Ein universeller Trendtest. Welche Kriterien sollte ein Trendtest erfüllen?

Ludwig A. Hothorn
Leibniz University Hannover, Germany

Jan 9, 2017

Kriterien I

- Angemessene Güte gegenüber allen Formen einer einseitigen Trendalternative- nicht nur gegenüber der lineare Dosis-Wirkungs-Funktion wie z.B. der Cochran-Armitage (1955) Trend Test
- Verallgemeinerbar im glmm
- Zulässiges Niveau und akzeptable Güte, auch bei kleineren Fallzahlen
- Einleuchtende Interpretierbarkeit für Nichtstatistiker, z.B. plausible Effektgröße
- Numerische Verfügbarkeit: library()
- Verfügbarkeit für korrelierte multiple Endpunkte, vorzugsweise in verschiedenen Skalen
- Dosis als qualitativer Faktor oder/und als quantitative Kovariable formulierbar
- Robust gegenüber Umkehreffekten bei hohen Dosen

Kriterien II

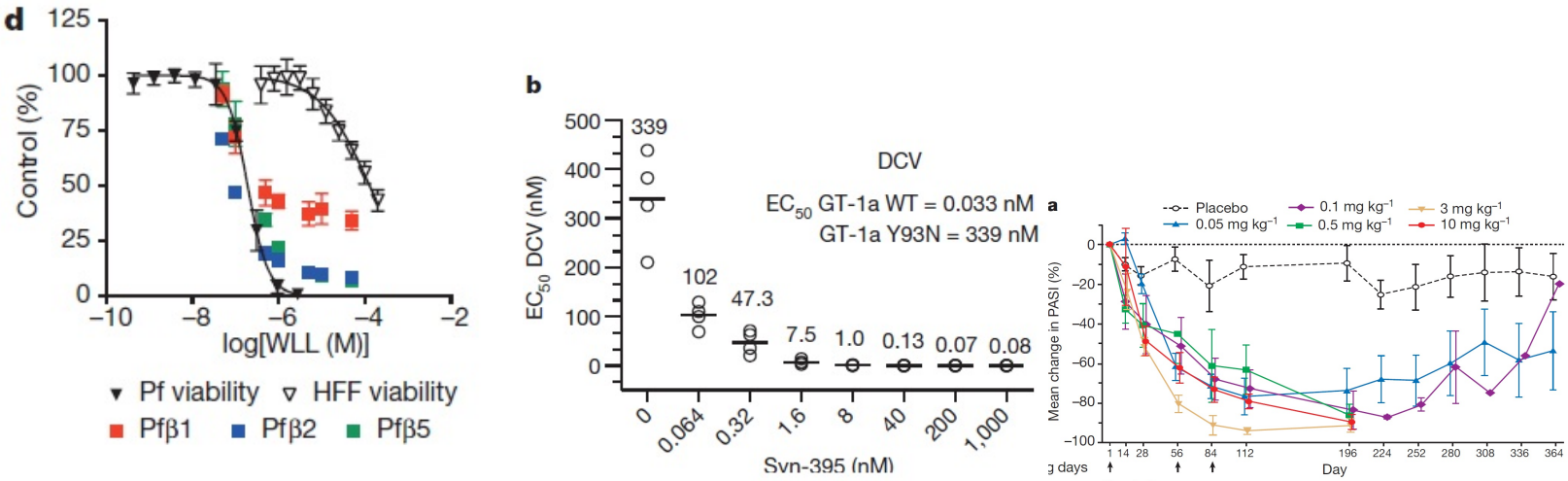
- Nichtparametrische Variante verfügbar
- Geschlossene Powerberechenbarkeit
- ...

- Focussing:
- How dose is considered in selected real data examples today?
- Multiple contrast tests vs. quasi-linear regression models
- Tukey trend test
- Didactical concept: explaining R code

Motivating examples I

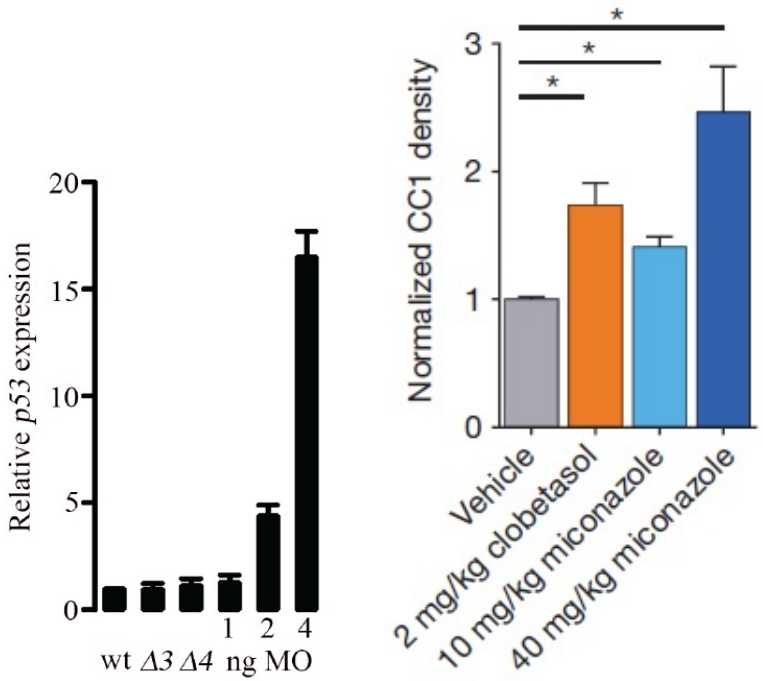
Five randomly selected 2016 Nature papers ([16]):

