Statistical aspects of equivalence testing in biosimilar studies: multiple, differently-scaled endpoints

Ludwig A. Hothorn
hothorn@biostat.uni-hannover.de
retired Leibniz University Hannover, Germany

9. Oktober 2018

Before going into details I

- Expecting a mixed audience:
  my slides varies from naive to rather complex
- Apologies for the others
- No just biosimilar issues, more general multiple endpoint equivalence. Even looking over the edge

State of the art in bioequivalence I

- Claiming bioequivalence for a single, selected PK endpoint.
  Commonly AUC within $[0.8, 1.25]$
- Claiming equivalence for two PK measures, commonly AUC and Cmax independently, each at level α

- **Challenge today: simultaneous claim**

- Commonly COD. In pre-clinical in-vivo studies commonly parallel group design (my background) $\Rightarrow$ today

State of the art in bioequivalence II

- **Recent example:** Pharmacokinetics..... equivalence trial of the biosimilar MYL-1401H vs. reference pegfilgrastim [40]

Table 1 Summary of the pharmacokinetic parameters for pegfilgrastim in serum (PK analysis set)

Parameter	MYL-1401H ($N=204$)	EU-reference ($N=203$)	US-reference ($N=207$)	MYL-1401H/EU-reference		MYL-1401H/US-reference	
				LS mean ratio	90% CI	LS mean ratio	90% CI
Primary pharmacokinetic end points							
C_{max} (%CV), pg/mL	36.7 (72.1)	34.2 (72.1)	37.3 (67.6)	1.07	0.98–1.16	0.99	0.91–1.07
AUC_{0-inf} (%CV), h·ng/mL	869 (69.1)	833 (70.1)	876 (66.3)	1.04	0.98–1.11	1.00	0.94–1.07
Secondary pharmacokinetic end points							
AUC_{0-t} (%CV), h·ng/mL	827 (71.4)	787 (72.7)	832 (68.6)	1.05	0.98–1.13	1.00	0.93–1.07
t_{max}, median (range), h	12.0 (6.0–24.0)	12.0 (6.0–48.0)	12.0 (4.0–24.0)	–	–	–	–
k_{el} (%CV), h^{-1}	0.014 (31.0)	0.014 (39.1)	0.014 (40.1)	1.03	0.98–1.08	1.04	0.99–1.09
$t_{1/2}$ (%CV), h	49.3 (36.5)	51.1 (48.9)	51.0 (42.5)	0.97	0.93–1.02	0.97	0.92–1.01
V_d/F (%CV), L	164 (100)	177 (101)	168 (113)	0.93	0.85–1.02	0.99	0.89–1.06

- i 2 primary endpoints: **each independent at level** α
  ii 3 formulations, i.e. multiple comparisons: **without any adjustment**

State of the art in bioequivalence III

iii Large n_i, narrow CI. No a-priori power information ($n_i \Uparrow \Leftrightarrow CI \Downarrow$)
Let us discuss RCT without a-priori power consideration now

iv Variance homogeneity? (almost balanced design)

vi $CV_{AUC} << CV_{Cmax}$ **without any consequences**

* Remember a different story in equivalence testing: 30 day toxicity in-vivo assay with up to 100 endpoints (on different scales).

1. Actually, a multivariate equivalence test should be used.
2. But one takes per-endpoint PoH, each independent at α level
3. Since the CVs are naturally very different, very different f- rates result, which are of primary interest (*Be safe in negative results*).
4. Creepy, but recent guidelines, publications and big companies do so. Even creepier

vii Classification between primary (for claim) and secondary endpoints (just to report) is missing

viii T_{max} without estimates and CI

ix ....

Multiplicity issues I

- **At least three sources of multiplicity**
  - i q primary endpoints. Correlated endpoints (otherwise stop here with Bonferroni)
  - ii k formulations, i.e at the most $k(k-1)/2$ all pairs comparisons (or user-defined less). Correlated treatment comparisons (otherwise stop here with Bonferroni)
  - iii lower AND upper test in each TOST, ie joint testing
  - * Notice, joint correlations between treatment comparisons, endpoints and two-one-sided tests should be taken into account, at least partly

Multiplicity issues II

- **Consequences of multiplicity adjustment**
  - i We pay inherently a price when taking multiplicity into account, i.e. confidence interval becomes conservative.
    I.e. from the point of view of a poor industrial statistician, I would only do a multiplicity adjustment when the authorities require it.
    Although, of course, the claim will be better (see below the wording of WHO guideline)
  - ii General strategy. Controlling FWER, but one tries to limit conservativeness by
    1. minimum number of tests (e.g. a-priori or importance order)
    2. consideration the correlation
    3. choosing most appropriate test statistics (confidence interval estimates)
  - iii This usually becomes complicated and requires numerous assumptions which are either not verifiable or not given in real (small n_i) data

Multiplicity issues III

iv On the other hand, Bonferroni works always and a simple, but unnecessarily conservative

v Notice, NHST and sCI not always compatible. Prefer sCI!

vi However, looking to current phase III efficacy RCT: the majority use a single primary endpoint, although a claim on multiple endpoints would be needed (or relevant). It would be more obvious to use a multivariate test for 2 primary efficacy endpoints than two PK parameters for biosimilars. **First, the big ones should do their homework**

Properties of motivating examples I

- **Example 1:** Two formulations of ticlopidine in COD [26]

TABLE 1 Ticlopidine hydrochloride data: results of univariate analyses separately for AUC_{0-t}, $AUC_{0-\infty}$, and C_{max}: point estimates, standard errors, TOST P values, lower and upper bounds of 90% confidence intervals

	$\widehat{\theta}$	SE($\widehat{\theta}$)	P	Lower	Upper
AUC_{0-t}	−0.080 (0.923)	0.059	0.012	−0.181 (0.834)	0.020 (1.021)
$AUC_{0-\infty}$	−0.068 (0.934)	0.064	0.012	−0.178 (0.837)	0.042 (1.042)
C_{max}	−0.094 (0.910)	0.066	0.031	−0.207 (0.813)	0.018 (1.019)

Estimates and interval bounds are given on the logarithmic scale (in round brackets on the original scale).

* *3 highly correlated PK measures*
  $\rho_{AUC,AUCinf} = 0.97; \rho_{AUC,Cmax} = 0.81$
* Assuming log-normal distribution for all 3 measures
* Here $CV_{AUC} \approx CV_{Cmax}$
* Disadvantage of p-value as criterion obvious: $p_{AUC} = p_{AUCinf}$ but $R = 0.923$ would make me nervous
* Details in Phillip's talk

Properties of motivating examples II

- **Example 2:** Biosimilar RCT [28] **PK and efficacy** of innovator infliximab (INX) and biosimilar CT-P13 in patients with active ankylosing spondylitis.

* Ratios of
AUC $[94\%; 116\%)]$
Cmax $[95\%; 109\%)]$
ASAS20-criterion: $p_{CTP13} = 70.5\%, p_{INX} = 72.4\%$ (no CI!)
* Conclusions: PK profiles of CT-P13 and INX equivalent as well as efficacy (ASAS20) comparable. Independent, each at level α
* No raw data, no correlation, nor really multiple endpoint equivalence- but appropriate biosimilar example
* In fact, even multiple efficacy endpoints and safety endpoints. How unlikely is that ALL are equivalent? (later discussion)
* Different-scaled endpoints: continuous, log-normal and proportion(s)!

