Nonclinical Statistics Conference Paris 2018

Lessons learned from the glyphosate case to evaluate **long-term carcinogenicity assays**: multiple studies with different dose levels and multiple correlated binary endpoints

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

20. September 2018

Biostatistical issues for the next 40 mins I

- Generalized linear mixed effect models for weighted binomials
- Trend test considering dose quantitatively- remember my Cambridge 2016 tutorial
- Trend test and pairwise tests
- Max-test for multiple correlated tumor incidences
- Use R!
  Sorry for non-R-user: explaining ideas by R-code

A data example- to clarify the problems I

- Glyphosate male mouse malignant lymphoma [12, 15]
- The story of Glyphosate (an agro-chemical) is an example only. No statement whether positive or not in this talk

Year	Strain	Dura	Doses	Crude prop $p_i = r_i/n_i$	p^{Poly3}
1983	Crl:CD1	24	0/157/814/4841	2/50, 5/49, 4/50, 2/50	0.51
1993	CD1	24	0/100/300/1000	4/50, 2/50, 1/50, 6/50	0.08
1997	CrJ:CD1	18	0/165/838/4348	2/50, 2/50, 0/50, 6/50	0.012
2001	SW	18	0/15/151/1460	10/49,15/49,16/49,19/49	0.09
2009	Crl:CD1	18	0/71/234/810	0/51, 1/51, 2/51, 5/51	0.005

A data example- to clarify the problems II

- Is a joint analysis feasible?

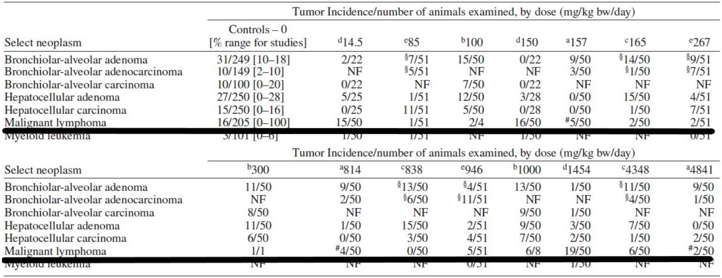
1. Over 30 years
2. Different strains
3. Different durations
4. Still the same NTP design (no. D, n_i)
5. **Quite different doses**: D_3 by factor 6, $D_{j,3} < D_{j,2}$
6. Different dose spacings $D_3/D_1 = 27.6, 10, 26.4, 97.3, 11.4$
7. Quite different shapes: monotone 0,1,2,5 to non-monotone 2,5,4,2
8. Extreme different spontaneous rate $0/50,...,10/49$. Remember p_0 effect in prop tests!
9. (No. animal at risk unrealistic uniformly)?
10. **Mortality** data not available
11. Historical controls per assay not available

- Isn't it all simple? Use many Fisher exact tests. No!

A data example- to clarify the problems III

- **Issue I**: Conclusions of the German Toxicology Chief [4] i) *all rates within range of historical controls*, ii) *lack of a dose-response across the several orders of magnitude*, i.e. monotone d-r-pattern as criterion in an inappropriate super-pooled data table

Table 22. Summary of select neoplasms in male mice (Studies 10–14).