- left) strong log; many concentrations; no control; comparing 2
  mid) log; zero dose control; displayed qualitatively
  right) placebo-controlled RCT; few doses; log (nonconstant); but presented qualitatively; dose-by-time interaction



Motivating examples II

- left) both qualitative factor and quantitative dose levels
  right) reference drug and 2 quantitative dose levels mixed



- Qualitative factor **and/or** quantitative covariable
- Assuming log transformation (to some extend)
- **Almost regression model I (x_{ij} non-random), i.e. grouped dose levels**
- Why log-transformation? How log(C=0)?

Classification I

- I) Dose as **qualitative factor** $\Rightarrow$ multiple contrast test
- II) Dose as **quantitative covariable** $\Rightarrow$ lin-log or nonlinear regression model
- Choice (I,II) depends on number doses, aim, dose vs. concentration at target cells, ... No a-priori universal best approach
- Main aim of this talk: both MCT and lin-log model
- MCPMod approach: ...
- Tukey's idea:
  three dose metameters **arithmetic, ordinal, lin-log**
  covers most shapes of dose response, is simple and easy to interpret [23]
- Main results: i) MCT **and** Tukey trend test, ii) extensions within glmm, iii) R library `tukeytrend`

MCT: Motivation I

- **Dose-response studies** manifold aims:
  - Just any heterogeneity between the groups
  - Comparison vs. control (0) for several effect sizes $\mu_i - \mu_0$, μ_i/μ_0, π_i/π_0, ...
  - A dose-related trend (global only)
  - Estimate a particular dose: MED (efficacy), LOAEL (MAXSD) (safety)
  - ...

- Why order restriction?
  - Ordered alternative: $H_1 : \mu_0 \leq \mu_1 \leq ... \leq \mu_k | \mu_0 < \mu_k$,
  - Increase the power and/or
  - Achieve a specific claim, such as increasing monotone trend, or identification MED/LOAEL.
  - Notice, trade-off between restriction and robustness

Order restricted tests and related simultaneous confidence intervals I

- What means trend?
- Decomposition the monotone $H_1 : \mu_0 \leq \mu_1 \leq ... \leq \mu_k$ into all linear-elementary alternatives; e.g. k=3
  $H_1^a : \mu_0 = \mu_1 = \mu_2 < \mu_3$
  $H_1^b : \mu_0 = \mu_1 < \mu_2 < \mu_3$
  $H_1^c : \mu_0 < \mu_1 = \mu_2 < \mu_3$
  $H_1^d : \mu_0 < \mu_1 = \mu_2 = \mu_3$
  $H_1^e : \mu_0 < \mu_1 < \mu_2 = \mu_3$
  $H_1^f : \mu_0 < \mu_1 = \mu_2 < \mu_3$
  $H_1^g : \mu_0 = \mu_1 < \mu_2 = \mu_3$
- Called isotonic H_1
- *Different H_1 definitions available*

Order restricted tests and related simultaneous confidence intervals II

- A trend test should be **sensitive against all possible elementary alternatives**.
  Not against just one, e.g. the linear trend- as the wide-spread used Cochran-Armitage trend test [3] for proportions or the Jonckheere trend test for pairwise ranks
  $\Rightarrow$ crazy
- At least two approaches:
  i. MLE-test **quadratic test statistics** [4]
  ii. MCT **linear test statistics**
- **A specific trend test, which compares vs. control:**
  Williams trend test [25] **typically in nonclin and clin dose finding studies**

Order restricted tests and related simultaneous confidence intervals III

- **MCT I)**: A contrast is a suitable linear combination of means (or other effect sizes, e.g.π_i): $\sum_{i=0}^{k} c_i \bar{x}_i$
- Here $i = 0...k$, focusing on comparisons vs. control (placebo) (more general possible)
- **MCT II):** A contrast test is standardized

$$t_{Contrast} = \sum_{i=0}^{k} c_i \bar{x}_i / S \sqrt{\sum_{i}^{k} c_i^2 / n_i}$$

where $\sum_{i=0}^{k} c_i = 0$ guaranteed a $t_{df,1-\alpha}$ distributed level-α-test.

