Design of Comparative Bioavailability Studies

Excursion: Assumptions in Statistics

- All models rely on assumptions
  - Log-transformation allows for additive effects required in ANOVA
  - No carry-over effect in the model of crossover studies
    - Cannot be statistically adjusted
    - Has to be avoided by design (suitable washout)
    - Shown to be a statistical artifact in meta-studies
    - Exception: Endogenous compounds (biosimilars!)
  - Between- and within-subject errors are independently and normally distributed about unity with variances σ_b^2 and σ_w^2
    - If the reference formulation shows higher variability than the test, the ‘good’ test will be penalized (higher sample size) for the ‘bad’ reference
  - All observations made on different subjects are independent
    - No monocygotic twins or triplets in the study!

Error(s)

- All *formal* decisions are subjected to two 'Types' of Error
  - α = Probability of Type I Error (a.k.a. Risk Type I)
  - β = Probability of Type II Error (a.k.a. Risk Type II)
- Example from the justice system – which presumes that the defendant is *not guilty*:

Verdict	Defendant *innocent*	Defendant *guilty*
Presumption of innocence *rejected* (considered *guilty*)	wrong decision	correct decision
Presumption of innocence *accepted* (considered *not guilty*)	correct decision	wrong decision

Hypotheses

- In statistical terminology
  - Null hypothesis (H_0): Devendant is innocent
  - Alternative hypothesis (H_a a.k.a. H_1): Devendant is guilty

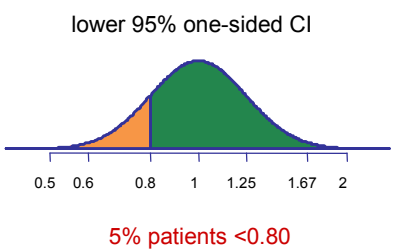
Decision	Null hypothesis *true*	Null hypothesis *false*
H_0 rejected	Type I Error	Correct (accept H_a)
Failed to reject H_0	Correct (accept H_0)	Type II Error

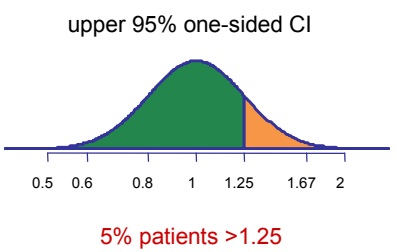
- In BE the Null hypothesis is bioinequivalence ($\mu_T \neq \mu_R$)!

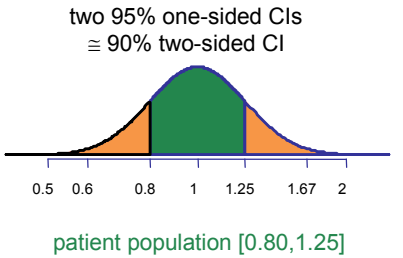
Decision	Null hypothesis *true*	Null hypothesis *false*
H_0 rejected	Patient's risk (α)	Correct (BE)
Failed to reject H_0	Correct (not BE)	Producer's risk (β)

Type I Error

- α Patient's risk to be treated with an inequivalent formulation (H_0 *falsely* rejected)
  - BA of the test compared to reference in a *particular* patient is considered to be risky *either* below 0.80 *or* above 1.25
    - If we keep the risk of *particular* patients at α 0.05 (5%), the risk of the entire *population* of patients (where BA <0.80 *and* >1.25) is 2α (10%) – expressed as a confidence interval: $100(1 - 2\alpha) = 90\%$
    - However, since in a patient BA cannot be <0.80 and >1.25 *at the same time*, the patient's risk from a 90% CI is still only 5%!







