A combined approach for claiming equivalence and difference for multiple endpoints with application in GMO-trials [1]

Ludwig A. Hothorn

Institute of Biostatistics
Leibniz University Hannover
e-mail: hothorn@biostat.uni-hannover.de

[1] *Sponsored by the German Federal Ministry of Science and Technology*

University of Hohenheim 2007
Agriculture Field Trials - Today and Tomorrow

The problem I

- At least two kinds of field trials for demonstrating harmlessness of new GMO for:
  i) non-target species, ii) compositional components
- Complex designs including blocks, locations, years (random factors) where an evaluation using a mixed model is appropriate.
  For simplicity here considering a completely randomized one-way layout with the two treatments: GMO and near-isogenic variety
- The long-term acceptance of the isogenic variety is supposed in the environment, for feeding animals and human consumption.
  Therefore inference $\mu_{GMO} - \mu_{iso}$ is appropriate for demonstrating harmlessness of GMO

The problem II

- A multiple endpoint problem exists: **hundreds of species** (including both sexes and development stages) or **hundreds of compositional components** will be observed/ measured: $y_1, ..., y_k$.
- Question: Why is the commonly used non-significance of a point-zero hypothesis test, such as t- or Wilcoxon test, inappropriate? Simply: *Absence of evidence is not evidence of absence* (Altman and Bland, 2004)
- I.e. this proof of hazard is inappropriate, particularly because sample size is not defined (EFSA working group), but sample size matters seriously
- Therefore, the **proof of safety** should be used. The objective here is to formulate a proof of **safety approach for multiple endpoints**

Examples I

- Gathmann et al. (2006): nontarget lepidopteran larvae

Molecular Ecology (2006) **15**, 2677–2685 doi: 10.1111/j.1365-294X.2006.02962.x

Impact of Bt maize pollen (MON810) on lepidopteran larvae living on accompanying weeds

ACHIM GATHMANN,*LUDGER WIROOKS,†LUDWIG A. HOTHORN,‡DETLEF BARTSCH* and INGOLF SCHUPHAN*

*Aachen University, Institute of Environmental Research, Chair of Ecology, Ecotoxicology and Ecochemistry, Worringerweg 1, 52062 Aachen, †Steinkaulstr. 46, 52070 Aachen, ‡University of Hanover, Unit Biostatistics, Herrenhaeuser Str. 2, D-30419 Hanover, Germany

Abstract

Environmental risks of Bt maize, particularly pollen drift from Bt maize, were assessed for nontarget lepidopteran larvae in maize field margins. In our experimental approach, we carried out 3-year field trials on 6 ha total. Three treatments were used in a randomized block design with eight replications resulting in 24 plots: (i) near-isogenic control variety without insecticide (control), (ii) near-isogenic control variety with chemical insecticide (Baytroid) and (iii) Bt maize expressing the recombinant toxin. We established a weed strip (20 × 1 m) in every plot consisting of a *Chenopodium album* (goosefoot)/*Sinapis alba* (mustard) mixture. In these strips we measured diversity and abundance of lepidopteran larvae during maize bloom and pollen shed. *C. album* hosted five species but all in very low densities; therefore data were not suitable for statistical analysis. *S. alba* hosted nine species in total. Most abundant were *Plutella xylostella* and *Pieris rapae*. For these species no differences were detected between the Bt treatment and the control, but the chemical insecticide treatment reduced larval abundance significantly. Conclusions regarding experimental methodology and results are discussed in regard to environmental risk assessment and monitoring of genetically modified organisms.

Keywords: Bt maize, insecticide, Lepidoptera, monitoring, nontarget effects, risk assessment

Received 22 November 2005; revision accepted 6 March 2006

Examples II

- George et al.(2004): compositional components of corn MON 863

JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

J. Agric. Food Chem. 2004, 52, 4149−4158

Composition of Grain and Forage from Corn Rootworm-Protected Corn Event MON 863 Is Equivalent to That of Conventional Corn (*Zea mays* L.)

CHERIAN GEORGE,*,† WILLIAM P. RIDLEY,† JANET C. OBERT,† MARGARET A. NEMETH,† MATTHEW L. BREEZE,‡ AND JAMES D. ASTWOOD†

Product Safety Center, Monsanto Company, 800 North Lindbergh Boulevard, St. Louis, Missouri 63167, and Covance Laboratories, Inc., 3301 Kinsman Boulevard, Madison, Wisconsin 53704