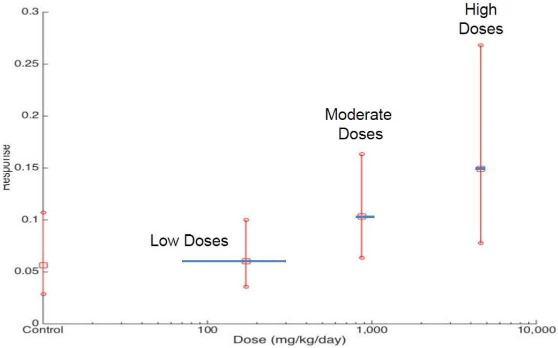
	Tumor Incidence/number of animals examined, by dose (mg/kg bw/day)							
Select neoplasm	Controls – 0 [% range for studies]	[d]14.5	[e]85	[b]100	[d]150	[a]157	[c]165	[e]267
Bronchiolar-alveolar adenoma	31/249 [10–18]	2/22	[§]7/51	15/50	0/22	9/50	[§]14/50	[§]9/51
Bronchiolar-alveolar adenocarcinoma	10/149 [2–10]	NF	[§]5/51	NF	NF	3/50	[§]1/50	[§]7/51
Bronchiolar-alveolar carcinoma	10/100 [0–20]	0/22	NF	7/50	0/22	NF	NF	NF
Hepatocellular adenoma	27/250 [0–28]	5/25	1/51	12/50	3/28	0/50	15/50	4/51
Hepatocellular carcinoma	15/250 [0–16]	0/25	11/51	5/50	0/28	0/50	1/50	7/51
Malignant lymphoma	16/205 [0–100]	15/50	1/51	2/4	16/50	[#]5/50	2/50	2/51
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

	Tumor Incidence/number of animals examined, by dose (mg/kg bw/day)							
Select neoplasm	[b]300	[a]814	[c]838	[e]946	[b]1000	[d]1454	[c]4348	[a]4841
Bronchiolar-alveolar adenoma	11/50	9/50	[§]13/50	[§]4/51	13/50	1/50	[§]11/50	9/50
Bronchiolar-alveolar adenocarcinoma	NF	2/50	[§]6/50	[§]11/51	NF	NF	[§]4/50	1/50
Bronchiolar-alveolar carcinoma	8/50	NF	NF	NF	9/50	1/50	NF	NF
Hepatocellular adenoma	11/50	1/50	15/50	2/51	9/50	3/50	7/50	0/50
Hepatocellular carcinoma	6/50	0/50	3/50	4/51	7/50	2/50	1/50	2/50
Malignant lymphoma	1/1	[#]4/50	0/50	5/51	6/8	19/50	6/50	[#]2/50
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

- *i) Pooling $p_{j,0}$ inapprop., ii) pooling studies inapprop., iii) ignoring mortality inapprop., iv)....*

A data example- to clarify the problems IV

- **Issue II**: Most recent paper [12]: *trend test results should not be played off against those from pairwise comparisons*. See ⇓
- **Issue III**: [1] EchA categorization of quite different study-specific dose levels into a single pseudo study- problematic!



A data example- to clarify the problems V

- **Issue IV**: *Be safe in negative results*. But proof of safety not used in routine. Today **proof of hazard**, still considering a specific false +/- relationship
- **Issue V**: Interpretation and joint analysis of **multiple bioassays** NOT defined in a guidance or publication
- **Issue VI**: Historical controls: div papers including [8], [11], [13]
- **Issue VII**: Asymmetry of chi2 test: depends on p_0
- **Issue VIII**: Using pooled 2-by-k table data for each tumor site and naive Fisher tests? **These bioassays are complex and therefore appropriate and complex tests and their specific interpretation should be mandatory**

A data example- to clarify the problems VI

- **Issue IX**: Multiplicity
  - i Multiple doses, tumor sites, sex (males,females), studies, classifications (pre-neoplasia, adenoma, carcinoma, combined), trend and pairwise tests
  - ii Missing relevance criteria: [3] *Because of the large number of comparisons involved (usually 2 species, 2 sexes, and 30 or more tissues examined), a great potential exists for finding statistically significant positive trends or treatment-placebo differences due to chance alone (i.e., a false positive). Therefore, it is important that an overall evaluation of the carcinogenic potential of a drug take into account the multiplicity of statistical tests of significance for both trends and pairwise comparisons.*
  - iii (NTP 2 species, 2 sexes) Overall 10% false+: i) **Trend test** common and rare tumors are tested at 0.005 and 0.025 levels ii) **Control-High Pairwise Comparisons** 0.01 and 0.05

A data example- to clarify the problems VII

iv Criterion positive trend:

1. common CA-test is for linear regression (optimal power when linear, but similar sensitive for sublinear shapes , up to $[0, 0, 0, 0, \delta]$ that is a trend, but less sensitive for supralinear trends $[0, \delta, ..., \delta]$
2. trend and pairwise tests (several definitions pairwise: only vs. Dmax, pairwise vs. control at α, Dunnett-type tests) extreme inconsistent from stats view FWER and CWER. Why they do this at all? Probably because be sensitive for downturn effects; partly still changing the underlying test principles (exact, asymp)
3. 2-sided vs. 1 sided (tumor trend inherently 1-sided for an increase) [12] but NTP *neoplasm: reported P values are one sided* trend test at all 1-sided

- **Issue X**: p-value of a test is still used as a relevance criterion, e.g. p=0.003 for a single trend test. This is only the second best choice, but if you use the NTP design (n_i. no doses, dose choice, etc.), adjust the spontaneous rate with the historical controls, and just take appropriate tests, acceptable

Tumor development and mortality I

- Primary endpoint: number of tumors (of a certain classification) in relation to number of animals at risk $p_i = r_i/n_i$
- Primary inference $p_i > p_0$, any i in a NTP design $D_0, D_1, D_2, D_3, i = 0, 1, 2, 3$
- Specific relationship between tumor development and mortality

  i Most tumors can be diagnosed in dead animals only

  ii Tumor can be fatal (ie cause for mortality) or incidental (no cause of death, but found in dead animals). But microscopic classification into fatal/incidental can be difficult

  iii Early death prevents the development of tumors that may occur at a later stage, ie high early mortality can increase f- for tumor incidences!

- In history (and guidelines) stratified 2-by-k table test for fatal and incid. tumors (and their combination)- too complex for toxicologists

Tumor development and mortality II

- Use poly-k adjustment

i A modification of the Cochran –Armitage test [2] modeling survival time by a 2P Weibull distribution.

ii To account for censoring due to treatment-specific mortality by individual weights $w_{ij} = (t_{ij}/t_{max})^k$ reflecting individual mortality pattern (t_{ij} is time of death of animal j in treatment i).

iii Weibull shape parameter $k = 3$ seems to be a good empirical choice. Is it really?

iv These weights result in adjusted sample sizes $n_i^* = \sum_{j=1}^{n_i} w_{ij}$ (which are used instead of the randomized number of animals n_i)

v Therefore adjusted proportions $p_i^* = y_i/n_i^*$ are used instead of the crude tumor proportions $p_i = y_i/n_i$

vi Not a perfect adjustment for all shapes of survival functions, but acceptable [9]

vii But, CA-trend test is sensitive to near-to-linear shapes only. We need a test, which is sensitive to most shapes ⇒ today

viii We need a generalization in the glmm ⇒ today

Tumor development and mortality III

Summary I: poly-k adjustment results in weighted glm or glmm models for log(OR) as effect size

Trend test for multiple studies with quite different dose levels I

- **Very different doses** leads to the question: how to evaluate a **dose-response relationship** at all for adjusted proportions?
- Primarily a trend test should be used, sensitive to all possible shapes (including a downturn effect), adjusted for possible group-specific mortality, adjusted against spontaneous rates of suitable selected historical controls, taking into account the distance of this $\hat{p_{0i}}$ from zero, expandable for multiple tumors within a study as well as for multiple studies
- Coming back to my Cambridge 2016 talk

Trend test for multiple studies with quite different dose levels II

- Tukey's trend test [16] based on ξ multiple linear regression models for the ξ dose transformation functions $\psi^{\xi}(D_j)$ (for the arithmetic, ordinal, and linear-log dose metameters) 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

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

- 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 an union-intersection test (UIT).