Properties of motivating examples III

- **Example 3:** Nutritional assessment of GMO vs. isogenic varieties [13]

Non-parametric ratio estimates and marginal $(1 - 2\alpha^{\text{satety}})$ confidence interval for all components

Component	Ratio	LowerCI	UpperCI
Moisture	0.98	0.95	1.02
Total fat	1.00	0.98	1.01
Protein	0.99	0.95	1.03
Ash	0.99	0.98	1.02
Carbohydrates	1.01	1.00	1.02
Acid detergent fiber	0.96	0.92	1.00
Neutral detergent fiber	1.10	1.05	1.17
Calcium	0.96	0.94	0.98
Phosphorus	0.97	0.95	0.99
Potassium	1.01	0.98	1.04
Magnesium	0.99	0.96	1.03
Sodium	0.94	0.89	1.00
Iron	1.02	0.99	1.05
Manganese	1.00	0.96	1.04
Copper	1.02	1.00	1.05
Zinc	1.02	0.97	1.08
α-Tocopherol	1.05	1.02	1.08
β-Tocopherol	n.e.	n.e.	n.e.
γ-Tocopherol	0.99	0.96	1.03
δ-Tocopherol	n.e.	n.e.	n.e.
Phytic acid	0.98	0.95	1.01
Alkenyl glucosinolate	0.84	0.77	0.91
Aromatic glucosinolate	0.99	0.98	1.03
Indolyl glucosinolate	0.94	0.82	1.09
Total glucosinolate	0.87	0.79	0.93
Alanine	1.00	0.96	1.04
Arginine	0.98	0.94	1.03
Aspartic acid	1.03	0.98	1.08
Cystine	0.94	0.90	0.98
Glutamic acid	0.97	0.92	1.02
Glycine	1.00	0.96	1.04
Histidine	0.97	0.93	1.01
Isoleucine	0.98	0.94	1.02
Leucine	0.99	0.95	1.03
Lysine	0.98	0.95	1.02
Methionine	0.97	0.94	1.00
Phenylalanine	0.99	0.94	1.04
Proline	0.97	0.93	1.01
Serine	1.03	0.99	1.06
Threonine	1.01	0.98	1.05
Tryptophan	0.99	0.95	1.04
Tyrosine	0.99	0.96	1.03
Valine	0.98	0.94	1.03
C16:0 Palmitic	1.05	1.03	1.07
C16:1 Palmitoleic	1.02	0.95	1.05
C17:1 Heptadecenoic	1.04	0.93	1.07
C18:0 Stearic	1.01	0.99	1.03
C18:1 Oleic	1.00	0.98	1.02
C18:2 Linoleic	1.02	1.01	1.04
C18:3 Linolenic	1.00	0.97	1.01
C20:0 Arachidic	1.01	0.98	1.03
C20:1 Eicosenoic	1.02	0.99	1.03
C22:0 Behenic	1.02	0.96	1.04
C22:5 Docosapentaenoic	1.03	1.00	1.06
C22:6 Docosahexaenoic	1.03	1.00	1.06
C24:0 Lignoceric	0.97	0.94	1.06

n.e., not estimable.

Properties of motivating examples IV

* Assuming the same normal distribution for all endpoints?
* Here, it is obviously unrealistic to demand equivalence for all endpoints simultaneously
* For all endpoints the same thresholds $[1/\delta, \delta]$. Really?

Facts on bioequivalence I

- Bioequivalence definition, e.g. WHO

> *...two pharmaceutical products are bioequivalent if ...their bioavailabilities, in terms of rate (Cmax and tmax) and extent of absorption (area under the curve) ... are similar to such a degree that their effects can be expected to be essentially the same*

(WHO Technical Report Series No. 996, 2016, Annex 9)

- Already:
  1. 3 primary endpoints
  2. degree of similarity not formal defined (no explicit choice of δ)
  3. one could already understand the and as simultaneous
- Endpoint- and/or condition-depending choice of δ:
  1. WHO: $[0.8, 1.25]$ for AUC and Cmax
  2. High variable drugs: Cmax $[0.698, 1.43.2]$ for
  3. FDA: Tmax should be similar (no formal δ at all)

Facts on bioequivalence II

- δ definition implies μ_T/μ_R as effect size

Issue I: Different-scaled endpoints on the same multiplicative effect size- really comparable?

- TOST, ie. 90% two-sided CI **within** [lower, upper], ie interval inclusion criterion

Issue II: Claiming equivalence depends both on estimates (data, statistics) **and** choice of δ. Notice, a general principle, only in case of superiority test one cheats through with $\delta = 0$. **Was and is a devastating concept, still common used**

Facts on bioequivalence III

- **TOST is IUT**, i.e. $H_A : H^1$ AND H^2
  Notice, the opposite is the common-used UIT (OR...OR...OR).
  See slides below
- Majority assumes AUC is log-normal distributed $\Rightarrow$ CI(t-test) for $y_{ij} = log(AUC_{ij})$
- Alternatives exists, e.g. ratio-to-control CI assuming normal distribution (allowing variance heterogeneity) [1] or any distribution (rank statistics)

- Claiming equivalence for multiple correlated endpoints is challenging. The pros and cons of multivariate tests and intersection-union tests are discussed. Sometimes bioassays with multiple biosimilars with respect to a single comparator are considered. The pros and cons of multiplicity adjustment will be discussed.

Questions so far?

Design Issues I

- May be many endpoints, even up to $p > n$ problem (GMO risk assessment) $\Rightarrow$ bivariate today only
- May be k-sample design $\Rightarrow$ 2-sample design today only
- May be complex layouts (COD) $\Rightarrow$ simple randomized design today only
- May be adaptive [23] . Not adaptive today only
- I.e. the following approach is without limitation of generalizability

Multiple endpoints issues I

- My selection criteria:
  - i sCI instead of tests
  - ii per-endpoint sCI connected with specific per-endpoint thresholds (even to be more conservative)
  - iii taking the correlation between endpoints into account
  - iv not only global claim, but also subset claims (eg. similar for AUC, but not for Cmax (at the lower limit only)). Ie. more directional (due to importance) as noninferiority claim
  - v not assuming Gaussian distribution with homogeneous variances,
  - vi not only considering location effects
  - vii not believing $[0.8, 1.25]$ is valid for all endpoints

Multiple endpoints issues II

- Wording *multiple endpoint* vs. *multivariate*:
  i) the second implies multivariate normal distribution and a global test,
  ii) the first implies any-distributed y_{ij} as well as global and endpoint-wise decisions $\Leftarrow$ today

Multiple endpoints issues III

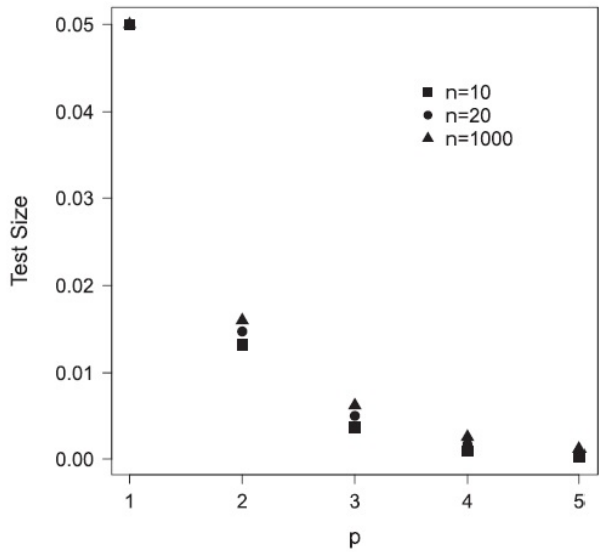
- Strategy:
  - i all endpoints equivalent (just one global IUT) or
  - ii at least k of p equivalent [32] (using Bonferroni for the subsets) or
  - iii at least one, any-one, up to all (UIT(IUT)) complete power approach [11]
- Relevant published tests
  1. Most advanced approach: [26]
     i) realistic: smallish sample sizes, unknown and potentially heterogeneous variances, and highly correlated PK measures (0.973 AUC,AUC; 0.808 AUC,Cmax),
     ii) assuming multivariate normal,
     iii) simult CI for each endpoint,
     iv) *recommend treating it as joint confidence regions for the joint parameter vector with respect to predefined margins* $[-\Delta, \Delta]$ *rather than marginal simultaneous CIs for individual... and there is no inherent need to derive univariate intervals* I disagree

Multiple endpoints issues IV

v) projection into rectangular regions make it conservative, but keep it interpretable (oh yes!),

vi) all-or-nothing criterion contradiction to sCI

1a IUT is an extreme conservative approach by function `iutsize`



2 A PK parameter AND a efficacy parameter (adaptive seamless design for establishing PK and efficacy equivalence in biosimilars) [36]

Multiple endpoints issues V

3 Ratio and difference of two AUCs in 2-sample design allowing variance heterogeneity assuming normal distribution [15]; Already in [12]
4 Power for TOST with correlated endpoints [30]. Pioneering paper, was later ignored.
5 Equivalence for functional data (lung volume test). Frequentist Bonferroni pointwise $(1 - 2\alpha)$ nonparametric bootstrap [10]
6 Equivalence for high-dimensional expression data: F- or range test using standardized squared Euclidean distance, (implicitly assuming multivar normal homo vars) a single moment-based difference ratio (DR) criterion of 1.25 [43]. Direct comparisons only, simplified to randomized parallel group design (no interim analysis, no adaption)
7 Recent: [16] [18]

Multiple endpoints issues VI

8 [41] multivar test may be powerful, but for each endpoint within L; U needed
9 More: [26], [25], [3], [5], [6], [7], [8]
10 Which effect size?
11 Further more: [11], [**?**], [13], [17], [24], [32], [35], [37], [39]
12 Recent permutation approach [2]
13 Really FDR [38]?
14 Meta analysis on multiple references (proportion) [42]
15 2 AUC with log and t [15]
16 New [3] and [30] and [42] and [2]

Multiple endpoints issues VII

i Should we ignore σ? i) Certainly not, use studentization. ii) Should we use σ directly e.g $\frac{\mu_R - \mu_T}{\sigma}$? iii) should be test $\frac{\mu_R}{\mu_T}$ and $\frac{\sigma_R^2}{\sigma_T^2}$ e.g. by IUT?, iv) are the first 2 moments sufficient, yes if normal distribution, otherwise...., iv) consider any .. mlt

ii Population bioequivalence in the univariate case is a comparison of two distributions that simultaneously compares means and variances

iii margins symmetric, or?