Insect-protected corn hybrids containing event MON 863 protect corn plants against feeding damage from corn rootworm (*Diabrotica*), a major North American insect pest. Corn event MON 863 contains a gene that expresses an amino acid sequence variant of the wild-type Cry3Bb1 insecticidal protein from *Bacillus thuringiensis*. The purpose of this study was to compare the composition of corn containing event MON 863 with that of conventional nontransgenic corn. Compositional analyses were conducted to measure proximates, fiber, amino acids, fatty acids, minerals, folic acid, thiamin, riboflavin, vitamin E, antinutrients, and certain secondary metabolites in grain and proximates and fiber content in forage collected from a total of eight field sites in the U.S. and Argentina. Compositional analyses demonstrated that the grain and forage of event MON 863 are comparable in their nutritional content to the control corn hybrid and conventional corn. These comparisons, together with the history of the safe use of corn as a common component of animal feed and human food, support the conclusion that corn event MON 863 is compositionally equivalent to, and as safe and nutritious as, conventional corn hybrids grown commercially today.

KEYWORDS: Corn (*Zea mays* L.); corn rootworm; insect-protected corn; composition

One- or two-sided hypotheses?

- The formulation of two-sided hypotheses is quite common. However, from the power perspective in field trials with extreme small sample sizes, e.g. $n_i = 4$, the increase of power, and hence the decrease of false negative rate, is substantial when using **one-sided tests or one-sided confidence limits**
- Most endpoints reveal a direction of harmfulness, e.g. reduction of a vitamin, reduction of non-target larvae
- Therefore, one-sided hypotheses will be used primarily in the proof of safety, i.e. testing non-inferiority. Alternatively, for two-sided hypotheses, tests on equivalence will be described
- Notice, the terms *non-inferiority* and *superiority* come from primary clinical endpoints and are *reversely* used here in safety assessment
- Still, some controversy on this topic

Proof of safety I

- Proof of safety is appropriate because of direct control of the more important false negative rate, i.e. consumer's risk. Formulated for toxicology by Kirkland (1999) *be confident in negative results* (Lovell et al. 2000)
- The hypotheses on equivalence (2-sided) or non-inferiority (1-sided) need the a priori definition of an acceptance threshold δ for the difference to isogenic $\mu_{GMO} - \mu_{iso}$ respective θ for the ratio to isogenic μ_{GMO}/μ_{iso}. Particularly in the multiple endpoint problem described above the availability of δ_i (respective θ_i); this is unrealistic
- More realistic is the estimation of confidence limits and their related post-hoc interpretation: **what is still acceptable?**
- This is a heuristic approach only

Proof of safety II

- **Question:** Why is the interpretation of the acceptance threshold θ for the **ratio-to-isogenic** μ_{GMO}/μ_{iso} more appropriate compared with those for difference-to-isogenic (δ):
  i) the direct comparison of differently scaled multiple endpoints is possible
  ii) % change is easy to understand
- Notice problems: additive vs. multiplicative model, instability when mean in the isogenic control is low (given s_i, n_i), etc.
- **Approach I:** Claiming local safety by independent analysis of each endpoint, each at level α. I.e. the comparisonwise error rate will be controlled only
- **Approach II:** Claiming global safety (objective here)
  y_1 AND y_2 AND...AND y_k are safe.
  This is an IUT, hence each elementary test is performed at level α

