Proof of safety for a novel food-multiple endpoints approach

Ludwig A. Hothorn

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

Agrostat 2016

The problem I

- Approval of a novel food requires both a statistical proof of efficacy and **a proof of safety for possible side effects**
- Two examples to demonstrate harmlessness of new GMO: Non-target species (lepidopteran larvae) in field trials of Bt-corn (Gathmann2006)

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

The problem II

Nutritional components of oilseed rape seeds (Hothorn and Oberdoerfer 2006)

Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Regulatory Toxicology and Pharmacology xxx (2005) xxx–xxx
Regulatory Toxicology and Pharmacology
www.elsevier.com/locate/yrtph

Statistical analysis used in the nutritional assessment of novel food using the proof of safety

Ludwig A. Hothorn, Regina Oberdoerfer *

University of Hannover, LG Biostatistics, Herrenhaeuser Str. 2, D-30419 Hannover, Germany
Bayer CropScience GmbH, Industriepark Höchst, K607, D-65926 Frankfurt/Main, Germany

Received 28 July 2005

Abstract

The safety assessment of Novel Food, including GM biotechnology-derived crops, starts with the comparison of the Novel Food with a traditional counterpart that is generally accepted as safe based on a history of human food use. Substantial equivalence is established if no meaningful difference from the conventional counterpart was found, leading to the conclusion that the Novel Food is as safe and nutritious as its traditional counterpart. In general, the non-significance of p value is used for the proof of safety. From a statistical perspective, the problems connected with such an approach are demonstrated, namely that quite different component-specific false negative error rates result. As an alternative, the proof of safety is discussed with the inherently related definition of safety thresholds. Moreover, parametric and non-parametric confidence intervals for the difference and the ratio to control (conventional line) are described in detail. Finally, the treatment of multiple components for a global proof of safety is explained.

- Rather different aims, but similar statistical method

The problem III

- Classification from a statistical perspective
  1. Multiple secondary endpoints:
     i) a priori planned,
     ii) solicited
  2. Different scaled endpoints
  3. Aim: claiming similarity for almost all side effects
  4. Similarity $\Rightarrow$ equivalence
  5. Difficulties with statistical claim of equivalence for a single endpoint:
     i) choice of δ, ii) power π, n_i, iii) effect size
  6. Even more difficult with **multivariate equivalence**
- Complex designs including blocks, locations, years (random factors!).
  For simplicity here considering a completely randomized one-way layout with the two treatments GMO and near-isogenic variety.

The problem IV

- 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
- A multiple endpoint problem exists: **hundreds of species** (including both sexes and development stages) or **compositional components** will be observed/ measured: $y_1, ..., y_k$.
- Why (the commonly-used) non-significance of a point-zero hypothesis test, such as t-test, is inappropriate?
  Simply: Absence of evidence is not evidence of absence (Altman and Bland, 2004)

The problem V

- 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.
  Objective: formulate a proof of safety approach for multiple endpoints

Claiming equivalence for a single endpoint I

- $\Rightarrow$ TOST (two-one-sided-tests). IUT $\Rightarrow$ both tests significant: a lower test **and** an upper test. Needs a-priori definition of δ
- OR: inclusion within a $(1 - 2\alpha)$ confidence interval. Allows post-hoc definition of what is still acceptable
- Serious mis-use of [0.8; 1.25] thresholds from AUC in common drugs (bioequivalence FDA-rule) for other (therapeutic) equivalence problems
- Even more extreme: in most cases δ is unknown. CI-inclusion approach can be seen as transformation of the test problem into an δ threshold problem: *what can be tolerate as acceptable non-similarity?*
- Notice: NHST p-value is a transformation into a probability of Poppers falsification approach
- But δ is needed to calculated n_i in advance (power approach).

Claiming equivalence for a single endpoint II

- Example: TOST for Sasabuchi test (homogeneous variances) at $H_0 = 1$ and $CV_{iso} = 0.25$

```
library(PowerTOST)
pilowvar<-power.RatioF(alpha = 0.05, theta1 = 0.5, theta2=2, theta0 = 1, CV=0.125, n=6,
pihighvar<-power.RatioF(alpha = 0.05, theta1 = 0.5, theta2=2, theta0 = 1, CV=0.25, n=6,
pilowvar08<-power.RatioF(alpha = 0.05, theta1 = 0.8, theta2=1.25, theta0 = 1, CV=0.125,
pihighvar08<-power.RatioF(alpha = 0.05, theta1 = 0.8, theta2=1.25, theta0 = 1, CV=0.25,
```