- To guarantee comparable simultaneous confidence intervals needed: $\sum sign^{+}(c_i) = 1, \sum sign^{-}(c_i) = 1$

Order restricted tests and related simultaneous confidence intervals IV

- **MCT III):** A **multiple contrast test is defined as maximum test**:

$$t_{MCT} = max(t_1, ..., t_q)$$

which follows jointly $(t_1, \ldots, t_q)'$ a q-variate t- distribution with degree of freedom df and the correlation matrix $R \Rightarrow$ depending on c_i, n_i but also $s_i, \rho_i, ...$
May be complex

- **MCT IV):** Just the choice of a particular contrast matrix defines the respective MCT

Order restricted tests and related simultaneous confidence intervals V

Known examples (balanced design k=2)

- Dunnett one-sided [9]

c_i	C	T_1	T_2
c_a	-1	0	1
c_b	-1	-1	0

- Williams Procedure (as multiple contrast [6])

c_i	C	D_1	D_2
c_a	-1	0	1
c_b	-1	1/2	1/2

- Much more.... (interesting GrandMean [18])
- **MCT V):** One-sided (lower) simultaneous confidence limits:
  $[\sum_{i=0}^{k} c_i \bar{x}_i - S * t_{q,df,R,2-sided,1-\alpha} \sqrt{\sum_i^k c_i^2/n_i}]$

Modification: Effect sizes I

- **Different effect size**: sCI for $\omega_i = \mu_i/\mu_0$
- Sasabuchi's trick of a linear form $L(\omega_i) = \sum c_i \overline{Y}_i - d_i\omega_i \overline{Y}_0$ (nominator c_i, denominator d_i)
- Simultaneous Fieller-type confidence intervals for ω_i - solutions of the inequalities

$$T^2(\omega_i) = \frac{L^2(\omega_i)}{S^2_{L(\omega_i)}} \leq t^2_{q,\nu,R(),1-\alpha},$$

- $t_{q,\nu,R(i),1-\alpha}$ is a central q-variate t-distribution with ν degrees of freedom and correlation matrix $R(i) = [\rho_{ij}]$, where $\rho_{ij'}$ depend on c_{hi}, n_i **and on unknown ratios** ω_i: plug-in ML-estimators [8] $\Rightarrow$ second trick

Modification: Effect sizes II

- **Relative effect size**: $H_0^F : F_0 = ... = F_k$ formulated in terms of the distribution functions against simple tree $H_1^F : F_0 < F_i$
- But the distribution of the rank means is unknown under H_1, neither sCI nor power can be estimated
- Using relative effect size [7], [22]:

$$p_{01} = \int F_0 dF_1 = P(X_{01} < X_{11}) + 0.5P(X_{01} = X_{11}).$$

- p_{01} is a *win probability* in the sense of [10]
- **sCI:** [15] Let $R_{lj}^{(0l)}$ denote the rank of X_{lj} among all $n_0 + n_l$ observations within the samples 0 and l

Tukey's trend test- Intro I

- Up to now: dose as a qualitative factor only
- Now: considering dose as **quantitative covariate**

Tukey's trend test I

- Three decades ago: [23] max-test on three regression models for the arithmetic, ordinal, and linear-log dose metameters of the **covariate dose** in a randomized one-way layout- without multiplicity adjustment
- Joint distribution? Problem: $\mathbb{R}$
- Tukey's trend test based on ξ multiple linear regression models for the ξ dose transformation functions $\psi^{\xi}(D_j)$ for a vector of response variables y_{ijk} with $i = 1, ..., I$ multiple endpoints in $j = 0, ..., J$ dose levels with k_j unbalanced replicates

$$y_{ijk}^{\xi} = \alpha_{i\xi} + \beta_{i\xi}(\psi^{\xi}(D_{jk})) + \epsilon_{\xi ijk}$$

Tukey's trend test II

- A maximum test on the slope parameters $\beta_{i\xi}$ from multiple marginal models for a global null hypothesis is performed

$$H_0 : \beta_{i\xi}(\psi^{\xi}(D_j)) = 0$$

  representing a union-intersection test.
- In the mmm-framework [20] ξ marginal models for a **univariate** endpoint ($i = 1$) or ($\xi * I$) marginal models for I **multiple endpoints** are included.
- From these parameter estimates the correlation matrix is estimated and the test is on the ξ (respective ($\xi * I$)) slope parameters $\beta_{i\xi}$.
- Joint distribution of parameter estimates from **multiple marginal models** [20]- without assuming a certain multivariate distribution for the data

Tukey's trend test III

- Allows max-tests on multiple linear models and estimation of adjusted p-values or simultaneous confidence intervals
  without the explicit formulation of the correlation matrix in `lm`, `glm`, `lmm`- asymptotically

- For appropriate chosen df ν, finite versions works well (various simulations)

- Example: Bivariate normal endpoints: trend test for liver and body weight (to avoid relative organ weights)

Tukey's trend test IV

Dose	BodyWt	LiverWt
0	338	11
0	319	10
0	369	13
0	373	13
0	315	10
...	...	...
...	...	
1000	294	9
1000	294	9
1000	281	8
1000	317	9
1000	292	8