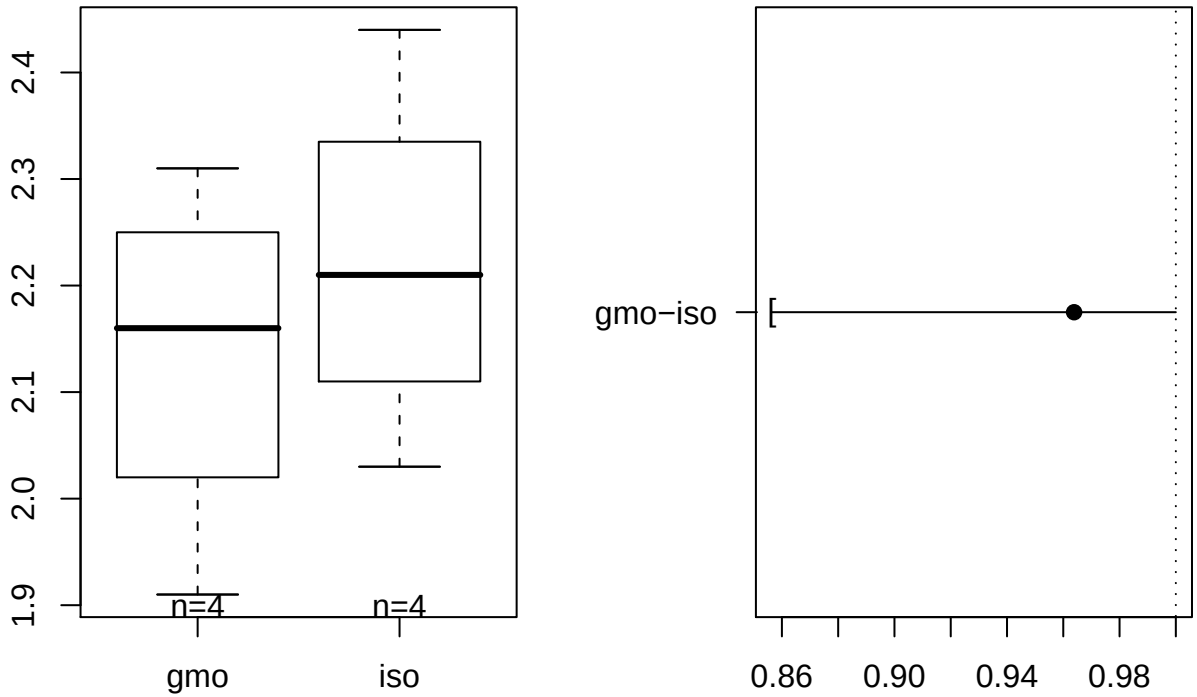
Proof of safety for multiple endpoints

- The outcome of global safety of hundreds of different endpoints is not likely in real field trials:
  i) from a practical point of view,
  ii) from the characteristics of the IUT: with increasing k the IUT becomes seriously conservative, remember $k > 100$
- Therefore, the need of the proof of safety for **a subset of** q **endpoints** is obviously $(q < k)$
- Consider an endpoint (vitamin) where harmfulness means *decrease* in GMO relative to isogenic.
  In the single endpoint case, non-inferiority, i.e. harmlessness, can be claimed when the lower confidence limit for μ_{GMO}/μ_{iso} is above an acceptance threshold (less than 1) say $\theta = 50\%$ or $\theta = 75\%$.
  Otherwise, this endpoint is inferior, i.e. harmful

Proof of safety: simple single endpoint case I

- **An example**: the compositional component Phytic Acid in one location of a field trial on genetically modified oilseed rape seeds (Hothorn and Oberdoerfer, 2006)
- Because different distributed multiple endpoints, non-parametric approaches will be used.
  Related two-sample confidence intervals are available in the R-library pairwiseCI (Schaarschmidt, 2007)

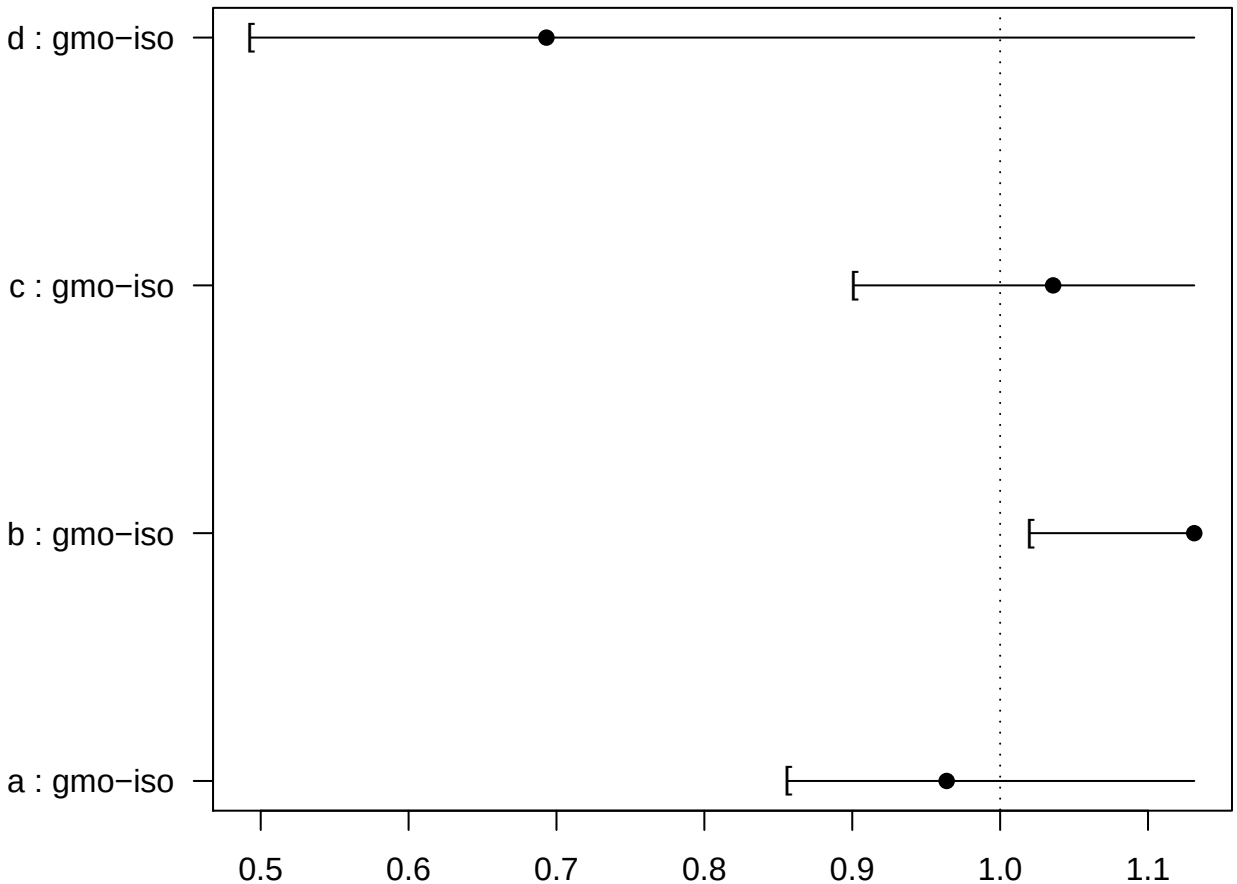
Proof of safety: simple single endpoint case II