Type II Error

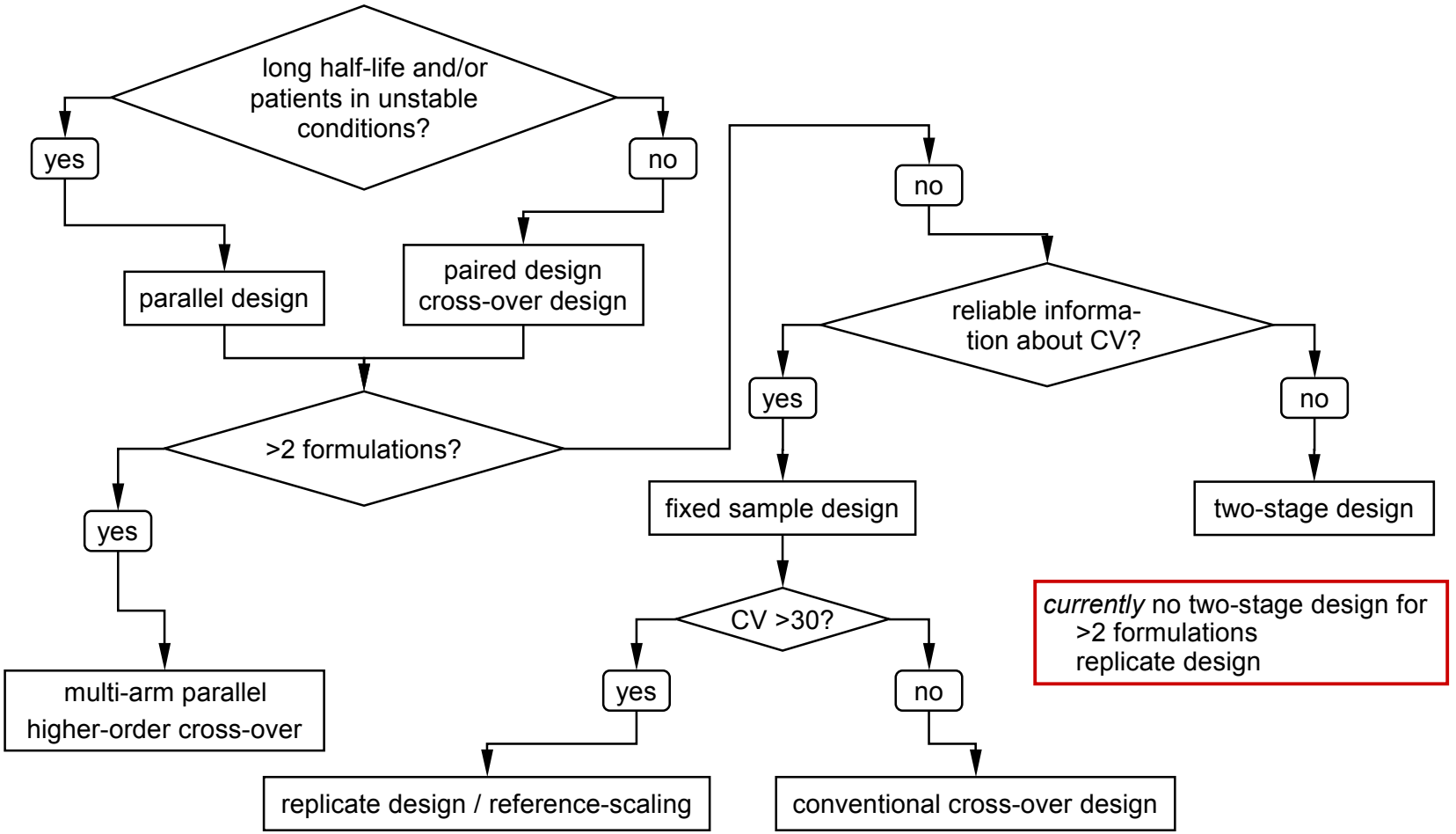
- β Producer's risk that an equivalent formulation is not approved (H_0 *falsely* not rejected)
  - Fixed in study planning to 0.1 – ≤0.2 (10 – ≤20%), where power = 1 – β = ≥80 – 90%
  - If all assumptions in sample size estimations turn out to be correct and power was fixed at 80%,
    one out of five studies will fail by pure chance!