- mmm formulated for slope-to-zero test for three dose scalings **and** 2 endpoints: six highly correlated tests on slope parameters- *by means of R-code*

Tukey's trend test V

- elementary code

```
bN <-lm(LiverWt~DoseN, data=liv)     # arithm
bO <-lm(LiverWt~DoseO, data=liv)     # ordinal
bLL <-lm(LiverWt~DoseLL, data=liv)   # log-lin

lN <-lm(BodyWt~DoseN, data=liv)
lO <-lm(BodyWt~DoseO, data=liv)
lLL <-lm(BodyWt~DoseLL, data=liv)
library("multcomp")
BoLi <- glht(mmm(covarLiv=lN, ordinLiv=lO, linlogLiv=lLL,
                 covarBody=bN, ordinBody=bO,  linlogBody=bLL),
             mlf(covarLiv="DoseN=0", ordinLiv="DoseO=0", linlogLiv="DoseLL=0",
                 covarBody="DoseN=0", ordinBody="DoseO=0", linlogBody="DoseLL=0"))
```

- function

```
data("liv", package="SiTuR")
fitLl <-lm(LiverWt~Dose, data=liv)
fitLb <-lm(BodyWt~Dose, data=liv)
ttLl <- tukeytrendfit(fitLl, dose="Dose", scaling=c("ari", "ord", "arilog"))
ttLb <- tukeytrendfit(fitLb, dose="Dose", scaling=c("ari", "ord", "arilog"))
cttL <- combtt(ttLl, ttLb)
EXA11<-summary(asglht(cttL))
```

- package `tukeytrend`

Tukey's trend test VI

- Interpret the adjusted p-values!

Model	Test stats	p-value
covarLiv: DoseN	-5.38	0.0000002
ordinLiv: DoseO	-3.98	0.0001642
linlogLiv: DoseLL	-3.98	0.0001611
covarBody: DoseN	-6.04	0.0000000
ordinBody: DoseO	-5.17	0.0000003
linlogBody: DoseLL	-5.17	0.0000003

Table : Tukey trend test for bivariate normal: body and liver weights

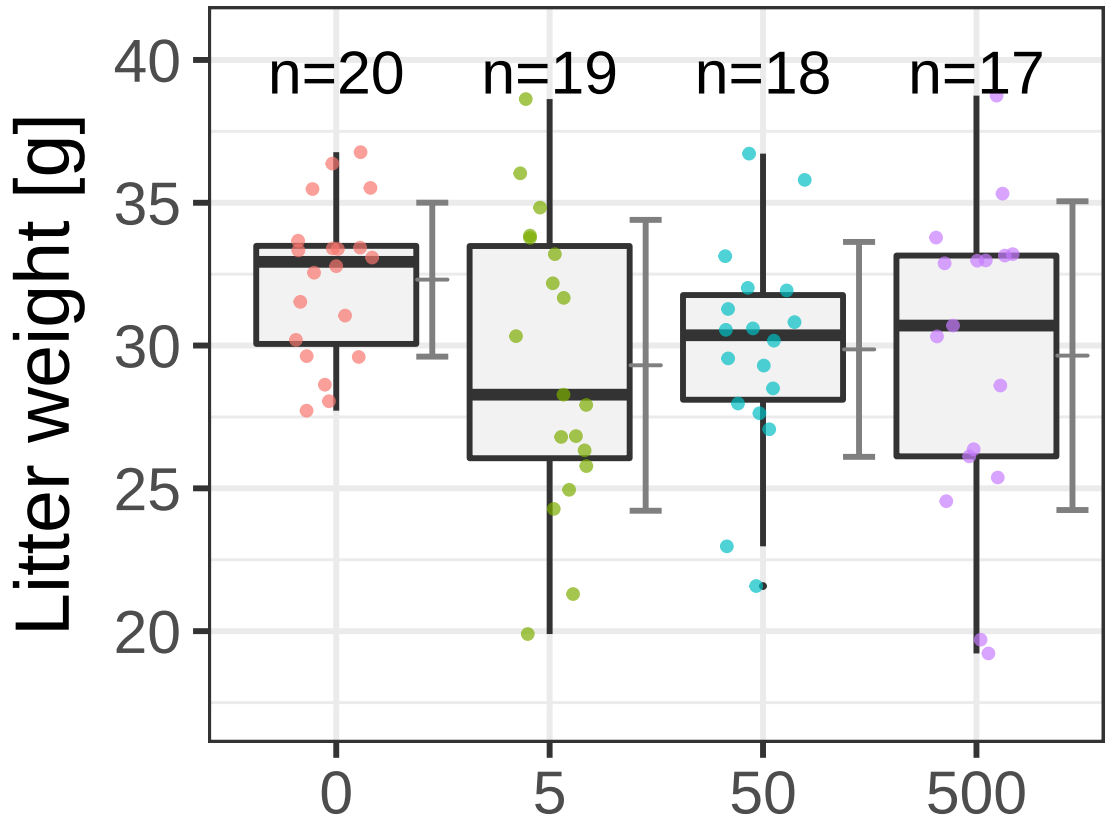
- Trends for both endpoints
- Alternatively, simultaneous confidence intervals for the slopes available- more appropriate for interpretation (not shown)