Trend test for multiple studies with quite different dose levels III

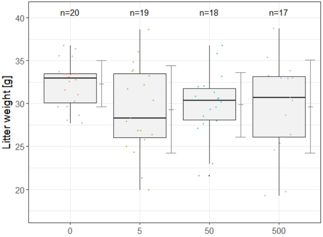
- From these parameter estimates the correlation matrix is estimated and the test is on the ξ (respective $(\xi * I)$) slope parameters $\beta_{i\xi}$.
- Joint distribution of parameter estimates from **multiple marginal models** [14]- without assuming a certain multivariate distribution for the data
- Available as function `mmm` in library(multcomp)
- I.e. correlations between different parameter estimates obtained from different model fits to the same data. No explicit calculation of R needed!
- Alternatively, simultaneous confidence intervals for the single parameter `slope` available- more appropriate for interpretation!
- Remark: nonlinear models try an optimal fit, but need several parameters. Remember: *All models are wrong, some are useful*

Trend test for multiple studies with quite different dose levels IV

- **Covers a wide range of dose-response patterns**
- Recent GLMM-generalization and CRAN-library(tukeytrend)
- For appropriate chosen df ν, finite versions works well (various simulations by Drs. Pallmann, Schaarschmidt, Ristl and me)
- 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* → belonging to the same lm-class. The difficult problem of estimating $\mathbb{R}$ can be easily solved by mmm

Trend test for multiple studies with quite different dose levels V

- Extension of the Tukey trend test:
  i) three regression models for the arithmetic, ordinal, and logarithmic-linear dose metameters [16] **AND** ii) Williams multiple contrast
- Example: litter weight data [7]



- Decreasing weights is the possible toxic effect. No clear trend. A possible dose plateau?