Coffee break about here

Sorry, I must first torture you with multiplicity

Some background about multiplicity I

- The ultimate goal: control FWER
- What is a family? Some people believe in all-pair comparisons when considering k treatments. Eg., for a design $[P, D_1, D_2, D_3, C]$
- I like claimwise error rate [29]. Possible claims:
  1. Trial sensitivity $\mu_C > \mu_P$?
  2. Superiority which $\mu_{D_i} > \mu_P$
  3. At least noninferiority μ_{D_i} is at least noninferior to μ_C
     * But not all-pairs comparisons
- For biosimilars with 2 primary endpoints and 3 formulations (see the example above): how to define claimwise error rate?
- Most textbook ugly: only treatment comparisons, only UIT

Some background about multiplicity II

- **I) Much more sources of multiplicity:**
  - i endpoints
  - ii time points
  - iii subgroups
  - iv adaptive designs
  - v test principles (eg with or without considering baseline covariate adjustment)
  - vi multiple tuning parameters (eg. poly-k adjustment in carcinogenicity bioassays)
  - ... etc.
- Goal: joint considering all sources of multiplicity and using the entire correlation matrix, see treatment-by-time eg. Phillip's PhD thesis on www.biostat.uni-hannover.de or the paper [27]

Some background about multiplicity III

- **II) UIT** $\Rightarrow$ max-t-test: $T^{max} = max(t_1, t_2, ..., t_\phi)$
- Hypothesis $H_0 = \bigcap_{i=1}^{\phi} H_{0i}$
  Simple reject whether H_0^1OR H_0^2,OR,....,H_0^ϕ **at least one, anyone**
- Eg. Dunnett-type comparison $[0, i]$ rejection when
  $T_i = \frac{|\bar{X}_i - \bar{X}_0| - \delta}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_0}}} > t_{\phi,\nu,\mathbf{R},1-\alpha}$
  with the lower $(1-\alpha)$ quantile $t_{\phi,\nu,\mathbf{R},1-\alpha}$ of an underlying ϕ-variate t-distribution with correlation matrix $\mathbf{R}$
- $\mathbf{R}$ can be simple (Dunnett) or complex (see below)
- First, it is an univariate test (univariate skewed t-distributed, see M Hasler PhD thesis on www.biostat.uni-hannover.de)

Some background about multiplicity IV

- Because Gabriel's theorem (monotonicity of quantiles in ϕ) can the elementary hypotheses rejected- not only a global. I.e. sCI are available:
  $(\bar{X}_i - \bar{X}_0 \pm t_{\phi,\nu,\mathbf{R},1-\alpha}\, S\sqrt{\frac{1}{n_i} + \frac{1}{n_0}})$
  And adjusted p-value; in relevant cases compatible (eg. but not for stepwise procedures)
- Estimating $t_{\phi,\nu,\mathbf{R},1-\alpha}$ with known, a-priori calculated, **R** `library(multcomp)` or estimated from data(model) `mmm()`, allows correlated models, ie. lmm
- Basic property: $t_{\phi,\nu,\mathbf{R},1-\alpha} \Rightarrow t_{\nu,1-\alpha/\phi}$ when $\mathbf{R} \Rightarrow 0$
  and $t_{\phi,\nu,\mathbf{R},1-\alpha} \Rightarrow t_{\nu,1-\alpha}$ when $\mathbf{R} \Rightarrow 1$
- Everything we do in UIT-research
  1. Reducing dimension ϕ, e.g. stepup, stepdown, a-priori importance, claimwise formulation,...
  2. Using $\mathbf{R} > 0$

Some background about multiplicity V

- **An example for correlated models using mmm**:
  Dose-response with two primary endpoints: weight (normal) and malformations (proportion)

	Litter	Dose	Weight	Malformation
1	60	0	0.90	0
2	60	0	0.83	0
3	60	0	0.95	0
4	60	0	0.95	0
5	60	0	1.07	0
6	60	0	1.06	0
..	...	...	...	...
1017	156	3000	0.81	1
1018	156	3000	1.00	0
1019	156	3000	0.88	0
1020	156	3000	0.79	0
1021	156	3000	0.90	0
1022	156	3000	0.86	0
1023	156	3000	0.86	0
1024	156	3000	0.80	0
1025	156	3000	0.84	0
1026	156	3000	0.87	0
1027	156	3000	0.72	0
1028	156	3000	0.83	0

Doses: 0,750,1500,3000

Some background about multiplicity VI

```
data("wuleon", package="SiTuR")
library("multcomp"); library(tukeytrend)
Wul <- dosescalett(wuleon, dose="Dose", scaling=c("ari", "ord", "arilog"))$dat
wN <-lm(Weight~Doseari, data=Wul)
wO <-lm(Weight~Doseord, data=Wul)
wLL <-lm(Weight~Dosearilog, data=Wul)
pN <-glm(Malformation~Doseari, family=binomial(),data=Wul)
pO <-glm(Malformation~Doseord, family=binomial(),data=Wul)
pLL <-glm(Malformation~Dosearilog, family=binomial(),data=Wul)
wuWeMa <- glht(mmm(covarWe=wN, ordinWe=wO, linlogWe=wLL,
          covarMa=pN, ordinMa=pO, linlogMa=pLL),
mlf(covarWe="Doseari=0", ordinWe="Doseord=0", linlogWe="Dosearilog=0",
covarMa="Doseari=0", ordinMa="Doseord=0", linlogMa="Dosearilog=0"))
```

	Model	Test stats	p-value
1	covarWe: Doseari	-28.6898	0.0000
2	ordinWe: Doseord	-30.6301	0.0000
3	linlogWe: Dosearilog	-30.6301	0.0000
4	covarMa: Doseari	14.6421	0.0000
5	ordinMa: Doseord	14.2400	0.0000
6	linlogMa: Dosearilog	14.2400	0.0000

Some background about multiplicity VII

- **III) IUT:** $\Rightarrow$ min-t-test: $T^{min} = min(t_1, t_2, ..., t_\phi)$
  (Notice, min-p-test is similar to maxT test!)
- Hypothesis $H_0 = \bigcup_{i=1}^{\phi} H_{0i}$
  Simple reject whether H_0^1AND H_0^2,AND,....,H_0^ϕ **for all**
- Reject $t_1 < t_{\nu,1-\alpha} \bigcap t_2 < t_{\nu,1-\alpha} ... \bigcap ...$
- Bioequivalence TOST=2-sample-IUT

Some background about multiplicity VIII

- **IV) Comparing UIT and IUT**
  - i Both become conservative with $\Uparrow \phi$:
    i) UIT in terms of $t_{\nu,1-\alpha/\phi}$,
    ii) IUT in terms of $t_1 < t_{\nu,1-\alpha} \bigcap t_2 < t_{\nu,1-\alpha} ... \bigcap ...$ **even more clearly** (see below)
  - ii UIT allows information on all elementary tests by adj.p-values or sCI, but IUT allows only ONE global outcome. Serious limitation of IUT- remember multiple endpoint equivalence: either ALL endpoints are equivalent or...
  - iii Compromise approach:**all-pairs power UIT**. Used for multiple endpoint noninferiority test [11]. In the sense of the Lui 2011 approach [20]- never referenced
    Common UIT is any-pairs power assumption: at least one H_1^i, anyone. Here: all H_1^i

Some background about multiplicity IX

An rather recent example (not yet published)

1. The US-FDA 2001 guidance recommended the evaluation of individual tumors in long-term carcinogenicity bioassays by a trend test **or** pairwise comparison of high dose with control
2. Alternative decision rule for a strict monotone dose-response relationship, a trend test **and** pairwise test simultaneously was proposed recently [19] (joint test)
3. This logical AND operation represents an IUT. The elementary tests within an IUT are performed at level α to control FWER.
4. Simulation normal distributed homoscedastic errors in a balanced $k = 3 + 1, n_i = 20$ design
5. Tests used: i) LinReg ... linear regression alone, ii) UIT...lin regression OR HvsC contrast (UIT), iii) IUT... lin regression AND HvsC contrast (IUT)