Trend test using both a covariate and a factor I

- To assume dose as a **qualitative factor** or a **quantitative covariate** result in quite different- disjoint- approaches: trend tests or non-linear models
- Common perception: trend test and (non)linear models are completely separate approaches -*not necessarily!*
- Extension of the Tukey trend test:
  i) three regression models for the arithmetic, ordinal, and logarithmic-linear dose metameters [23] **AND ii) Williams multiple contrast**
- Background: mmm based on linear models. Tukey assumes linear models after transformation, Williams MCT assumes a factor, both belong to the lm-class. Exciting: estimation the correlation implicitely. To calculate the correlation may be rather complex!

Trend test using both a covariate and a factor II

- Example: litter weight data [12]



- Directed: decreasing weights. No clear trend. A possible dose plateau

Trend test using both a covariate and a factor III

- 4 marginal models for 6 hypotheses needed:
  3 regression models for arithmetic, ordinal and log-linear dose metameters **and** 3 Williams-type multiple contrasts

```
litter$dosen <- as.numeric(as.character(litter$dose)) # add a nume
fitc <- lm(weight ~ dosen, data=litter)
dfn<-fitc$df.residual
ttw <- tukeytrendfit(fitc, dose="dosen",
 scaling=c("ari", "ord", "arilog", "treat"),ctype="Williams")
exa1<-summary(glht(ttw$mmm, ttw$mlf), df=dfn)
```

Dose metameter	Test statistics	*p*-value
dosenari: dosenari	-0.818	0.726
dosenord: dosenord	-1.703	0.214
dosenarilog: dosenarilog	-1.128	0.519
dosentreat: C 1	-1.863	0.156
dosentreat: C 2	-2.287	0.061
dosentreat: C 3	-2.759	0.018

- Look how insensitive any regression model for a plateau shape is!

Trend test using both a covariate and a factor IV

- More general:
  1. Power of Tukey trend test depends on dose metameters and design (unbalancedness, k) and ...
  2. A tiny simulation study (log-scaled doses, 10000 runs) (rather new rank versions in a recent manuscript)

shape	Williams	Tukey	TukeyWil
dose	quali	quanti	both
linear	0.85	0.89	0.87
supralin	0.95	**0.76**	0.87
sublin	0.81	0.96	0.89

  3. Serious power loss for plateau profiles for dose as quantitative covariate
  4. TukeyWilliams max-test: no serious power loss for any shape of dose response. Robustification!
  5. TukeyWilliams max-test: interpreting covariate vs. factor (or pairwise comparison $Cvs.D_{max}$)

Formulated for non-monotonic trends I

- A downturn effect at high doses is likely in toxicology, i.e. a non-monotonic dose-response relationship occurs [12].
- Trend tests may be seriously biased and tests without any order restriction, such as Dunnett test, allows only groupwise inference, no claiming trend
- A double maximum test can be used, a maximum over all possible peak point doses (from $k, (k-1), (k-2), ..., 1$) and a maximum on the multiple contrasts (e.g. Williams-type contrasts) [5].
- This idea can be easily transferred to the Tukey trend test approach by
  A) defining an so-called `UmbrellaWilliams` contrast, i.e. taking dose qualitatively or,

Formulated for non-monotonic trends II

B) using a double maximum test on the possible peak points and the three regression models, where these pseudo dose metameters up to certain dose are generated with `NA` (see below object `comptt`).

- Although $3k$ instead of 3 models are subject to a maximum test, the power will be not sacrificed because these many models are highly correlated.
- Dose qualitatively

```
tt10 <- tukeytrendfit(fitw, dose="dosen",
          scaling=c("ari", "ord", "arilog", "treat"),
          ctype="UmbrellaWilliams", ddf="residual")
Exa10<-summary(glht(model=tt10$mmm, linfct=tt10$mlf, alter
```

Formulated for non-monotonic trends III

- Dose quantitatively

```
dl$dos500<-dl$dosen; dl$dos50<-dl$dosen
dl$dos500[dl$dosen==500] <-NA
dl$dos50[dl$dos50==50] <-NA
fitall<-lm(weight ~ dosen, data=dl)
fit500 <- lm(weight ~ dos500, data=dl)
fit50 <- lm(weight ~ dos50, data=dl)
tt50<- tukeytrendfit(fit50, dose="dos50",
            scaling=c("ari"), ddf="residual")
tt500<- tukeytrendfit(fit500, dose="dos500",
            scaling=c("ari", "ord", "arilog"), ddf="residual")
ttall<- tukeytrendfit(fitall, dose="dosen",
            scaling=c("ari", "ord", "arilog"), ddf="residual")
combi10 <- combtt(tt50,tt500,ttall)
comptt<- summary(glht(model=combi10$mmm, linfct=combi10$mlf, alte
```