- Interpretation: Only if we accept about a 86% decrease as still tolerable, harmlessness can be concluded; otherwise Phytic acid is harmfully reduced.

Proof of safety: simple single endpoint case III

- Four virtual scenarios:



Proof of safety: simple single endpoint case IV

- Interpretation:

- Scenario d: harmfull (GMO inferior vs. isogenic) because $\theta < 50\%$ too small to accept
- Scenarios a and c: harmless (GMO non-inferior vs. isogenic) because lower limits large enough, e.g. above $\theta = 80\%$, irrespective whether the point estimator is above 1 or not
- Scenario b: harmless (GMO superior vs. isogenic) because lower limit even above 1

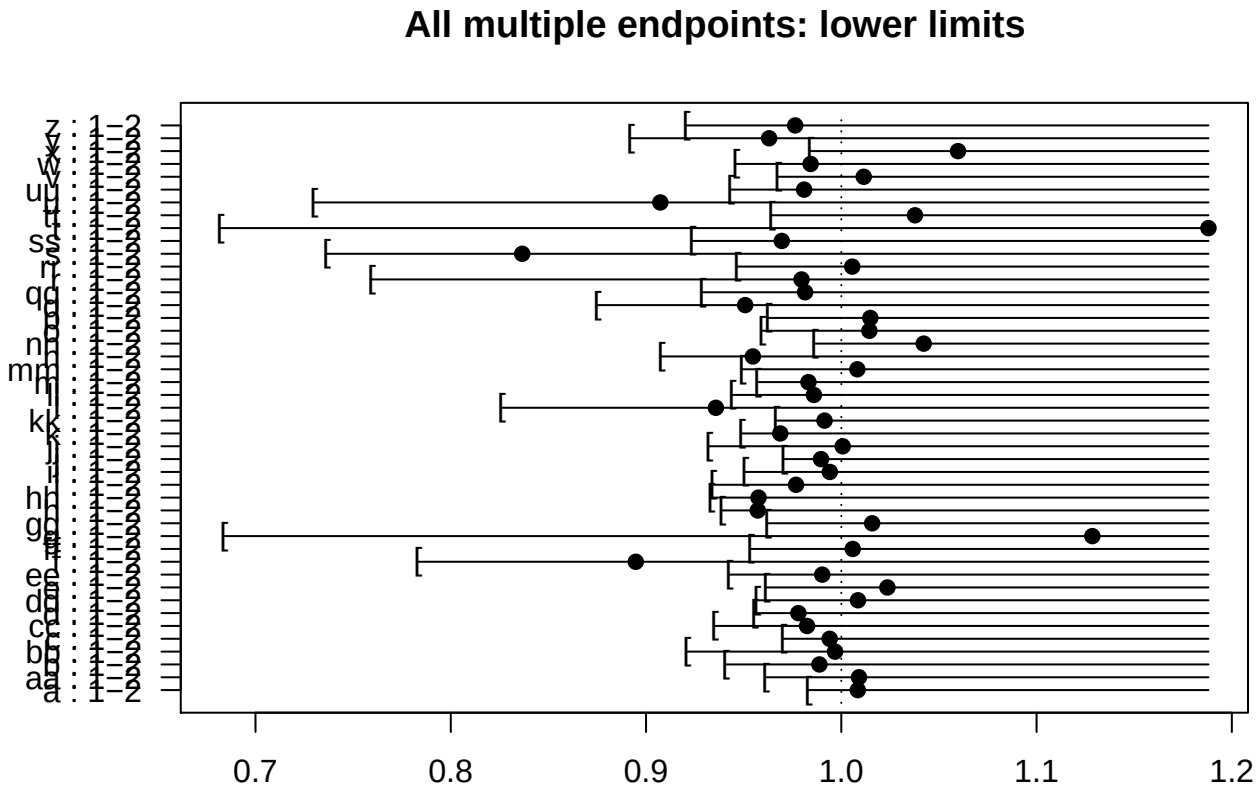
Proof of safety for subsets of multiple endpoints I