Trend test for multiple studies with quite different dose levels VI

- Therefore, 4 marginal models for 6 hypotheses needed:
  3 regression models for arithmetic, ordinal and log-linear dose metameters **and** 3 Williams-type multiple contrasts
- Notice, small sample t-distributed version!

```
litter$dosen <- as.numeric(as.character(litter$dose)) # add a numeric dose var
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	0.727
dosenord: dosenord	-1.703	0.212
dosenarilog: dosenarilog	-1.128	0.519
dosentreat: C 1	-1.863	0.156
dosentreat: C 2	-2.287	0.062
dosentreat: C 3	-2.759	0.018

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

Trend test for multiple studies with quite different dose levels VII

- More general:

1. Power of Tukey trend test depends on dose metameters, design ...
2. Some simulation results

shape	Williams	Tukey	TukeyWil
dose	quali	quanti	both
linear	0.85	0.89	0.87
plateau	0.95	0.76	0.87
sublinear	0.81	0.96	0.89

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

Trend test for multiple studies with quite different dose levels VIII

Summary II: Trend test for dose as quantitative covariate (allows different dose levels in ζ biossays) AND/OR qualitative levels is available in this framework

A test for strict monotone trend I

- US-FDA 2001 draft guidance [3] recommended the evaluation of individual tumors by a trend test **or** pairwise comparison C vs. D_{high}: trend test $\alpha = 0.005$; pairwise tests $\alpha = 0.01$ for common tumors (for rare tumors 0.025, 0.05) (to achieve an overall false positive rate of about 10%)
- Recently an alternative decision rule for a strict monotone trend: trend test **and** pairwise test C vs. D_{high} simultaneously [10] (joint test).

  This logical AND operation represents an intersection-union test (IUT).

  The elementary tests within an IUT are performed at level α to control FWER.

A test for strict monotone trend II

- However, IUT's are conservative by definition, which is an undesirable property for the specific ratio of $f+/f-$. Conservativity $\Rightarrow$ reduced by using correlation between the tests, unfortunately this is not yet available for the IUT [5]. Moreover, they allow only the global decision *trend and pairwise*
- Alternative: max-t test, an UIT, specifically defined for an all-pairs power alternative [6]. In principle, every UIT allows all patterns of elementary decisions: both the trend test and the pairwise test
- Advantages max-t test: i) quantile $\Downarrow$ with $\Uparrow$ correlation, ii) adjusted p-value as well as compatible simultaneous confidence intervals are available
- Armitage and Williams trend test are formulated for a monotone alternative, but they are significant for nonmonotone shapes, e.g. $\pi_0 = 0, \pi_1 = \delta, \pi_{2 \ldots k-1} = 2\delta, \pi_k = \delta$

A test for strict monotone trend III

- Simulations assuming normal distributed homoscedastic errors in a balanced $k = 3 + 1, n_i = 20$ design

1. LogR ... linear regression
2. LogH ... linear regression jointly with HvsC contrast (UIT- or)
3. LogIU ... linear regression jointly with HvsC contrast (IUT - and)
4. Tuk ... Tukey type trend test (max(ari,ord,log))
5. TukH ... Tukey type trend test jointly with HvsC contrast (UIT- or)
6. IUT ... Tukey type trend test jointly with HvsC contrast (IUT- and)