α 0.05	BE
not BE	β 0.20

0.20 = 1/5

  - *A posteriori* (a.k.a. *post hoc*) power is irrelevant
    - *Either* a study has demonstrated bioequivalence *or not*
    - Calculating / reporting *a posteriori* power demonstrates a lack of statistical knowlege of the CRO

Designs: Selection



Designs: Background

- 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)
    2×2×2 crossover (TR | RT)
      Parallel (T | R)

- Variances which can be estimated

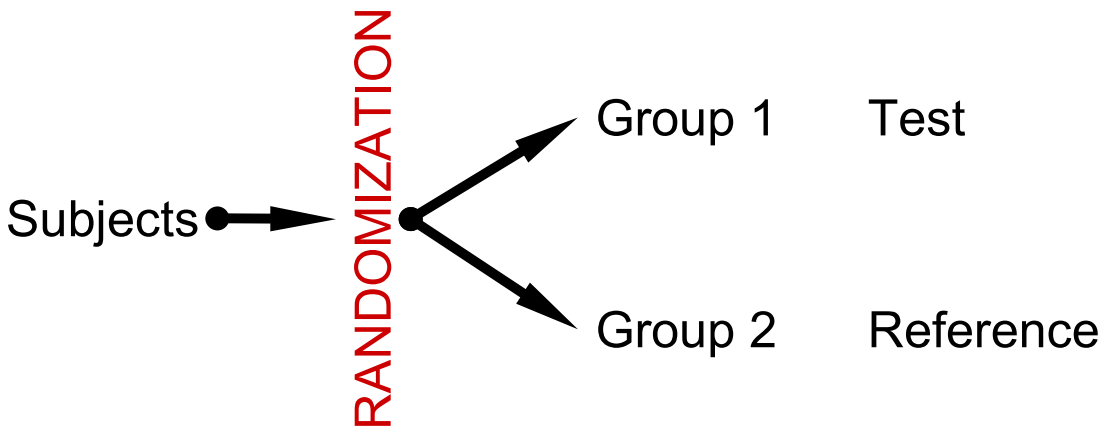
Parallel	*total* variance (*pooled* of between + within subjects)
2×2×2 crossover	+ *between*, *within* subjects
Partial replicate	+ *within* subjects (of R only)
Full replicate	+ *within* subjects (of R *and* T)

Parallel Designs

- One group is treated with the test formulation and another group with the reference
- Quite common that – due to dropouts the data set of eligble subjects is imbalanced, *i.e.*, $n_1 \neq n_2$
  - Equal variances should never be assumed (details in Presentation № 6)