- In the line of the Quan et al. (2001) approach for three clinical endpoints, according to Hasler and Hothorn (2007) follows:
  i) **In a first step**, calculate the $(1-\alpha)$ lower confidence limits for all k endpoints. If each limit is above $\theta = 50....??\%$CI, all endpoints are at least non-inferior and harmless. The procedure stops with the claim of global safety for all endpoints.
  If not, all endpoints failing this demand - say j - are not at least non-inferior and hence, harmful.
  ii) The remaining $(p-j)$ not decided endpoints are taken for next step.
  Calculate $(1-\alpha/(j+1))$ lower confidence limits
  iii) etc.
  This procedure ends with not later than the p-th step where the possibly last undecided endpoint comes to a conclusion using a $(1-\alpha/k)$

An example I

- **An example**: all non-equal and above-detection-limit compositional components in one location of a field trial on genetically modified oilseed rape seeds (Hothorn and Oberdoerfer, 2006)

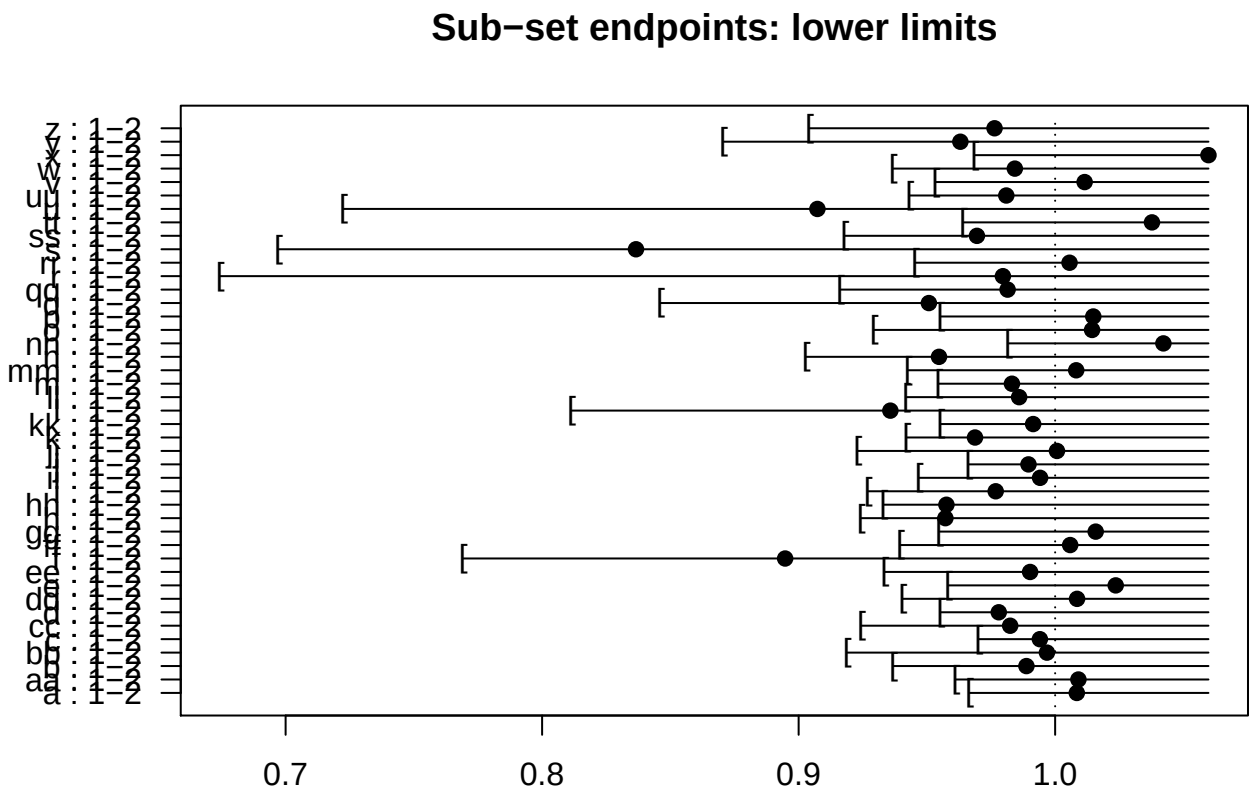
An example II



- Interpretation: Assume we accept a decrease to 70% as tolerable, two endpoints are not harmless: g and t