Dose metameter	Test statistics	*p*-value
tt50.lm.weight.dos50ari: dos50ari	-0.910	0.388
tt500.lm.weight.dos500ari: dos500ari	-1.047	0.327
tt500.lm.weight.dos500ord: dos500ord	-1.922	0.076
tt500.lm.weight.dos500arilog: dos500arilog	-1.047	0.327
ttall.lm.weight.dosenari: dosenari	-0.818	0.429
ttall.lm.weight.dosenord: dosenord	-1.703	0.117
ttall.lm.weight.dosenarilog: dosenarilog	-1.128	0.295

Inclusion of the control vs. high dose comparison I

- The power comparison between control vs. highdose (CvsH) and the Tukey test:
  CvsH-test is superior to Tukey's contrast test only when all the middle doses have identical or near identical response ([2]).
- Because such pattern occur not too frequently, they recommend Tukey's trend test alone.
- Instead focusing on the decision between CvsH-test and Tukey-test: CvsH-test can be included as 4^{th} model in Tukey's trend test.
- A two-sample CvsH test is identical to a linear regression model for only two groups (by ignoring all other dose groups- technically in R transforming into NA's).

Inclusion of the control vs. high dose comparison II

- The multiplicity penalty for including a further model will be small, because a CvsH-model is highly correlated with the 3 other models.
- Notice, the comparison of the CvsH-test alone (i.e. ignoring all other doses) has a special meaning in trend testing. It is a single contrast within the [25] or [9] multiple contrast test and its combination ([13]) and is the first test in the closed testing procedure under order restriction ([11]).

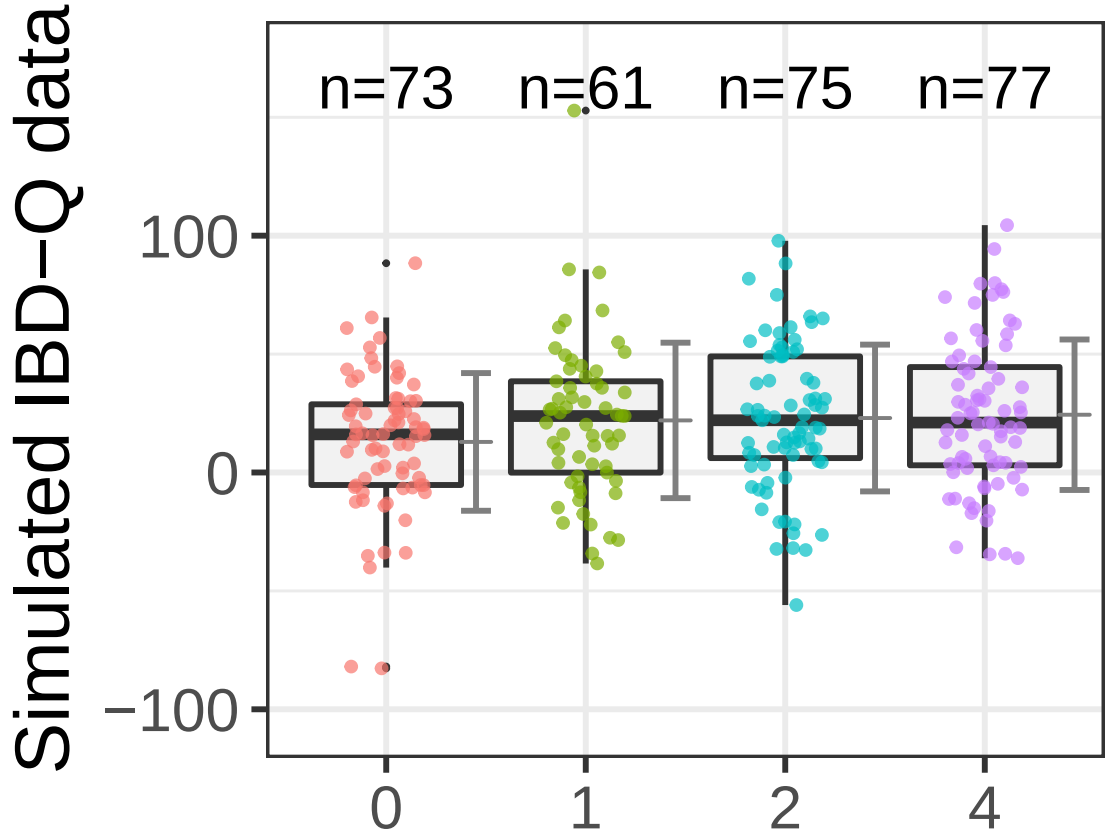
```
ttw6 <- tukeytrendfit(fitw, dose="dosen",
  scaling=c("ari", "ord", "log", "arilog", "highvslow"))
Exa6 <- summary(glht(model=ttw6$mmm, linfct=ttw6$mlf, alte
```

Additional testing vs. pooled doses I

- For plateau-shaped dose-response relationships testing vs. the pooled doses may be of interest.
- E.g., in the randomized placebo-controlled dose finding trial with golimumab to treat ulcerative colitis [21] unadjusted p-values for comparisons of all individual doses vs. placebo and as well against the pooled doses.
- As an example the health-related quality according to IBD-Questionnaire (change from baseline after 6 weeks) was used (see their Table 2).