Parallel Designs

- Example Two-Group Parallel Design



Parallel Designs

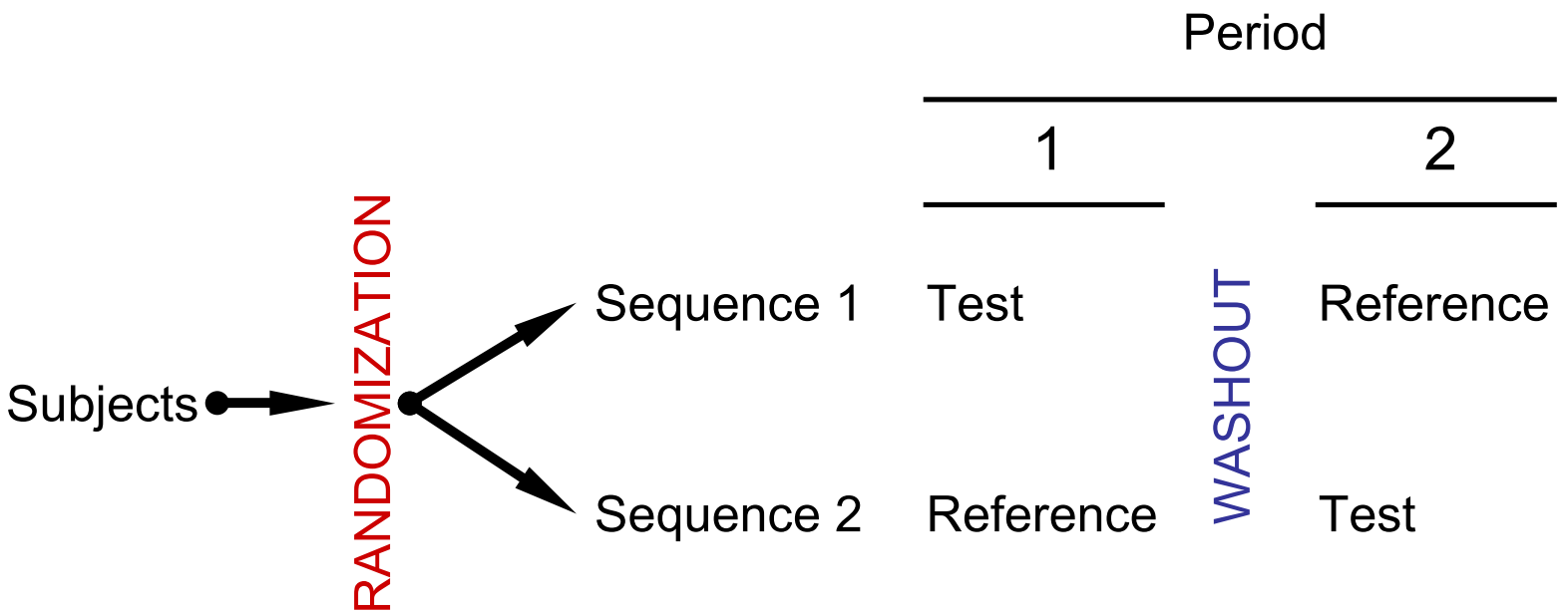
- Advantages
  - Clinical part – sometimes – faster than cross-over
  - Straigthforward statistical analysis
  - Drugs with long half life
  - Potentially toxic drugs or effect and/or AEs unacceptable in healthy subjects
  - Studies in patients, where the condition of the disease irreversibly changes
- Disadvantages
  - Lower statistical power than crossover design (rule of thumb: sample size should *at least* be doubled)
  - Pheno-/genotyping highly recommended for drugs showing polymorphism in metabolism

Crossover Designs

- Every subject is treated with all formulations
- In the most simple case (two formulations) subjects are randomized into two groups
  - One is receiving the formulations in the order TR and the other one in the order RT
  - These two orders are called *sequences*

Crossover Designs

- Standard 2×2×2 (2 treatments, 2 sequences, 2 periods)



