answer:
    The outlier cannot be removed from evaluation […] but should not be taken into account for calculation of within-subject variability and extension of the acceptance range.
    An outlier test is not an expectation of the medicines agencies but outliers could be shown by a box plot. This would allow the medicines agencies to compare the data between them.

Outliers (EMA)

- Data set I (full replicate)
  - CV_{WR} 46.96%
    EL 71.23–140.40%
    Method A: 107.11–124.89%
    Method B: 107.17–124.97%
  - But there *are* two outliers!
    By excluding subjects 45 and 52
    CV_{WR} drops to 32.16%.
    EL 78.79–126.93%
    Almost no more gain compared to conventional limits…



Thank You!

Statistical Analysis of BE Data

Open Questions?

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

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