An example III

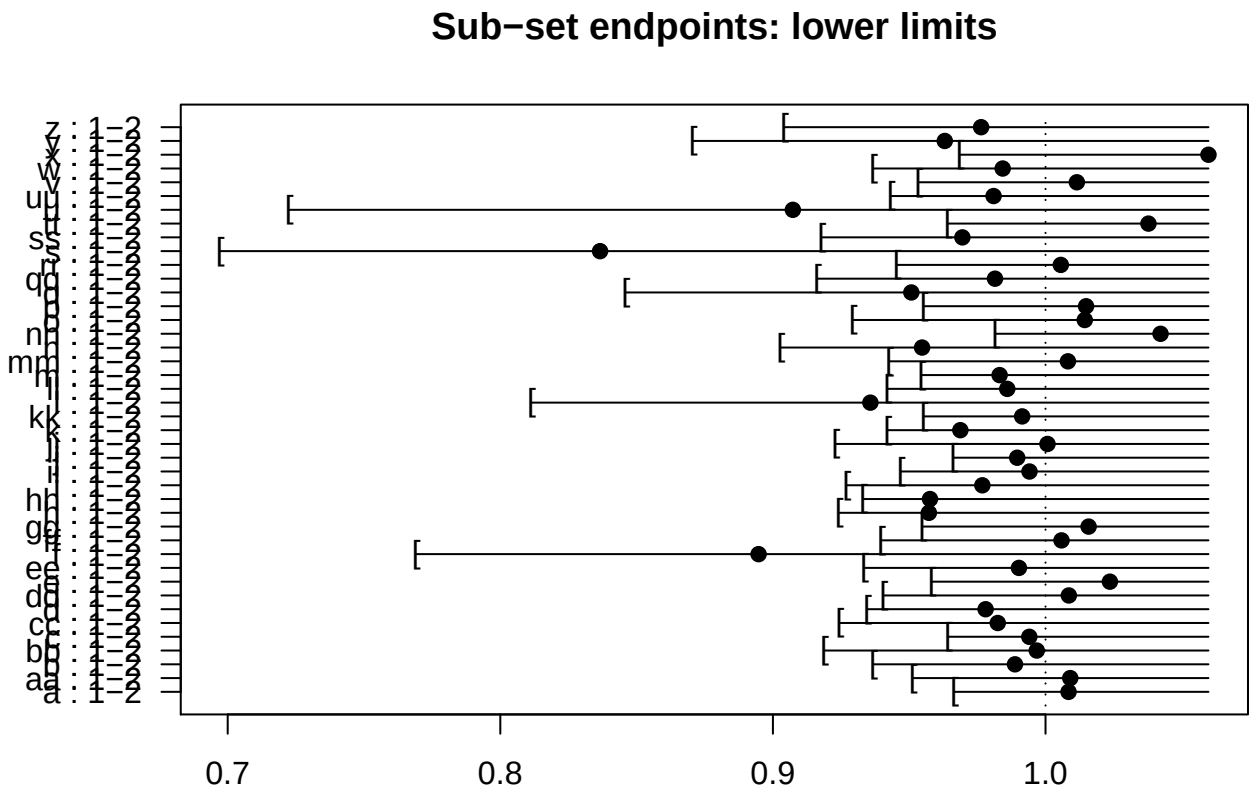
- **Second step:** estimate IUT-$(1-\alpha/(2+1))$ confidence intervals for the remaining $k-j=45-2=43$ endpoints



- A further endpoint, namely r is not harmless

An example IV

- **Third step:** estimate IUT-$(1 - \alpha/(3 + 1))$ confidence intervals for the remaining $k - j = 45 - 3 = 42$ endpoints