Standard 2×2×2 Design

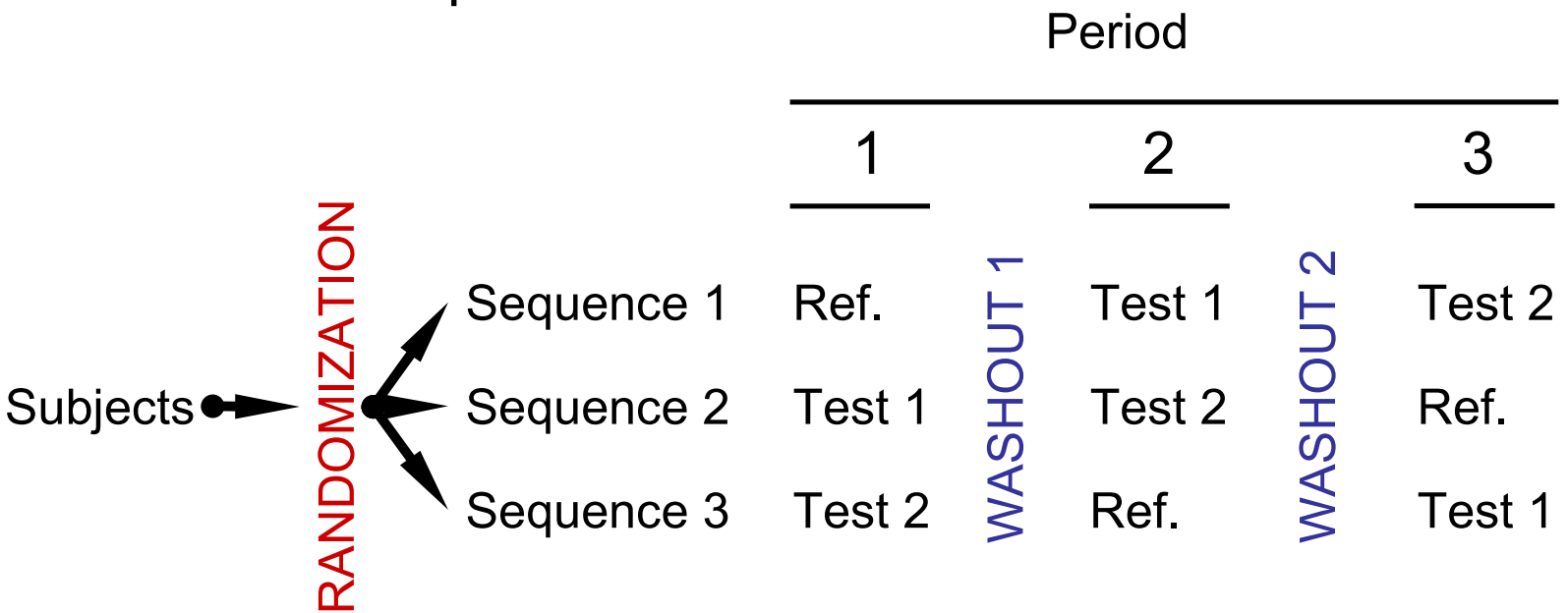
- Advantages
  - Globally applied standard protocol for bioequivalence, drug-drug interaction, and food effect studies
  - Healthy subjects and patients with a stable disease
  - Straigthforward statistical analysis
- Disadvantages
  - Not suitable for drugs with long half life (→ parallel design)
  - Not optimal for studies in patients with instable diseases (→ parallel design)
  - Not optimal for HVD(P)s (→ replicate designs)

Crossover Designs

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

Crossover Designs

- 3×3×3 Latin Square



3×3×3 Latin Square

- Advantages
  - Allows to choose between two candidate test formulations in a pilot study or comparison of one test formulation with two reference formulations (*e.g.*, the FDA's RLD and a European originator)
  - Number of subjects in the study is a multiplicative of three
  - Design for establishment of dose proportionality
- Disadvantages
  - Statistical analysis more complex (especially in the case of dropouts and a small sample size)
  - Not available in all software
  - Not mentioned in any guideline

Crossover Designs

- Variance balanced designs
  Example 3×6×3 Williams' design (three treatments)

Sequence	Period		
	1	2	3
1	Ref.	Test 2	Test 1
2	Test 1	Ref.	Test 2
3	Test 2	Test 1	Ref.
4	Test 1	Test 2	Ref.
5	Test 2	Ref.	Test 1
6	Ref.	Test 1	Test 2

Crossover Designs

- Variance balanced designs
  Example 4×4×4 Williams' design (four treatments)

Sequence	Period			
	1	2	3	4
1	Ref.	Test 3	Test 1	Test 2
2	Test 1	Ref.	Test 2	Test 3
3	Test 2	Test 1	Test 3	Ref.
4	Test 3	Test 2	Ref.	Test 1

Williams' Designs

- Advantages
  - Allows to choose between two candidate test formulations in a pilot study or comparison of one test formulation with two reference formulations (*e.g.*, the FDA's RLD and a European originator)
  - Design for establishment of dose proportionality
  - Mentioned in Brazil's (ANVISA), the EMA's, and the WHO's GLs
- Disadvantages
  - Mores sequences for an *odd* number of treatment needed than in a Latin Squares design (but equal for *even* number)
  - Statistical analysis more complex (especially in the case of dropouts and a small sample size)
  - Not available in all software

Interlude: Failed Studies

- Studies fail due to
  1. true bioinequivalence (CI completely outside the BE-limits)
  2. poor study conduct (increasing variability)
  3. pure chance (producer's risk…)
  4. over-optimistic assumptions about the variability and/or T/R-ratio
- Remedies
  1. Reformulate (another study is futile)
  2. Find a 'better' CRO
  3. – 4. Another study?
     Possibly unethical to repeat the study in a larger group of subjects