7. TuW ... Tukey type trend test jointly with Williams contrasts (UIT- or)
8. Wil ... Williams multiple contrast test
9. TWIUT ... Tukey type trend test jointly with Williams and HvsC contrast (IUT- and)
10. LinRa ... LinRahman type test: IUT linear logistic regression and t-test, each at alpha

Shape	Mo	LogR	LogH	LogIU	Tuk	TuH	**UIT**	TuW	Wil	TWIUT	LinR
H0	y	0.049	0.049	0.027	0.049	0.049	0.037	0.050	0.046	0.027	0.035
0,0,0,d	y	0.894	0.920	0.849	0.887	0.887	0.902	0.890	0.886	0.881	0.884
lin	y	0.946	0.941	0.893	0.947	0.947	0.898	0.934	0.919	0.880	0.909
0,0,d,d	y	0.988	0.984	0.910	0.991	0.991	0.904	0.982	0.966	0.890	0.923
0,d,d,d	y	0.906	0.926	0.859	0.929	0.929	0.915	0.982	0.982	0.897	0.894
0,0,d,2/3d	no	0.219	0.171	0.035	0.340	0.340	0.033	0.471	0.502	0.024	0.048
0,0,d,1/3d	no	0.000	0.000	0.000	0.000	0.000	0.000	0.026	0.029	0.000	0.000
0,0,d,4/5d	no	0.808	0.752	0.446	0.842	0.842	0.434	0.821	0.804	0.378	0.502
0,d,d,4/5d	no	0.386	0.412	0.280	0.500	0.500	0.382	0.899	0.904	0.333	0.355
0,d,d,2/3d	no	0.037	0.044	0.020	0.084	0.084	0.040	0.737	0.750	0.027	0.032

A test for strict monotone trend IV

Interpretation:

1. Both UIT and IUT are conservative; IUT even more
2. My favorite IUT (Tukey type trend test jointly with HvsC contrast (IUT- and) reveals a similar power behavior as LinRahman test, but allows a conclusion on trend only (or C vs. High) within the FEWR control
3. UIT-joint test allows adjusted p-values and /or simultaneous confidence intervals
4. All joint tests are extreme sensitive to downturn shapes
5. TuW (Tukey type trend test jointly with Williams contrasts) is more powerful for plateau shapes than any regression tests alone.
6. Power differences became smaller when power $1-\alpha$ approaching
7. UIT-joint test can be recommended, when testing for strict monotone trend
8. Consideration for adjusted proportions next

A test for strict monotone trend V

- Notice, substantial different f- rates for
  trend AND C vs. D_{high}**,**
  trend OR C vs. D_{high}**,**
  trend OR C vs. D_i
- Overdosing is an issue in tox at all (to limit f- decisions), to some extend in long-term carcinogenicity studies, too.
  I.e. downturn effect at the high dose possible.
  I.e. an UIT for
  $max(trend^{C,D_1,D_2,D_3}, pairw(C-D_3), trend^{C,D_1,D_2}, pairw(C-D_2))$
  can be formulated easily

A test for strict monotone trend VI

Summary III: Joint test [Trend test AND/OR pairwise C vs. D_{max}] can be recommended and is available in this framework

A glmm version of Tukey type trend test for poly-k adjusted proportions for multiple studies I

- No access for me to Glyphosate raw data (animal-specific tumor, death,...) A shame
- Toy example: males and females in US-NTP data base. Zymbal adenoma or carcinoma in TR365 for male and female rats
- Data snippet

	sex	dose	zymbal	time	weightpoly3
1	male	0	0	69	0.28
2	male	0	0	77	0.38
3	male	0	0	81	0.45
4	male	0	0	83	0.48
5	male	0	0	85	0.52
...	male	...	...	...	...
108	male	50	1	99	1.00
140	male	50	0	106	1.00
141	female	0	0	72	0.31
142	female	0	0	80	0.43
143	female	0	0	81	0.45
144	female	0	0	81	0.45
145	female	0	0	88	0.57
146	female	0	0	90	0.61
...	female	...	...	...	...
225	female	50	1	99	1.00

A glmm version of Tukey type trend test for poly-k adjusted proportions for multiple studies II

- glmm: using partial least square algorithm in library(MASS)