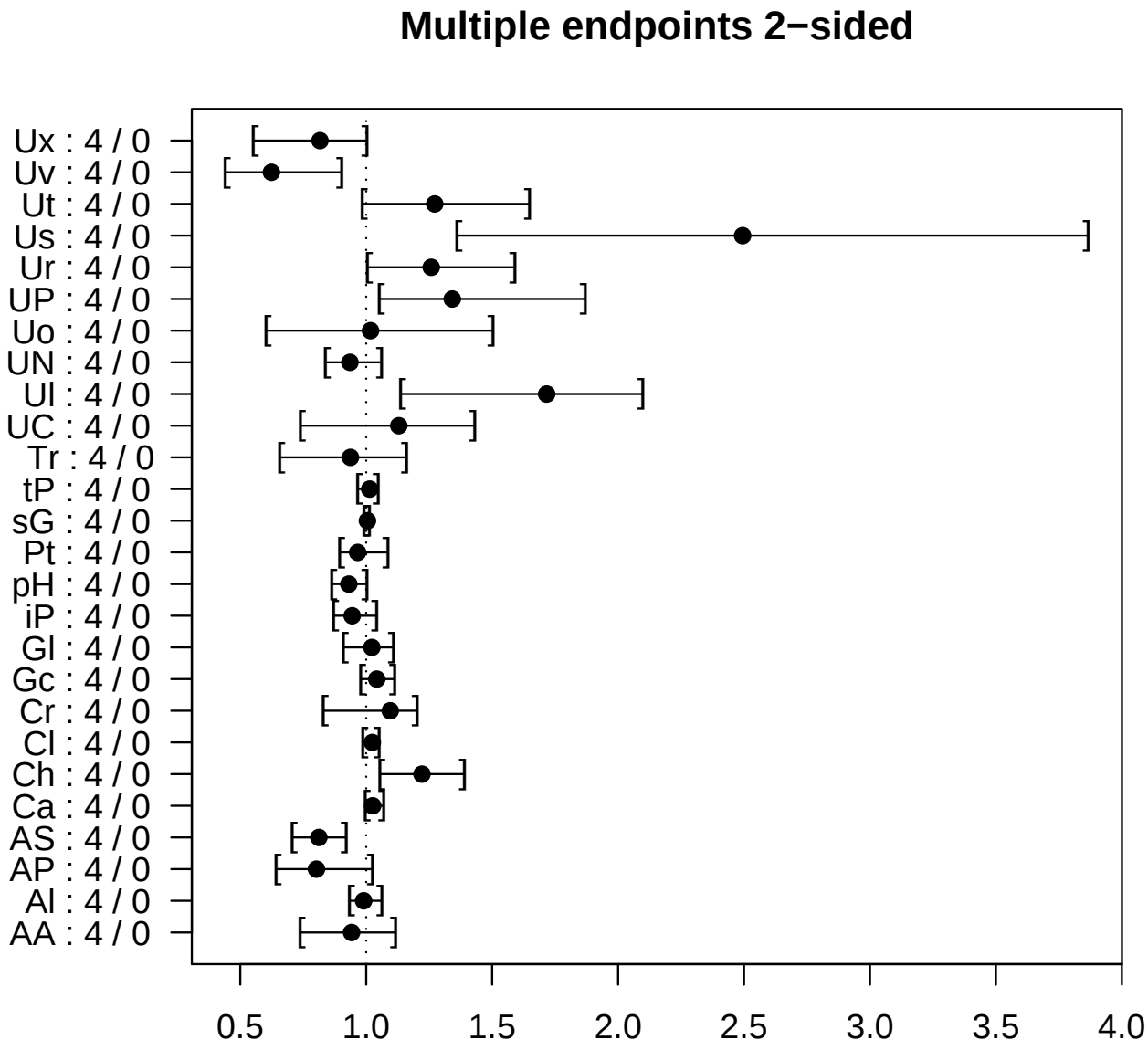
An example V

- After three steps, 42 of 45 endpoints are harmless given a 70% acceptance threshold
- This example shows **how difficult a proof of safety for all or a sub-set of many endpoints is**
- Notice, this is a teaching example using $n_i = 4$
  Real field trials include several locations and years, and the mixed model confidence limits are commonly shorter

Proof of safety for two-sided hypotheses I

- Toxicological feeding studies include endpoints where no specified direction of harmlessness can be defined, e.g. organ weights. Therefore two-sided hypotheses are needed
- **Equivalence hypotheses**
  $H_0^1 : \mu_{GMO}/\mu_{iso} < 1/\theta$ OR $H_0^2 : \mu_{GMO}/\mu_{iso} > \theta$
  $H_A^1 : \mu_{GMO}/\mu_{iso} > 1/\theta$ AND $H_A^2 : \mu_{GMO}/\mu_{iso} < \theta$
  Each individual test at level α, TOST
  or $(1 - 2\alpha)$ two-sided confidence interval
- Multiple endpoint equivalence problem: **IUT(IUT) test**
- An example: High dose and control for females data from a 90 days feeding study (EFSA stats working group, 2007)