Interlude: Failed Studies

- Add-On Designs
  - Assess the study for BE
  - If it fails,
    - recruit another group of subjects
    - pool the data and assess for BE again
  - The patient's risk must be controlled
    - Already noticed at Bio-International Conferences (1989, 1992) and guidelines from the 1990s
    - Add-On Designs were shown to inflate the patient's risk *
    - Currently only recommended in Japan and Mexico

* Schütz H. *Two-stage designs in bioequivalence trials*. Eur J Clin Pharm. 2015;71(3):271–81. doi:10.1007/s00228-015-1806-2.

Group Sequential Designs

- Long and accepted tradition in phase III
  - Based on Armitage *et al.* (1969), McPherson (1974), Pocock (1977), O'Brien/Fleming (1979), Lan/DeMets (1983), Jennison/Turnbull (1999), …
- Fixed total sample size (*N*) and – in BE only one – interim analysis
  - Requires two assumptions
    - A 'worst case' *CV* for the total sample size and
    - A 'realistic' *CV* for the sample size in the interim

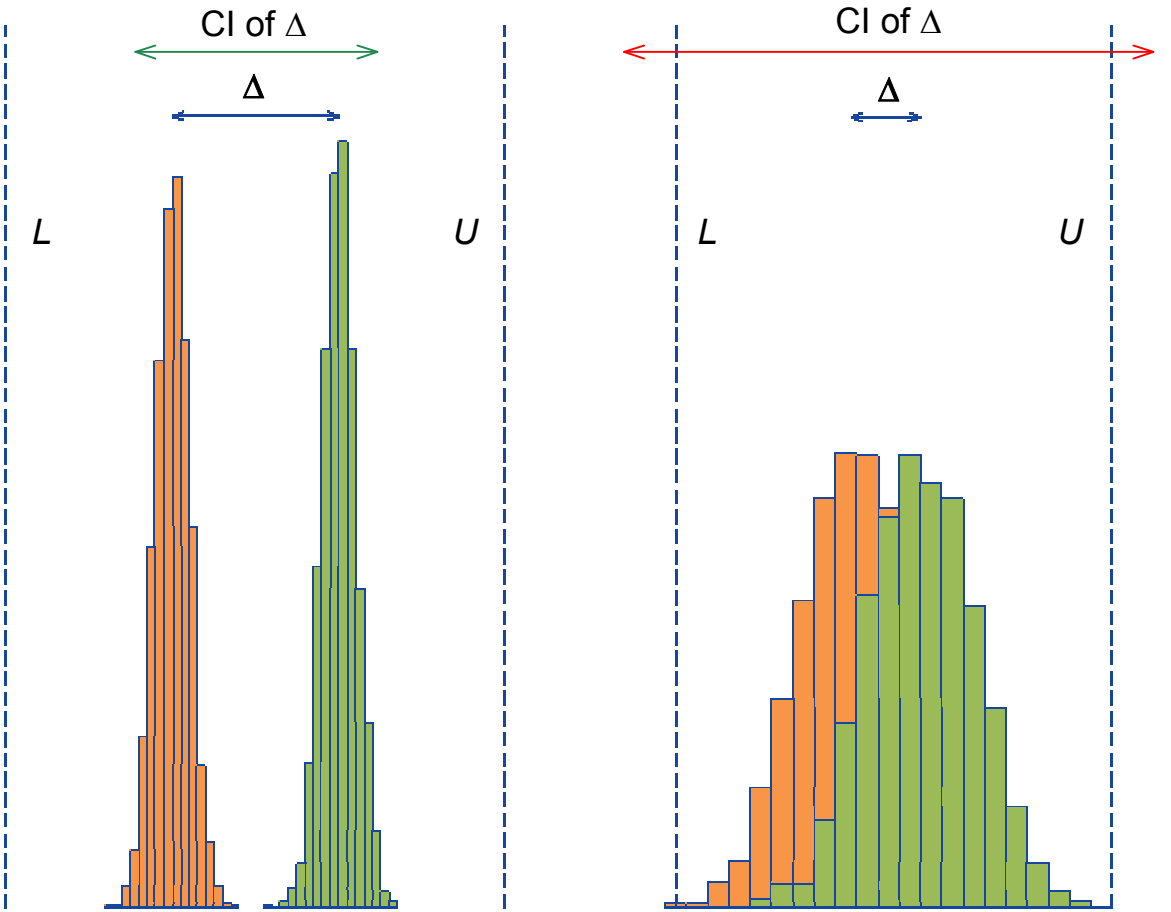
Group Sequential Designs

- All published methods were derived for superiority testing, parallel groups, normal distributed data with known variance, and the interim analysis at exactly *N*/2
  - That is not what we have in BE
    - Testing for equivalence (generally crossover) and lognormal data with unknown variance
    - Due to dropouts, the interim might not be exactly at *N*/2 (might inflate the Type I Error)
- Proposal by Gould (1995) in the field of BE did not get regulatory acceptance in Europe

(Adaptive) Sequential Two-Stage Designs

- Fixed stage 1 sample size (n_1) and sample size re-estimation in the interim analysis
  - Generally a fixed *GMR* is assumed
  - Published methods are valid only for a range of combinations of stage 1 sample sizes, *CV*s, *GMR*s, and desired power
  - With one exception (inverse normal method) no analytical proof of controlling the Type I Error exists
    - It is the responsibility of the sponsor to demonstrate (*e.g.*, by simulations) that the patient's risk is controlled
- Accepted by the WHO, FDA, EMA, Health Canada, Russian Federation, Eurasian Economic Union

Highly Variable Drugs / Drug Products



Counterintuitive concept of BE:

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