```
library(MCPAN); library(multcomp); library(tukeytrend); library(MASS)
# study-specific poly3 weights
zymF$weightpoly3 <- 1 # Compute the poly-3 (-k)- weights at the level of singl
wt0f <- which(zymF$zymbal == 0)
zymF$weightpoly3[wt0f] <- (zymF$time[wt0f]/max(zymF$time))^3
#.... dito for males
ZYM<-rbind(zymM,zymF) # joint data with poly3 weights
TN1 <- dosescalett(ZYM, dose="dose", scaling=c("ari", "ord", "arilog"))$data
glmmari1T <- glmmPQL(fixed=zymbal ˜ doseari, random = ˜ 1 |sex,
glmmord1T <- glmmPQL(fixed=zymbal ˜ doseord, random = ˜ 1 |sex,
                   family = binomial, data=TN1, niter = 100)
glmmarilog1T <- glmmPQL(fixed=zymbal ˜ dosearilog, random = ˜ 1 |sex,
                   family = binomial, data=TN1)
lmari1T <- tukeytrend:::lmer2lm(glmmari1T)
lmord1T <- tukeytrend:::lmer2lm(glmmord1T)
lmarilog1T <- tukeytrend:::lmer2lm(glmmarilog1T)
linf <- matrix(c(0,1), ncol=2)
mm1T <- glht(mmm("mari"=lmari1T, "mord"=lmord1T, "marilog"=lmarilog1T),
```

A glmm version of Tukey type trend test for poly-k adjusted proportions for multiple studies III

- Result (do'nt be surprized: $D_i = 0, 25, 50$)

Model	Test stats	p-value
ari: 1	2.14	0.016
ord: 1	2.14	0.016
arilog: 1	2.14	0.016

Tabelle: Tukey-type test for poly3 estimates using a mixed effect model for 2 studies

A glmm version of Tukey type trend test for poly-k adjusted proportions for multiple studies IV

Summary IV: Four approaches

i) glmmPQL (library(MASS),

ii) mmm (library(tukeytrend)),

iii) poly3-weights (library(MCPAN)),

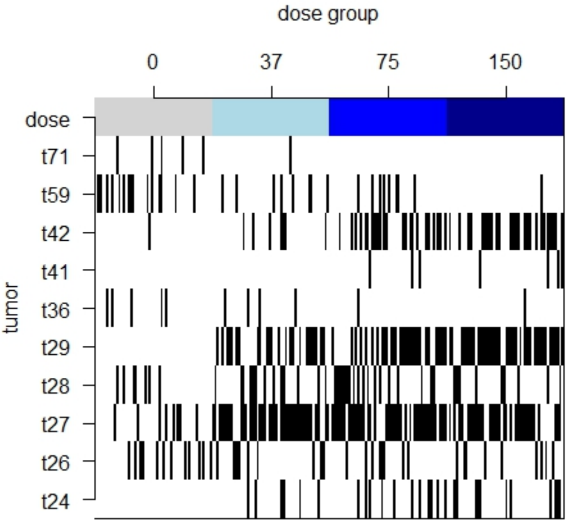
iv) max-test (mmm in library(multcomp))

allow joint analysis of mortality-adjusted tumor rates in repeated bioassays with different dose levels for a single selected tumor (or classification), robust against many patterns of dose-response

Multiple tumors I

- Commonly, up to about 30 tumor sites are diagnosed, where also classifications (adenoma, carcinoma, combined, body systems) are used
- Commonly, univariate analysis, each at level α is performed. Alternatively, a max-test (commonly min-p) can be recommended
- Here, a max-test, taken the correlation into account is used
- Example ([7]), 4 treatment groups (doses 0, 37, 75, 150), each containing 50 mice, have been investigated for presence or absence of 89 different tumor classifications (t01,...,t89)- here restricted to those 10 tumor classifications, that show an overall abundance more than 5.

Multiple tumors II



Multiple tumors III

- Max-test over correlated proportion for Tukey-type trend test