Additional testing vs. pooled doses II

- Only summary data are available and therefore simulated normal distributed raw data were generated for the analysis here.



Additional testing vs. pooled doses III

```
lmRU <-glm(IBDQ~dose, data=rut)
Cmat3 <-  c(-3, 1, 1, 1)
EXRU <- tukeytrendfit(lmRU, dose="dose",
  scaling=c("ari", "ord", "arilog", "treat"), ctype=Cmat3)
EXRRU <- summary(glht(model=EXRU$mmm, linfct=EXRU$mlf))
```

Table : Tukey Trend Test and pooled dose vs. placebo test

Dose metameter	Test statistics	*p*-value
doseari: doseari	2.0435	0.0742
doseord: doseord	2.2206	0.0494
dosearilog: dosearilog	2.2206	0.0493
dosetreat: 1	2.4785	0.0246

- The p-values for a trend (log-transformed) is about 0.05, but for the pooled doses vs. placebo 0.025, not surprizing for such a plateau shape.

Considering of different covariance-adjusting models I

- Interesting: simultaneous consideration of multiple models with different formulations for the adjusting covariate
- Nonclin example: Analysis of organ weights in toxicology. It is a priori not clear how to consider the body weight: i) as relative organ weight, ii) body weight as covariate or iii) ignoring body weight (quite different biological background!)
- Liver weights from a 13-week study on female F344 rats administered with sodium dichromate dihydrate [1]

```
data("liv", package="SiTuR")
liv$relLiv <-liv$LiverWt/liv$BodyWt
LIVmod1<-lm(LiverWt~Dose, data=liv)
LIVmod2<-lm(relLiv~Dose, data=liv)
LIVmod3<-lm(LiverWt~Dose+BodyWt, data=liv)
tt1<- tukeytrendfit(LIVmod1, dose="Dose", scaling=c("ari", "ord",
tt2<- tukeytrendfit(LIVmod2, dose="Dose", scaling=c("ari", "ord",
tt3<- tukeytrendfit(LIVmod3, dose="Dose", scaling=c("ari", "ord",
cttC <- combtt(tt1,tt2, tt3)
LIVExa4 <- summary(glht(model=cttC$mmm, linfct=cttC$mlf))
```

Considering of different covariance-adjusting models II

	Model	Test stats	p-value
1	tt1.lm.LiverWt.Doseari: Doseari	-6.0358901	0.0000000
2	tt1.lm.LiverWt.Doseord: Doseord	-5.1678800	0.0000006
3	tt1.lm.LiverWt.Dosearilog: Dosearilog	-5.1678800	0.0000017
4	tt2.lm.relLiv.Doseari: Doseari	-4.7739259	0.0000055
5	tt2.lm.relLiv.Doseord: Doseord	-4.6579781	0.0000495
6	tt2.lm.relLiv.Dosearilog: Dosearilog	-4.6579781	0.0000074
7	tt3.lm.LiverWt.Doseari: Doseari	-2.4553875	0.0507073
8	tt3.lm.LiverWt.Doseord: Doseord	-2.9008284	0.0149210
9	tt3.lm.LiverWt.Dosearilog: Dosearilog	-2.9008284	0.0142968

- We pay a tiny penalty to consider several, similar models- because they a highly correlated
- We achieve: the *best* model. Here: just liver weight
- Hard to solve this problem with model selection approaches (different models, but also different endpoints)
- Can be used e.g. in clinical Phase II dose findings studies with baseline values

Generalizations of Tukey trend test I

1. Modification for variance heterogeneity using sandwich estimator of var-cov matrix $\Rightarrow$ vignette
2. Extension to multiple endpoints: normal, binary, multinomial, different-scaled
3. Modification for different arithmetic-logarithmic scores $\Rightarrow$ vignette
4. GLM: Proportions, Poisson (overdispersed) $\Rightarrow$ vignette
5. Rank regression: both normal or rank $\Rightarrow$ paper with Frank Konietschke UniTexas soon
6. Mixed effects model: repeated measures $\Rightarrow$ see the vignette
7. Nonlinear regression models (next Dr. Christian Ritz, Copenhagen)

Further recent general mmm applications I

- Multiple normal distributed endpoints in multi-arm trials: Dunnett-type sCI
- Multiple binary endpoints in multi-arm trials: Williams-type trend test
- Multiple regression models in genetic association test [14]
- Composite binary endpoints [17]
- Subgroup analysis with claim for total, targeted and complementary populations [24]
- Inference on dose (randomized) and time (dependent) [19]

Take home message I

- Trend tests for several possible shapes are available with in glmm framework $\rightarrow$ rather relevant for nonclin and clin dose-finding studies
- TukeyWilliams trend test can be recommended: R library available, simple interpretation
- Flexibility is amazing
- Use confidence intervals (not shown above). Interpret the effect size first, e.g. slopes
- Comparison with nonlinear models (incl. model averaging resp. model selection) needed
- Basic property of mmm used: no explicit formulation of $\mathbb{R}$ needed!