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

Highly Variable Drugs / Drug Products

- It may be almost impossible to demonstrate BE of HVD(P)s with a reasonable sample size
  - Example: *CV* 70%, *GMR* 0.90, target power 80%, 2×2×2 design

```
library(PowerTOST)
sampleN.TOST(CV=0.7, theta0=0.9, targetpower=0.9, design="2x2x2")
++++++++++++ Equivalence test - TOST ++++++++++++
            Sample size estimation
-----------------------------------------------
Study design:  2x2 crossover
log-transformed data (multiplicative model)
alpha = 0.05, target power = 0.8
BE margins = 0.8 ... 1.25
True ratio = 0.9,  CV = 0.7
Sample size (total)
 n     power
358   0.801175
```

  - Since HVD(P)s are considered to be safe and efficacious some jurisdictions accept a larger ‘not clinically relevant’ difference
    - The BE limits can be *scaled* based on the variability of the reference product

Highly Variable Drugs / Drug Products

- Requires a replicate design, where at least the reference product is administered twice (though not necessarily to all subjects)
  - Smaller sample sizes compared to the standard 2×2×2 design but outweighed by increased number of periods
  - Similar total number of individual treatments (hence, study costs drived by bioanalytics similar)
- Any replicate design can be evaluated for ‘classical’ (unscaled) Average Bioequivalence (ABE) as well

Highly Variable Drugs / Drug Products

- Reference-scaling (*i.e.*, widening the acceptance range based of the variability of the reference) accepted in many juriscdictions
  - *AUC* and C_{max}
    - FDA
  - C_{max} only
    - EMA, ASEAN States, Australia, Brazil, Egypt, Russian Federation, Eurasian Economic Union, East African Community, New Zealand
  - C_{max} (*AUC* if justified)
    - WHO
  - *AUC* only
    - Canada

Reference-scaling for HVD(P)s

- Different statistical approaches
  - FDA Reference-scaled average bioequivaence (RSABE)
  - All others Average bioequivalence with expanding limits (ABEL)
- RSABE requires commercial software (SAS, Phoenix/WinNonlin, JMP)
- ABEL can be evaluated by the package replicateBE for the open-source software R *

* Schütz H, Tomashevskiy, Labes D. *replicateBE: Average Bioequivalence with Expanding Limits (ABEL)*. https://cran.r-project.org/package=replicateBE.