Some background about multiplicity X

Shape	n	LinReg	UIT	IUT
H0	20	0.049	0.049	**0.027**
lin	20	**0.946**	0.941	0.893
0,0,0,d	20	0.894	0.920	0.849
0,0,d,d	20	0.988	0.984	0.910
0,d,d,d	20	0.906	0.926	0.859
0,0,d,2/3d	20	0.219	0.171	0.035
0,0,1/3d	20	0.000	0.000	0.000
0,0,d,4/5d	20	0.808	0.752	0.446
0,d,d,4/5d	20	0.386	0.412	0.280

6. UIT is less conservative (not surprising, since he uses the correlation); both are tests for monotonic trends; UIT allows trend AND CvsH claim (adj p-value, sCI) IUT only ONE.

iv Valid only for these two scenario ([11]), but can be used as an idea

v Notice IUT, taking correlation into account is needed ... (I failed)

Some background about multiplicity XI

Summary of multiplicity issue

1. The actual test IUT(IUT(IUT))) is terribly conservative, especially because we cannot use the correlations, and it only allows a global statement (claim)
2. In biosimilar trials for multiple endpoints and multiple treatments I propose **UIT(UIT(IUT)))**:
   I) UIT() for correlated multiple endpoints,
   II) UIT(UIT()) for multiple treatment comparisons (within multiple endpoints),
   III) IUT ... TOST for equivalence
3. Can be analysed by function `mmm` (within `library(multcomp)`) for $(1 - 2\alpha)$ CIs (equi-TOST)
4. Allows all-hypotheses claim or any subset claim, while controlling FWER. Following perfectly the claimwise error rate concept
5. More work needed next

Questions so far?

Going on with provocations

A most extreme pseudo-multiple endpoints approach I

- Sometimes I look at data from the perspective of an empirical analyst quite naively: there is an **univariate endpoint, repeatedly measured on the same subject, several subjects within each factor level**.
- What am I doing? t-test-type intervals in the mixed model, with log(endpoint) to get into the multiplicative model.
- That's it. No kinetics assumptions, no multiple pseudo-endpoints
- Easily for variance heterogeneity (Satterthwhaite)
- Available for normal distribution (library(mratios)) and relative effect size (as ratio, but not yet in the mixed effect model) as well

A most extreme pseudo-multiple endpoints approach II

- **Toy example:** Hand and Crowder,Table A.14: blood glucose levels measured at 14 time points over 5 hours for 7 volunteers who took alcohol, where the same was repeated on a second date with the same subjects but with a dietary additive

	Subject	Date	Time	glucose
2	1	1	0	3.0
3	1	1	2	4.7
4	1	1	4	6.0
5	1	1	6	6.3
6	1	1	8	4.3
7	1	1	10	3.0
8	1	1	12	2.0
9	1	1	15	4.5
10	1	1	18	3.8
11	1	1	21	3.2
12	1	1	24	2.6
13	1	1	27	2.6
14	1	1	30	2.6
16	2	1	0	3.6
17	2	1	2	6.0
18	2	1	4	8.6
19	2	1	6	8.8
20	2	1	8	7.2
192	7	2	18	3.7
193	7	2	21	3.4
194	7	2	24	3.6
195	7	2	27	3.6
196	7	2	30	3.6

A most extreme pseudo-multiple endpoints approach III

- Naive mixed effect model (in COD or ... may be much more complex). Sorry, using Dunnett-type approach (here 2 sample t-test) because biosimilar assays compare commonly > 2 formulations

```
Gluc$lc<-log(Gluc$glucose+0.001)
library("lme4")
modM1 <-lmer(lc~group+Time +(Time|Subject), data=Gluc)
library("multcomp")
mix<-exp(confint(glht(modM1, linfct=mcp(group="Dunnett")),level=0.90)$
```

- Comparing the 90% confidence intervals:

Approach	Ratio	lwr	upr
Just repeated measures	1.02	0.96	1.07
AUC	1.03	0.95	1.12
Cmax	1.01	0.84	1.28

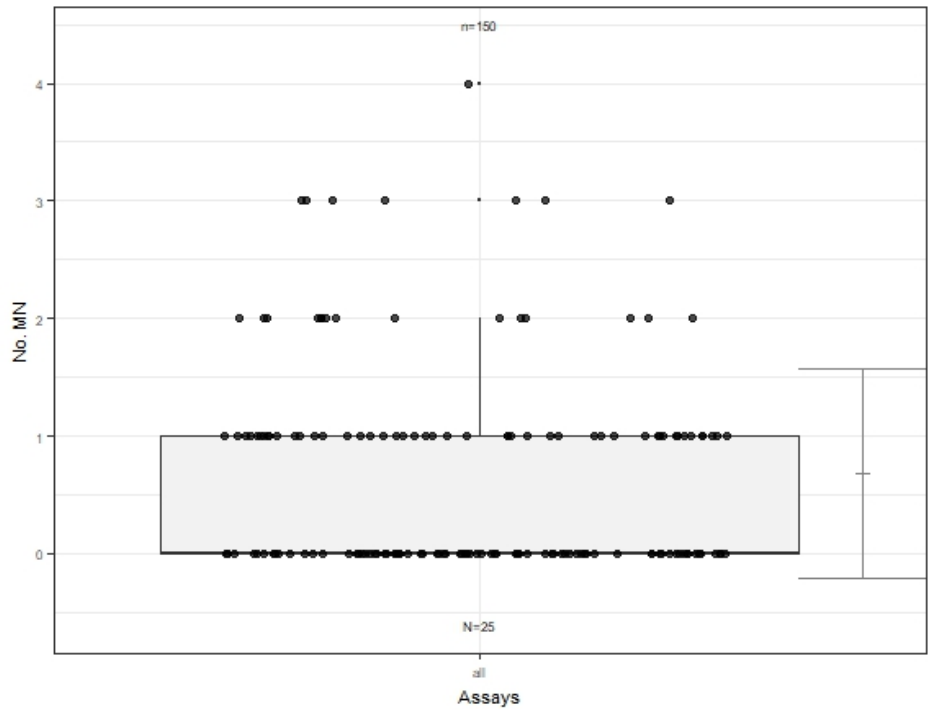
- Regulators will not be amused, I will always do this only for myself personally confidential in the future

Prediction intervals for future obs- a different story for a different problem? I

- In biosimilar testing not only subject-specific measures, such as AUC ar eof interest, also batch-specific quality parameters, e.g. purity
- Here prediction interval for a **single, k of n or all future** observations can be used. The reference products are used as historical data, the biosimilar as new. The rule is: calculate [*lower*; *upper*] from historical data and check whether the individual new data are within the claim similarity
- Commonly, replicates are measure for each batch, ie a random effect model should be used

Prediction intervals for future obs- a different story for a different problem? II

- A faked application: historical micronuclei counts: Runs within 25 historical assays(transformed into pseudo normal). Question: are all counts in the new assay with this prediction interval?



Prediction intervals for future obs- a different story for a different problem? III

- R code

```
library("mixADA")
library("lme4")
fiNC <-lmer(ftt ~ 1 + (1|run), data=HistNC)
fs <-mixADA:::predint_lmer(fit=fiNC, type="h1", level=0.95,
                           alternative="less")$predint
```

- The upper prediction limit in the random effects model is 3.36, ie if no MN in the new assay (any dose!) is < 3.36 this assay can be claimed as safe...
- But we know intervals are extreme sensitive to distribution misclassification
- Any multivariate extension was not found in the literature

Enough confusion: now comes a new approach

A proposal I

Principles:

1. Each primary endpoint can follow its **own specific distribution**, including skewed... censored
2. Primary endpoints can be even differently scaled $\Rightarrow$ find a **comparable transformation**
3. Use the same **comparable effect size: odds ratio** and its confidence interval for all primary endpoints. Another kind of ratio.
4. The **cut-off from continuous to dichotomous** should be 'optimal', specific for each primary endpoint
5. *Leaving the nice ratio* μ_T/μ_R *with its* $(0.8, 1.25)$ *thresholds into OR with* $(?, ?)$ *thresholds*

A proposal II

6. Do not only consider location effects only. Take **location, scale, shape** in a joint approach into account
7. Using simultaneous confidence intervals from an **UIT(IUT)-test** for multiple endpoints and multiple formulation in a complete power approach style
8. Take some **correlations** in the UIT(IUT) into account
9. Provide R-code

A proposal III

- **Issue I):** Using the most appropriate transformation function for each endpoint: most likely transformation approach [14]
- Both in univariate and multivariate analysis it is rather unrealistic to assume the same distribution for each of the many endpoint. The concept of *most likely transformation* (mlt) provides a comparable analysis of such quite different-scaled endpoints.
- The usual regression models estimate the conditional mean as a function of the covariate(s), assuming the higher moments can be ignored. Alternatively, mlt provides semiparametric regression models allows transformation functions to depend on the covariate for kernel-based non-parametric approaches or parametric generalized additive models for location, scale and shape.