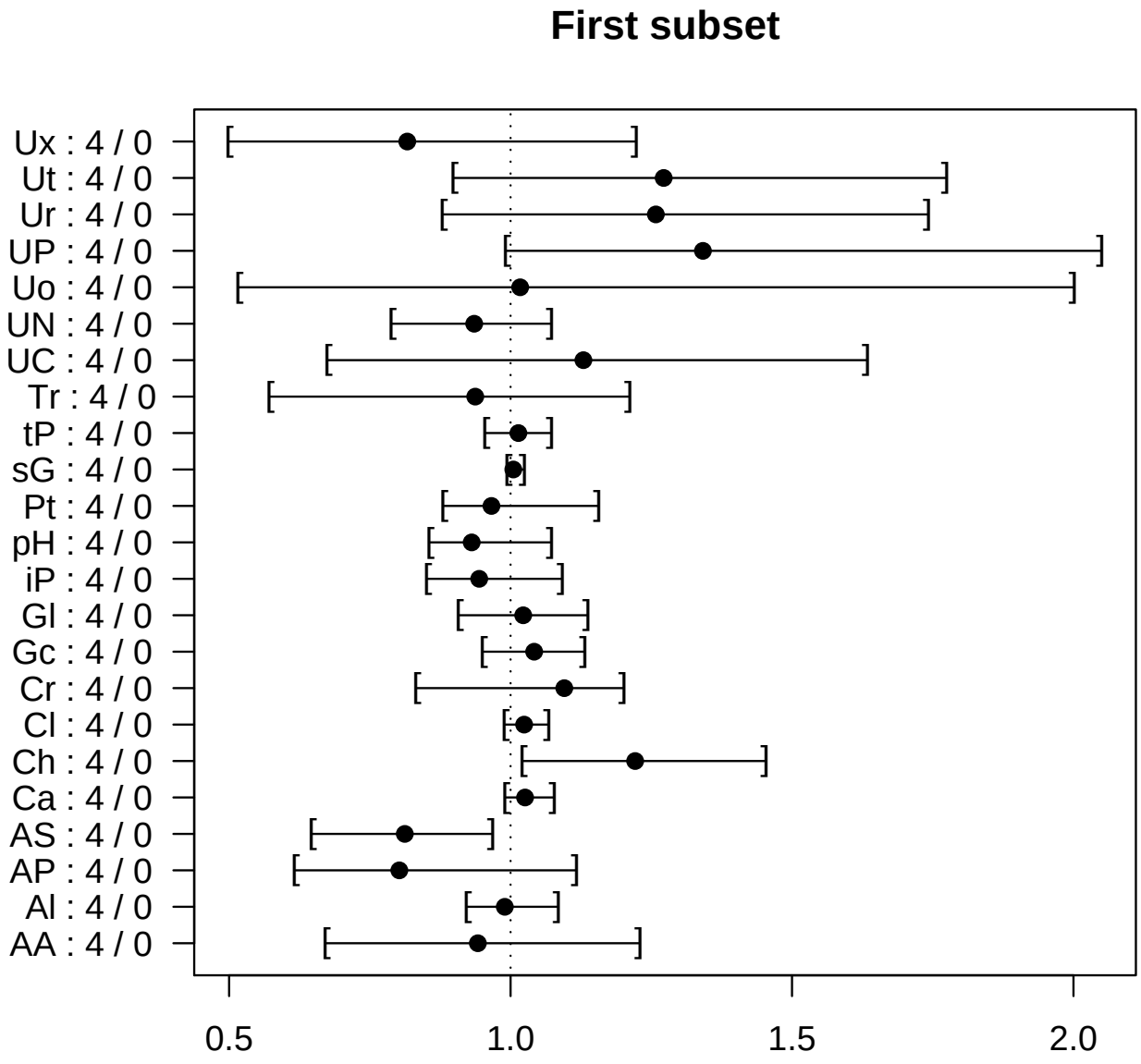
Proof of safety for two-sided hypotheses II



The second step I

- Looking at the estimated intervals we (a toxicologist and a biostatistician) may define:
  2-fold change is for these endpoints still acceptable
  Interpretation, i.e. equivalence region $[1/2; 2]$
- $k = 24$ endpoints are equivalent at this stage, but three endpoints: Us, UI, Uv are not
- **Second step:** estimate IUT-$(1 - \alpha/(3 + 1))$ confidence intervals for the remaining $k - j = 21$ endpoints

The second step II



Conclusions

- Proof of safety is a challenge for field trials with multiple endpoints
- A first proposal is presented here and demonstrated by several real data examples using available R libraries
- Many problems are unsolved up to now,
  e.g. definition of power, mixed model evaluation, small sample size problems, correlation between the endpoints
- Despite of all problems: the non-significance of a common t-test as a criterion for harmlessness
  for each individual endpoint should be avoided a.s.a.p.
- Related problems exist in long-term toxicity studies and safety endpoints in randomized clinical trials
- Finally: safety assessment using appropriate statistical methods is a relevant topic for further research