```
N24i <- glm(t24 ~ dose, data=miceF, family=binomial())
N26i <- glm(t26 ~ dose, data=miceF, family=binomial())
N27i <- glm(t27 ~ dose, data=miceF, family=binomial())
N28i <- glm(t28 ~ dose, data=miceF, family=binomial())
N29i <- glm(t29 ~ dose, data=miceF, family=binomial())
N36i <- glm(t36 ~ dose, data=miceF, family=binomial())
N41i <- glm(t41 ~ dose, data=miceF, family=binomial())
N42i <- glm(t42 ~ dose, data=miceF, family=binomial())
N59i <- glm(t59 ~ dose, data=miceF, family=binomial())
N71i <- glm(t71 ~ dose, data=miceF, family=binomial())

tu24i <- tukeytrendfit(N24i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu26i <- tukeytrendfit(N26i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu27i <- tukeytrendfit(N27i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu28i <- tukeytrendfit(N28i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu29i <- tukeytrendfit(N29i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu36i <- tukeytrendfit(N36i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu41i <- tukeytrendfit(N41i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu42i <- tukeytrendfit(N42i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu59i <- tukeytrendfit(N59i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu71i <- tukeytrendfit(N71i, dose="dose", scaling=c("ari", "ord", "arilog"))

tt10 <- combtt(tu24i, tu26i, tu27i, tu28i, tu29i, tu36i, tu41i, tu42i, tu59i, tu71i)
stt10 <- summary(asglht(tt10))
```

Multiple tumors IV

Model	Test stats	p-value
tu24i.glm.t24.doseari: doseari	3.05	0.03016
tu24i.glm.t24.doseord: doseord	3.29	0.01401
tu24i.glm.t24.dosearilog: dosearilog	3.29	0.01407
tu26i.glm.t26.doseari: doseari	-0.67	0.99938
tu26i.glm.t26.doseord: doseord	-0.80	0.99668
tu26i.glm.t26.dosearilog: dosearilog	-0.80	0.99673
tu27i.glm.t27.doseari: doseari	3.60	0.00465
tu27i.glm.t27.doseord: doseord	4.41	0.00027
tu27i.glm.t27.dosearilog: dosearilog	4.40	0.00027
tu28i.glm.t28.doseari: doseari	0.31	1.00000
tu28i.glm.t28.doseord: doseord	0.82	0.99613
tu28i.glm.t28.dosearilog: dosearilog	0.82	0.99613
tu29i.glm.t29.doseari: doseari	6.44	0.00000
tu29i.glm.t29.doseord: doseord	6.79	0.00000
tu29i.glm.t29.dosearilog: dosearilog	6.79	0.00000
tu36i.glm.t36.doseari: doseari	-1.84	0.52274
tu36i.glm.t36.doseord: doseord	-1.98	0.41782
tu36i.glm.t36.dosearilog: dosearilog	-1.98	0.41682
tu41i.glm.t41.doseari: doseari	2.26	0.24347
tu41i.glm.t41.doseord: doseord	2.21	0.26921
tu41i.glm.t41.dosearilog: dosearilog	2.21	0.26960
tu42i.glm.t42.doseari: doseari	5.65	0.00000
tu42i.glm.t42.doseord: doseord	5.78	0.00000
tu42i.glm.t42.dosearilog: dosearilog	5.79	0.00000
tu59i.glm.t59.doseari: doseari	-3.41	0.00937
tu59i.glm.t59.doseord: doseord	-3.45	0.00825
tu59i.glm.t59.dosearilog: dosearilog	-3.44	0.00827
tu71i.glm.t71.doseari: doseari	-2.02	0.38915
tu71i.glm.t71.doseord: doseord	-2.11	0.33298
tu71i.glm.t71.dosearilog: dosearilog	-2.10	0.33880

Multiple tumors V

Summary V: Max-test on correlated tumor incidences works, is conservative, can be extended to glmm

Take home I

- Available and mandatory:

  i use poly-k

  ii use best k (not discussed today, sorry)

  iii use trend test taking dose quantitatively into account

  iv use trend test protected against possible downturns

  v use trend test alone or trend test AND C vs D_{high} for strict monotone trend

  vi use max test for multiple tumors (or pooled classifications)

  vii use generalized linear mixed effect model over bioassays

  ... **use i)-vii) jointly**. CRAN packages available. More work needed for robustness and f^{-}/f^{+}

  viii use historical control p^{poly-k} (not discussed today, sorry)

  ix use-one-sided tests for an increase only

  x use NTP design only

  xi use odds ratios and its simultaneous confidence limits instead of p-value

Take home II

Finally

- We need a consensus conference with a following guideline (preferably within ICH) on +/- assessment of an assay: for a single tumor, taking into account competing mortality, for the joint examination of different tumors (classification, context,...), across multiple studies (animal species, strains, applications)
- And really right at the end: the problem is complex. Ends the naive evaluation, e.g. of glyphosate, because it is about life and death on the one hand and a lot of money on the other hand

References I

[1] *Portier, C.J. EchA Slides to Glyphosate 2017*.

[2] A. J. BAILER and C. J. PORTIER. Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. *Biometrics*, 44(2):417–431, June 1988.

[3] Center for Drug Evaluation and Research. Guidance for industry: Statistical aspects of the design, analysis, and interpretation of chronic rodent carcinogenicity studies of pharmaceuticals. Technical report, US Food and Drug Administration, 2001.

[4] H. Greim, D. Saltmiras, V. Mostert, and C. Strupp. Evaluation of carcinogenic potential of the herbicide glyphosate, drawing on tumor incidence data from fourteen chronic/carcinogenicity rodent studies. *Critical Reviews in Toxicology*, 45(3):185–208, March 2015.

[5] M. Hasler. *Extensions of Multiple Contrast Tests*. PhD thesis, Gottfried Wilhelm Leibniz Universität Hannover, 2009.

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

[7] L.A. Hothorn. *Statistics in Toxicoloyg- using R*. Chapman Hall, 2016.

[8] A. Kitsche, L. A. Hothorn, and F. Schaarschmidt. The use of historical controls in estimation simultaneous confidence intervals for comparisons against a concurrent control. *Computational Statistics and Data Analysis*, 56(12):3865–3875, 2012.

[9] R. L. Kodell. Should we assess tumorigenicity with the peto or poly-k test? *Statistics in Biopharmaceutical Research*, 4(2):118–124, May 2012.

References II

[10] 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.

[11] L. G. L. Novelo, A. Womack, H. X. Zhu, and X. W. Wu. A bayesian analysis of quantal bioassay experiments incorporating historical controls via bayes factors. *Statistics in Medicine*, 36(12):1907–1923, May 2017.

[12] Clausing P. Pesticides and public health: an analysis of the regulatory approach to assessing the carcinogenicity of glyphosate in the european union. *Epidemiol Community Health 2018;0:1–5.*, 2018.

[13] S. D. Peddada, G. E. Dinse, and G. E. Kissling. Incorporating historical control data when comparing tumor incidence rates. *Journal of the American Statistical Association*, 102(480):1212–1220, December 2007.

[14] 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.

[15] C. J. Portier and P. Clausing. Re: Tarazona et al. (2017): Glyphosate toxicity and carcinogenicity: a review of the scientific basis of the european union assessment and its differences with iarc. doi: 10.1007/s00204-017-1962-5. *Archives of Toxicology*, 91(9):3195–3197, September 2017.

[16] 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.