A proposal IV

- The estimated conditional distribution functions are consistent which allows a comparable analysis.
- Robust against any non-normal distributions (including discreteness), variance heterogeneity, extreme values, and (left)-censored observations
- By means of the CRAN package `mlt` this is quite simple for a selected endpoint `y` (within the data set `dat` and a grouped covariate `A`):

```
library(mlt)
yvar <- numeric_var("y", support = quantile(dat$y, prob = c(.1, .9)))
yb <- Bernstein_basis(yvar, ui = "increasing", order = 5)
mod <- ctm(yb, shifting = ~ X, todistr = "Normal", data = dat)
fmod <- mlt(mod, data = cc)
```

- The object `fmod` contains the parameter estimates of the most likely regression model for a single slope
- Depending an the scale of the endpoint, both the support region and the degree of Bernstein polynomial (`order=5`) are not too critical for the robustness of this approach.

Using odds ratio as effect size I

- Issue II: Using odds ratio as effect size
  Odds ratio in epidemiology widely used
  - Recently, the continuous outcome logistic regression was proposed [22] with estimating a continuous covariate distribution independent of both the endpoints scale and certain cut-offs for categorization and providing an odds-ratio and its confidence interval for the association between the continues covariate `X` and the arbitrarily distributed endpoint `y`.
  - A tiny change in the above R-code `todistr = "Logistic"` allows the odds-ratio estimation (or more comfortable the CRAN package `library(tram)` can be used)

```
COLRmod <- ctm(yb, shifting = ~ X, todistr = "Logistic", data = dat)
```

- Effect size is a ratio: comparable over multiple, different-scaled endpoints
- OR depends on an optimal cut-point (optimal in different senses: continuous $\Rightarrow$ dichotomous; location,scale,shape

Using odds ratio as effect size II

- The post-hoc categorization of the continuous covariate and its qualitative analysis is widely used, e.g. the use of four dietary reference quartiles [9]- despite all warnings [33, 4].
- It makes it possible to consider the association between a continuous covariate and an arbitrarily distributed endpoint, independently of the endpoints scale and of certain cut-offs for categorization of the covariate.
- It provides effect sizes, in terms of odds ratios with confidence intervals
- The dimensionless odds ratio is comparable over different-scaled analytes (endpoints)
- These odds ratios can be evaluated for all potential values or cut-off of the covariate function, which allows the associations for different categorization types.

Continuous outcome logistic regression (COLR) I

- Why not simply $Y = \alpha + \beta X$?
  1. linear relationship are rare in biosciences (remember Health=f(wine))
  2. $\epsilon = N(0, \sigma^2)$ unlikely, particularly for multiple endpoints
  3. $X_i (i > 10, cor(X_i) > ...)$ (today limited to naive X)
- Interesting in exposure epidemiology: categorization up to now, although many concerns published. Why they do this?
- **Continuous outcome logistic regression (COLR)** for the estimation of a continuous Y distribution. Parameters of interest, such as odds ratios for specific categories (or hazard ratios), can be extracted from this model post-hoc in a general way.
- Core idea of COLR: to model the entire conditional distribution of Y for all reasonable values b simultaneously
  $logit(P(endpoint \leq \phi | covariate)) = r(\phi | covariate)$
  This requires that the parameterization of the regression function is a smooth and monotonically increasing function of cut-off ϕ.

Continuous outcome logistic regression (COLR) II

- The odds ratio can be evaluated for all potential Y values $\phi > 0$, which allows the associations for different categorization schemes to be interpreted post hoc. The regression coefficients β are log-odds ratios of all possible events $Y \leq \phi$
- Advantage of COLR: the possibility of evaluating the likelihood of Y values obtained at different measurement scales or using different categorization schemes
- We can decide whether an association is $+$ or - by the sign of the regression coefficients β. But the nice interpretation: one-unit increase in covariate corresponds to an increase of the conditional mean of Y is *not more possible*
- We can interpret as log-OR, simultaneously for all possible binary logistic regression models
- Strong relationship between quantile regression models and transformation models.

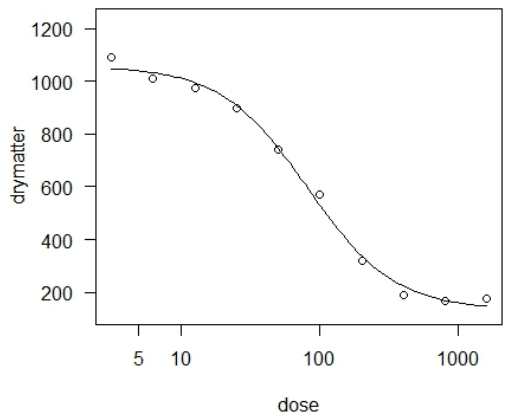
Continuous outcome logistic regression (COLR) III

- Continuous data are also ordinal, and ordinal regression models can be fit to continuous outcomes
- Belong to class of cumulative probability models (CPMs). [21]
- Attractive features:

1. Ordinal regression models are robust because they only incorporate the order information of response variables and are therefore invariant to any monotonic transformation of outcomes. This is particularly useful when the distributions of continuous responses are skewed
2. CPMs directly model the conditional cumulative distribution function, from which other components of the conditional distribution (eg, mean and quantiles) can be easily derived, whereas other regression models often only focus on one aspect mean only
3. CPMs can handle any orderable response, including those with mixed types of continuous and discrete ordinal distributions. eg.detection limits, eg. measurements censored at an assay detection limit resulting in a mixture of an undetectable category and detectable quantities

Continuous outcome logistic regression (COLR) IV

- The CRAN package `library(tram)` or rms, ordinal can be used to estimate such an odds ratio and its confidence interval
- A toy example: Kletten-Labkraut (Galium aparine) were treated with increasing concentration of a herbicide (data in library(drc))




```
library(tram)
myC<-Colr(dry ~ dose, data = mydat)
CI<-exp(confint(myC)); OR<-exp(coef(myC))
```

- The odds ratio is 1.0039 with [1.0028; 1.0050] (notice the huge concentrations)

Coffee break about here

Issue III: A maximum test for multiple correlated endpoints I

- Various multivariate methods exist for a multiple endpoint vector Y_{ijk} with $i = 1, ..., I$ correlated endpoints
- We prefer those based on well-understood univariate test statistics taking the correlation between the endpoints into account e.g. by means of a maximum contrast test [11].
- Whereas the correlation between these $i\xi$ linear models is estimated by the multiple marginal model approach [31]
- This union-intersection test is conservative (with a higher dimension I and less correlation), but offers not only a global statement like the IUT, but also for each individual test/confidence interval
- Again, the R-code is not complicated for two correlated endpoints `y1,y2` and their mlt-transformed models `mod2, mod2`:

```
library(multcomp)
summary(glht(mmm(ari=mod1, ari2=mod2), mlf(ari="X=0", ari2="X=0")))
```