References I

[1] *National Toxicology Program. 13 Weeks gavage study on female F344 rats administered with Sodium dichromate dihydrate (VI) (CASRN: 7789-12-0, Study Number: C20114,TDMS Number:2011402*.

[2] Girish Aras, Allen Xue, and Thomas Liu. Tukey's contrast test versus two-sample test in a dose-response clinical trial. *Statistics In Biopharmaceutical Research*, 3(1):31–39, February 2011.

[3] P. Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, 1955.

[4] D. J. Bartholomew. A test of homogeneity for ordered alternatives. *Biometrika*, 46(1-2):36–48, 1959.

[5] F. Bretz, L. A. Hothorn, and J. C. Hsu. Identifying effective and/or safe doses by stepwise confidence intervals for ratios. *Statistics in Medicine*, 22(6):847–858, March 2003.

[6] Frank Bretz. An extension of the Williams trend test to general unbalanced linear models. *Computional Statistics and Data Analysis*, 50(7):1735–1748, 2006.

[7] E. Brunner and U. Munzel. The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal*, 42(1):17–25, 2000.

[8] G. Dilba, F. Bretz, L. A. Hothorn, and V. Guiard. Power and sample size computations in simultaneous tests for non-inferiority based on relative margins. *Statistics in Medicine*, 25(7):1131–1147, April 2006.

[9] C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *J Am Stat Assoc*, 50(272):1096–1121, 1955.

[10] A. J. Hayter. Inferences on the difference between future observations for comparing two treatments. *Journal of Applied Statistics*, 40(4):887–900, APR 1 2013.

[11] L. A. Hothorn, M. Neuhauser, and H. F. Koch. Analysis of randomized dose-finding-studies: Closure test modifications based on multiple contrast tests. *Biometrical Journal*, 39(4):467–479, 1997.

[12] L.A. Hothorn. The two-step approach - a significant ANOVA F-test before Dunnett's comparisons against a control - is not recommended. *Communications in Statistics*, 2015.

[13] T. Jaki and L. A. Hothorn. Statistical evaluation of toxicological assays: Dunnett or Williams test-take both. *Archives of Toxicology*, 87(11):1901–1910, November 2013.

[14] A. Kitsche, C. Ritz, and L.A. Hothorn. A general framework for the evaluation of genetic association studies using multiple marginal models. *Human Heredity*, 2016.

References II

[15] F. Konietschke and L.A. Hothorn. Rank-based multiple test procedures and simultaneous confidence intervals. *Electron J Stat*, 6:738–759, 2012.

[16] T. Kopp, E. Riedl, C. Bangert, E. P. Bowman, E. Greisenegger, A. Horowitz, H. Kittler, W. M. Blumenschein, T. K. McClanahan, T. Marbury, C. Zachariae, D. L. Xu, X. S. Hou, A. Mehta, A. S. Zandvliet, D. Montgomery, F. van Aarle, and S. Khalilieh. Clinical improvement in psoriasis with specific targeting of interleukin-23. *Nature*, 521(7551):222–+, May 2015.

[17] GrosseRuse M., C. Ritz, and L.A. Hothorn. Simultaneous inference of a binary composite endpoint and its components. *J. Biopharm Statist*, 2016.

[18] P. Pallmann and L. A. Hothorn. Analysis of means: a generalized approach using r. *Journal of Applied Statistics*, 43(8):1541–1560, June 2016.

[19] 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.

[20] C. B. Pipper, C. Ritz, and H. 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.

[21] P. Rutgeerts, B. G. Feagan, C. W. Marano, L. Padgett, R. Strauss, J. Johanns, O. J. Adedokun, C. Guzzo, H. Zhang, J. F. Colombel, W. Reinisch, P. R. Gibson, and W. J. Sandborn. Randomised clinical trial: a placebo-controlled study of intravenous golimumab induction therapy for ulcerative colitis. *Alimentary Pharmacology & Therapeutics*, 42(5):504–514, September 2015.

[22] E. J. Ryu and A. Agresti. Modeling and inference for an ordinal effect size measure. *Statistics in Medicine*, 27(10):1703–1717, May 2008.

[23] J. W. Tukey, J. L. Ciminera, and J. F. Heyse. Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics*, 41(1):295–301, 1985.

[24] C. Vogel. Model-based simultaneous inference for multiple subgroups and multiple endpoints. *Report*, 2016.

[25] D.A. Williams. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, 27(1):103–117, 1971.