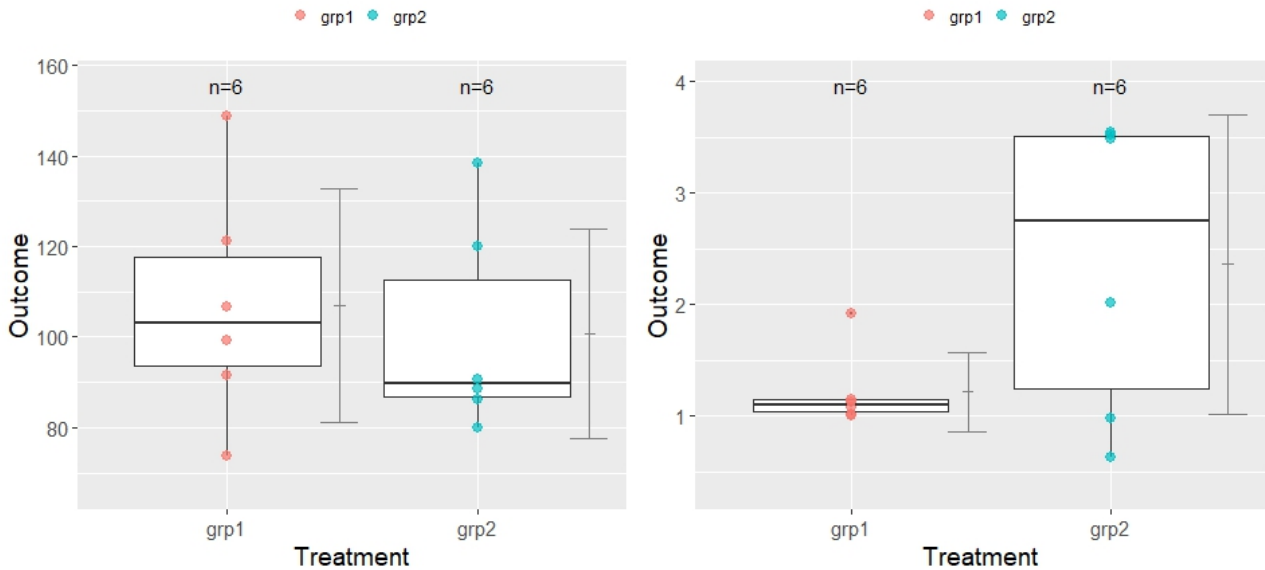
Issue III: A maximum test for multiple correlated endpoints II

- The outcomes of this proposal are for each endpoint an OR and its simultaneous (UIT) confidence interval, allowing a global equivalence claim (all sCI are within *lower*, *upper*) or any patterns of elementary claim. Hereby is the scale comparable, but more complex

	Odds ratio	lwr	upr
AUC	1.88	0.18	19.24
Tmax	0.09	0.00	2.34

- For these faked (tiny n_i) data not surprising

Issue III: A maximum test for multiple correlated endpoints III



Issue IV: Multiple formulations I

- Remember the 1st example

Table 1 Summary of the pharmacokinetic parameters for pegfilgrastim in serum (PK analysis set)

Parameter	MYL-1401H ($N=204$)	EU-reference ($N=203$)	US-reference ($N=207$)	MYL-1401H/EU-reference		MYL-1401H/US-reference	
				LS mean ratio	90% CI	LS mean ratio	90% CI
Primary pharmacokinetic end points							
C_{max} (%CV), pg/mL	36.7 (72.1)	34.2 (72.1)	37.3 (67.6)	1.07	0.98–1.16	0.99	0.91–1.07
AUC_{0-inf} (%CV), h·ng/mL	869 (69.1)	833 (70.1)	876 (66.3)	1.04	0.98–1.11	1.00	0.94–1.07
Secondary pharmacokinetic end points							
AUC_{0-t} (%CV), h·ng/mL	827 (71.4)	787 (72.7)	832 (68.6)	1.05	0.98–1.13	1.00	0.93–1.07
t_{max}, median (range), h	12.0 (6.0–24.0)	12.0 (6.0–48.0)	12.0 (4.0–24.0)	–	–	–	–
k_{el} (%CV), h^{-1}	0.014 (31.0)	0.014 (39.1)	0.014 (40.1)	1.03	0.98–1.08	1.04	0.99–1.09
$t_{1/2}$ (%CV), h	49.3 (36.5)	51.1 (48.9)	51.0 (42.5)	0.97	0.93–1.02	0.97	0.92–1.01
V_d/F (%CV), L	164 (100)	177 (101)	168 (113)	0.93	0.85–1.02	0.99	0.89–1.06

- Design: Biosimilar, EU-Reference, US-Reference quite common
- Claim? All similar?, at least against one reference?....
- I.e. either IUT(all) $\Rightarrow$ the same above shown UIT-complete power approach or UIT(IUT)(at least 1 equivalent)
- I favor the UIT-complete power approach, alone from the perspective of individual claims, e.g. $\mu_B/\mu_{EU} : [0.91, 1.12]$ but $\mu_B/\mu_{US} : [0.97, 1.26]$ but IUT would say: NO

Issue IV: Multiple formulations II

- On the other hand UIT(IUT) for comparison against control available (Bofinger, 1992). Quite complex due to curious correlation matrix
- A more general approach in `library(ETC)` (Hasler 2009) available

```
library(ETC)
 data(BW)
 comp <- etc.rat(formula=Weight~Dose, data=BW, margin.up=1.25, method="var.equ
 summary(comp)

Alternative hypotheses: ratios to control within
the margins 0.8 0.8 0.8 and 1.25 1.25 1.25
Method: var.equal
    estimate statistic  lower upper   p.value
2/1   0.9636    -9.248 0.9226 1.006 1.077e-12
3/1   0.9494    -8.448 0.9087 1.000 2.138e-11
4/1   0.8936    -5.291 0.8540 1.000 3.139e-06
```

- Could be extended to multiple correlated endpoints

Questions so far?

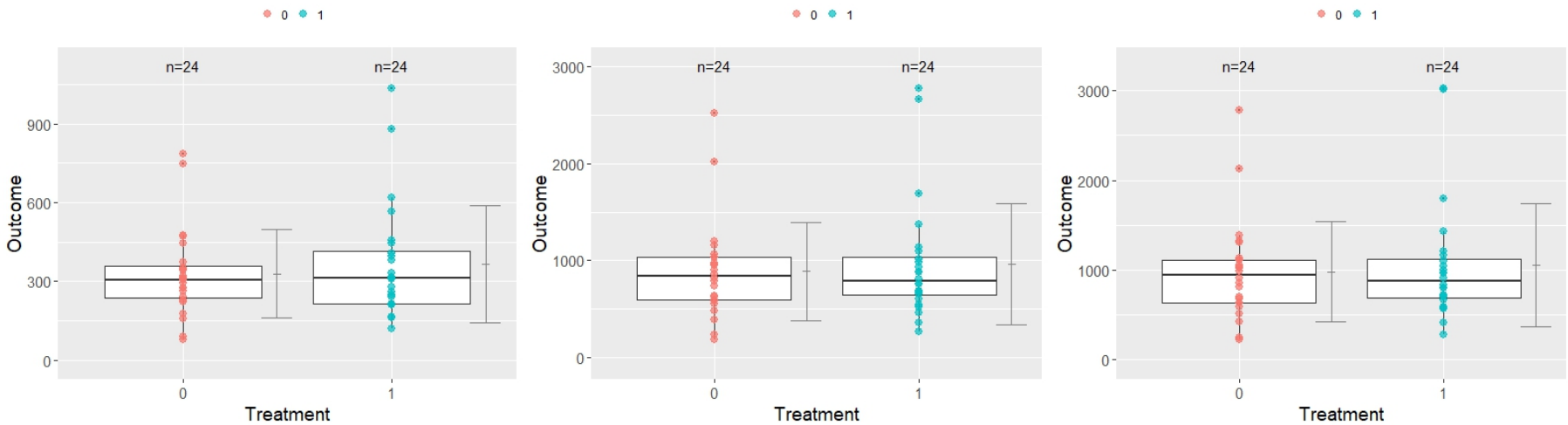
Using the max(max)(mlt) approach for faked ticlopidine data I

- Ticlopidine trial is a COD. Faked data as pseudo parallel group design (Remember: mixed effect model for `mlt` not yet available)
- Raw data quite naive with small sample sizes (see the wide CI)

	Dose	Cmax	Auc	Aucl
1	0	784.3	2021.7	2131.4
2	0	304.2	901.7	1107.9
3	0	307.3	741.4	806.2
4	0	156.7	475.6	509.9
5	0	745.6	2521.4	2784.0
6	0	295.1	1029.3	1391.2
7	0	89.6	232.4	248.1
8	0	321.1	629.6	672.4
9	0	310.8	1035.5	1105.8
10	0	475.4	1193.5	1321.8
...	...	...	...	...
37	1	166.8	260.9	280.2
38	1	455.0	932.0	960.9
39	1	406.5	1141.0	1208.3
40	1	240.9	681.0	715.8
41	1	378.8	812.4	910.6
42	1	278.4	756.7	801.8
43	1	209.6	539.0	577.5
44	1	212.4	606.2	662.9
45	1	162.1	359.2	414.0
46	1	212.1	518.0	572.0
47	1	394.2	1013.4	1101.4
48	1	567.0	1373.8	1433.7

Using the max(max)(mlt) approach for faked ticlopidine data II

- Distribution? Variance homogeneity? (remember log-transformation assumes (among others) homogeneous variances [34]



- Assuming any distribution. Modeling location, scale, shape

Using the max(max)(mlt) approach for faked ticlopidine data III

- mlt per endpoint. Estimation *generalized* log-OR, each

```
library(mlt)
vAUC <- numeric_var("Auc", support = quantile(Mydat$Auc, prob = c(.1, .9)), bo
vCmax <- numeric_var("Cmax", support = quantile(Mydat$Cmax, prob = c(.1, .9)))
vAUCI <- numeric_var("AucI", support = quantile(Mydat$AucI, prob = c(.1, .9)))

### Flexible baseline transformation functions
bAUC <- Bernstein_basis(vAUC, order = 5, ui = "increasing")
bCmax <- Bernstein_basis(vCmax, order = 5, ui = "increasing")
bAUCI <- Bernstein_basis(vAUCI, order = 5, ui = "increasing")

### P(AUC <= y | grp) = expit(h(y) + beta * grp2)
### => beta is y-independ log-OR for grp1
mauc <- ctm(bAUC, shifting = ~ Dose, todistr = "Logistic", data = Mydat)
mcmax <- ctm(bCmax, shifting = ~ Dose, todistr = "Logistic", data = Mydat)
mauci <- ctm(bAUCI, shifting = ~ Dose, todistr = "Logistic", data = Mydat)

auc5 <- mlt(mauc, data=Mydat)
cmax5 <- mlt(mcmax, data=Mydat)
auci5 <- mlt(mauci, data=Mydat)
```

Using the max(max)(mlt) approach for faked ticlopidine data IV

- Three independent simultaneous tests on slopes using mmm. Notice, extension to multiple formulations possible here (Here mmm used for univariate test!)

```
library(multcomp)

BB1 <- glht(mmm(au=auc5),  mlf(au="Dose=0"))
BB2 <- glht(mmm(cma=cmax5),  mlf(cma="Dose=0"))
BB3 <- glht(mmm(aui=auci5),  mlf(aui="Dose=0"))
```

- Max-test on 3 correlated endpoints using mmm to estimate 3 log-OR and its 90% 2-sided simultaneous confidence limits

```
BBB <- glht(mmm(au=auc5, cma=cmax5, aui=auci5),
            mlf(au="Dose=0", cma="Dose=0", aui="Dose=0"))
bbb<-exp(confint(BBB)$confint)
```

- We could estimate adjusted p-values

Using the max(max)(mlt) approach for faked ticlopidine data V

```
 Simultaneous Tests for General Linear Hypotheses
Linear Hypotheses:
                 Estimate Std. Error z value Pr(>|z|)
au: Dose == 0  -0.048224   0.502967  -0.096    0.992
cma: Dose == 0 -0.113613   0.502943  -0.226    0.941
aui: Dose == 0 -0.002994   0.502980  -0.006    1.000
(Adjusted p values reported -- single-step method)
```

- We should estimate OR and their 90% 2-sided simultaneous confidence limits

```
           Estimate       lwr      upr
au: Dose  0.9529200 0.3825822 2.373494
cma: Dose 0.8926030 0.3583815 2.223162
aui: Dose 0.9970109 0.4002751 2.483369
attr(,"conf.level")
[1] 0.9
```

- OR's near to 1, but sCI wide. I.e. we have to thing about the thresholds for this specific estimate

Needed I

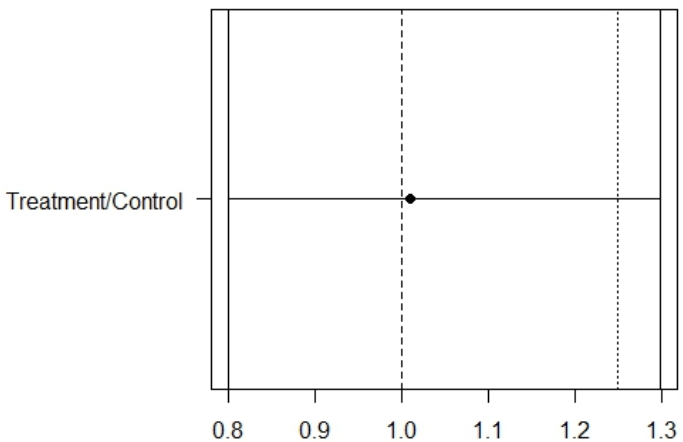
- Remember: claiming equivalence depends both on estimates: new effect size **and** choice of $\delta \Rightarrow$ new threshold needed (eg. by method comparison simulation)
- Extending to complex designs (eg. mixed effect model; not yet available)
-

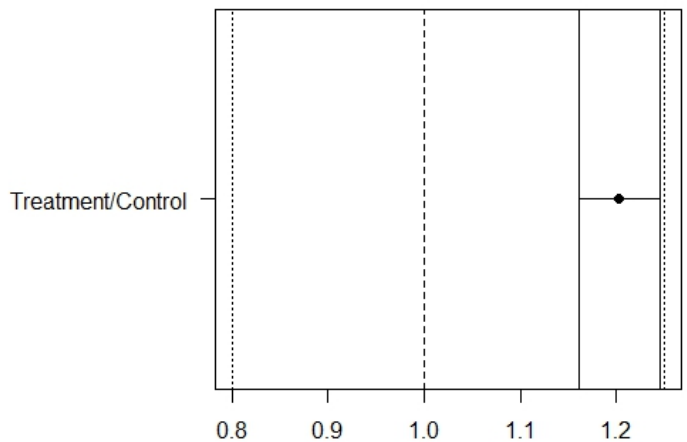
Provocation about equivalence criteria I

- For me is the primary criterion the closeness of odds ratio (or ratio of ...) to the **value of 1**. However, this criterion is missing completely in the interval inclusion approach of TOST. And an $OR = 0.997$ tells me a high degree of similarity....
- Choice of δ for TOST is rather complicated and this permanent relapse to FDA $0.8, 1.25]$ makes me unhappy.
- δ is NOT a simple function of variance (precision), it depends ALSO on the medical/biological meaning and it should be symmetric ONLY in rather rare cases (because to have more or less rate and extend of a biosimilar relative to the comparator has NOT the same consequences
- Why the p-values for $point - 0 H_0$ 0.992, 0.941 and 0.999 should be completely ignored? (Yes, I know the difference between PoH and PoS)

Provocation about equivalence criteria II

- Multiple endpoints intensify this choice of δ problem. Looking on the genetic modified varieties with > 100 endpoints: we will never be able to defined fair δ_{ij}
- My second counterargument against TOST interval approach





Take home I

- With simultaneous tests for multiple endpoints in biosimilars, we are not too late compared to RTC.
- There is enough time for controversies about different approaches before a guidance is prepared.
- This talk was a contribution to the controversy, I propose a, not the golden way, to Rome, via Aquincum
- Next, submitting a related paper
- Choice of δ in this odds ratio scale for multiple endpoints with difference importance and scales remains as an issue

Take home II

- Data, δ, n_i (Power), approach, objective (IUT, UIT) determines biosimilar analysis with correlated (PK) endpoints
- Counter argument: Don't act like that and take log-normal homoscedastic for granted: use multi var test. Contact Thomas and Phillip and use `library(jorce)`
- Finaly, sorry for too much questions and contradictions. But a proposal exists- and you can re-analysed easily your data with these CRAN packages and make up your own mind

References I

[1] *pairwiseCI: Confidence Intervals for Two Sample Comparisons, author = Frank Schaarschmidt and Daniel Gerhard, year = 2018, note = R package version 0.1-26, url = https://CRAN.R-project.org/package=pairwiseCI,.*

[2] R. Arboretti, E. Carrozzo, F. Pesarin, and L. Salmaso. Testing for equivalence: An intersection-union permutation solution. *Statistics in Biopharmaceutical Research*, 10(2):130–138, 2018.

[3] H. Y. Barnett, H. Geys, T. Jacobs, and T. Jaki. Comparing sampling methods for pharmacokinetic studies using model averaged derived parameters. *Statistics in Medicine*, 36(27):4301–4315, November 2017.

[4] J. L. Barnwell-Menard, Q. Li, and A. A. Cohen. Effects of categorization method, regression type, and variable distribution on the inflation of type-i error rate when categorizing a confounding variable. *Statistics in Medicine*, 34(6):936–949, March 2015.

[5] R. L. Berger and J. C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–302, November 1996.

[6] L. D. Brown, J. T. G. Hwang, and A. Munk. An unbiased test for the bioequivalence problem. *Annals of Statistics*, 25(6):2345–2367, December 1997.

[7] Y. Cao, D. Obeng, G. D. Hui, L. T. Xue, Y. K. Ren, X. J. Yu, F. Wang, and C. Atwell. Evaluating manufacturing process profile comparability with multivariate equivalence testing: Case study of cell-culture small scale model transfer. *Biotechnology Progress*, 34(1):187–195, January 2018.

[8] L. P. Du and L. Choi. Likelihood approach for evaluating bioequivalence of highly variable drugs. *Pharmaceutical Statistics*, 14(2):82–94, March 2015.

[9] R. Eriksen, R. Gibson, K. Lamb, Y. McMeel, A. C. Vergnaud, J. Spear, M. Aresu, Q. Chan, P. Elliott, and G. Frost. Nutrient profiling and adherence to components of the uk national dietary guidelines association with metabolic risk factors for cvd and diabetes: Airwave health monitoring study. *British Journal of Nutrition*, 119(6):695–705, March 2018.

[10] C. B. Fogarty and D. S. Small. Equivalence testing for functional data with an application to comparing pulmonary function devices. *Annals of Applied Statistics*, 8(4):2002–2026, December 2014.

[11] M. Hasler and L. A. Hothorn. Simultaneous confidence intervals on multivariate non-inferiority. *Statistics in Medicine*, 32(10):1720–1729, May 2013.

[12] D. Hauschke, M. Kieser, and L. A. Hothorn. Proof of safety in toxicology based on the ratio of two means for normally distributed data. *Biometrical Journal*, 41(3):295–304, 1999.

References II

[13] L. A. Hothorn and R. Oberdoerfer. Statistical analysis used in the nutritional assessment of novel food using the proof of safety. *Regulatory Toxicology and Pharmacology*, 44(2):125–135, March 2006.

[14] T. Hothorn, L. Most, and P. Buhlmann. Most likely transformations. *Scandinavian Journal of Statistics*, 45(1):110–134, March 2018.

[15] S. Y. Hua, D. L. Hawkins, and J. H. Zhou. Statistical considerations in bioequivalence of two area under the concentrationtime curves obtained from serial sampling data. *Journal of Applied Statistics*, 40(5):1140–1154, May 2013.

[16] S. Y. Hua, S. Y. Xu, and R. B. D'Agostino. Multiplicity adjustments in testing for bioequivalence. *Statistics in Medicine*, 34(2):215–231, January 2015.

[17] L. Kong, R. C. Kohberger, and G. G. Koch. Design of vaccine equivalence/non-inferiority trials with correlated multiple binomial endpoints. *Journal of Biopharmaceutical Statistics*, 16(4):555–572, July 2006.

[18] B. Lang and F. Fleischer. Comments on 'multiplicity adjustments in testing for bioequivalence'. *Statistics in Medicine*, 35(14):2479–2480, June 2016.

[19] K. K. Lin and M. A. Rahman. Comparisons of false negative rates from a trend test alone and from a trend test jointly with a control-high groups pairwise test in the determination of the carcinogenicity of new drugs. *J. Biopharm Statist*, 2018.

[20] F. Liu. Some correlations in intersection-union tests and their relationship with complete power. *Statistics & Probability Letters*, 81(4):518–523, April 2011.

[21] Q. Liu, B. E. Shepherd, C. Li, and F. E. Harrell. Modeling continuous response variables using ordinal regression. *Statistics in Medicine*, 36(27):4316–4335, November 2017.

[22] T Lohse, S Rohrmann, D Faeh, and T Hothorn. Continuous outcome logistic regression for analyzing body mass index distributions. *F1000Research*, 2017, 6:1933 (doi: 10.12688/f1000research.12934.1).

[23] W. Maurer, B. Jones, and Y. Chen. Controlling the type i error rate in two-stage sequential adaptive designs when testing for average bioequivalence. *Statistics in Medicine*, 37(10):1587–1607, May 2018.

[24] J. Mielke, B. Jones, B. Jilma, and F. Konig. Sample size for multiple hypothesis testing in biosimilar development. *Statistics in Biopharmaceutical Research*, 10(1):39–49, 2018.

[25] A. Munk and R. Pfluger. 1-alpha equivariant confidence rules for convex alternatives are alpha/2-level tests - with applications to the multivariate assessment of bioequivalence. *Journal of the American Statistical Association*, 94(448):1311–1319, December 1999.

References III

[26] P. Pallmann and T. Jaki. Simultaneous confidence regions for multivariate bioequivalence. *Statistics in Medicine*, 36(29):4585–4603, December 2017.

[27] P. Pallmann, M. Pretorius, and C. Ritz. Simultaneous comparisons of treatments at multiple time points: combined marginal models versus joint modeling. *Statistical Methods in Medical Research, accepted for publication.*, 2015.

[28] Won Park, Pawel Hrycaj, Slawomir Jeka, Volodymyr Kovalenko, Grygorii Lysenko, Pedro Miranda, Helena Mikazane, Sergio Gutierrez-Urena, MieJin Lim, Yeon-Ah Lee, Sang Joon Lee, HoUng Kim, Dae Hyun Yoo, and Juergen Braun. A randomised, double-blind, multicentre, parallel-group, prospective study comparing the pharmacokinetics, safety, and efficacy of ct-p13 and innovator infliximab in patients with ankylosing spondylitis: the planetas study. *Annals of the Rheumatic Diseases*, 72(10):1605–1612, October 2013.

[29] A. Phillips, C. Fletcher, G. Atkinson, E. Channon, A. Douiri, T. Jaki, J. Maca, D. Morgan, J. H. Roger, and P. Terrill. Multiplicity: discussion points from the statisticians in the pharmaceutical industry multiplicity expert group. *Pharmaceutical Statistics*, 12(5):255–259, September 2013.

[30] K. F. Phillips. Power for testing multiple instances of the two one-sided tests procedure. *International Journal of Biostatistics*, 5(1):15, 2009.

[31] Christian Bressen Pipper, Christian Ritz, and Hans Bisgaard. A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. *Journal of the Royal Statistical Society Series C-applied Statistics*, 61:315–326, 2012.

[32] H. Quan, J. Bolognese, and W. Y. Yuan. Assessment of equivalence on multiple endpoints. *Statistics in Medicine*, 20(21):3159–3173, November 2001.

[33] P. Royston, D. G. Altman, and W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25(1):127–141, January 2006.

[34] F. Schaarschmidt. Simultaneous confidence intervals for multiple comparisons among expected values of log-normal variables. *Computational Statistics and Data Analysis*, 58:265–275, FEB 2013.

[35] S. Y. Tian, H. H. Chang, D. Orange, J. K. Gu, and M. Suarez-Farinas. A bioequivalence test by the direct comparison of concentration-versus-time curves using local polynomial smoothers. *Computational and Mathematical Methods in Medicine*, page 4680642, 2016.

[36] R. Uozumi and C. Hamada. Adaptive seamless design for establishing pharmacokinetic and efficacy equivalence in developing biosimilars. *Therapeutic Innovation & Regulatory Science*, 51(6):761–769, November 2017.

References IV

[37] C. I. Vahl and Q. Kang. Equivalence criteria for the safety evaluation of a genetically modified crop: a statistical perspective. *Journal of Agricultural Science*, 154(3):383–406, April 2016.

[38] H. van der Voet. Safety assessments and multiplicity adjustment: Comments on a recent paper. *Journal of Agricultural and Food Chemistry*, 66(9):2194–2195, March 2018.

[39] H. van der Voet, P. W. Goedhart, and K. Schmidt. Equivalence testing using existing reference data: An example with genetically modified and conventional crops in animal feeding studies. *Food and Chemical Toxicology*, 109:472–485, November 2017.

[40] C. F. Waller, R. G. Tiessen, T. E. Lawrence, A. Shaw, M. S. Liu, R. Sharma, M. Baczkowski, M. A. Kothekar, C. E. Micales, A. Barve, G. M. Ranganna, and E. J. Pennella. A pharmacokinetics and pharmacodynamics equivalence trial of the proposed pegfilgrastim biosimilar, myl-1401h, versus reference pegfilgrastim. *Journal of Cancer Research and Clinical Oncology*, 144(6):1087–1095, June 2018.

[41] W. Z. Wang, J. T. G. Hwang, and A. Dasgupta. Statistical tests for multivariate bioequivalence. *Biometrika*, 86(2):395–402, June 1999.

[42] R. E. Weiss, X. M. Xia, N. Zhang, H. Wang, and E. Chi. Bayesian methods for analysis of biosimilar phase iii trials. *Statistics in Medicine*, 37(20):2938–2953, September 2018.

[43] C. Yang, A. A. Bartolucci, and X. Q. Cui. Multigroup equivalence analysis for high-dimensional expression data. *Cancer Informatics*, 14:253–263, 2015.