[1] 0.999 0.654 0.268 0.019

- Equivalence approach without δ using arbitrarily n_i is an insoluble problem

Claiming equivalence for a single endpoint III

- FDA thresholds $[0.8; 1.25]$ are defined for a multiplicative model (whereas majority of efficacy testing uses an additive model). Advantage: dimensionlessness.
- Common approach: log-transformed data $\Rightarrow$ t-test interval $\Rightarrow$ backtransformation- works only if data a log-normal distributed with homogeneous variances. Otherwise serious bias may occur. Alternatives: ratio-to-control tests and confidence intervals . Using library(mratios) (Dilba, 2004)
- Two-sided hypotheses common. But, 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
- Most endpoints reveal a direction of harmfulness, e.g. reduction of a vitamin, reduction of non-target larvae

Claiming equivalence for a single endpoint IV

- Therefore, **one-sided hypotheses** will be used primarily in the proof of safety, i.e. test on non-inferiority.
- Discussion: standarized vs. unstandarized test statistics. Considering Cohen's effect size

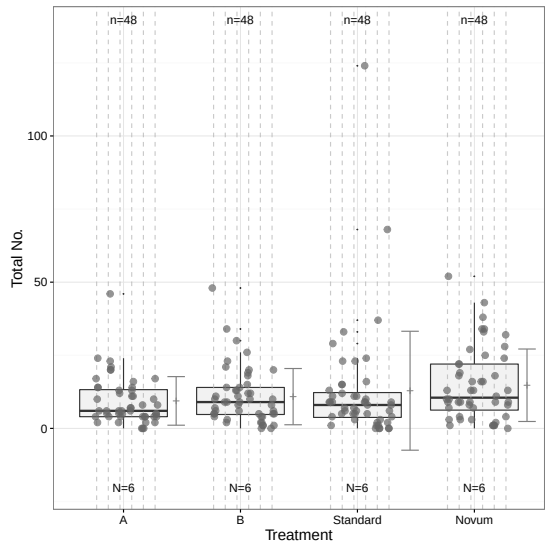


Abbildung : Boxplots total no. Nematocera

[1] -1.13 -0.30 0.68

Claiming equivalence for a single endpoint V

- Cohen vs. WinProb (effect size is probability, also relative effect size). Using internal WinProb R package [1] 0.22 0.45 0.69
- Test or confidence interval?
- Test on ratio-to-comparator vs. difference-to: i) ratio for lognormal endpoint, ii) ratio for nonparametric test, iii) impact of variance heterogeneity and values at detection limit
- Choice of delta: i) $f(\sigma)$ in field trials for genotype-by-variaty interaction EFSA (Vahl and Kang 2015)
  ii) without relation to variance for hazard consequence,
  iii) in principle asymmetric, e.g. vitamins,
  iv) Wellek's (1993) ϵ for Cohens effect size?
- Small sample size problem on power (Wellek table 6.2. p 104)

Claiming equivalence for a single endpoint VI

- **Summary: δ depends primarily of non-variance related consequence of hazard, for multiple endpoints hard to imagine. Solution marginal $(1 - 2\alpha)$ intervals for ratio-to-comperator with posthoc interpretation of the lower and upper limits, see below**

Claiming equivalence for a single endpoint VII

- Example rapseed (Oberdoerfer 2005)

Tabelle : CV for different endpoints.

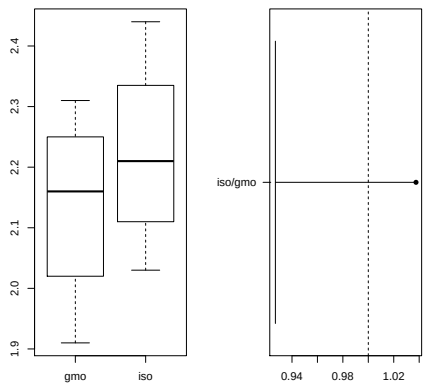
Endpoint	CV	My relevance score
Total fat	0.041	low
Eicosenoic acid	0.047	high
Arachidic acid	0.056	high
Ash	0.069	low
Linolenic acid	0.075	high
Cystine	0.12	very high
Protein	0.12	low
Behenic acid	0.12	very high
Alanine	0.13	very high
Arginine	0.13	very high
Aspartic acid	0.13	very high
Moisture	0.14	low
Total glucosinolate	0.24	very high

Claiming equivalence for a single endpoint VIII

- $f(\mathit{Variance})$ can not be recommended as choice of δ. But EFSA (van der Voet et al. 2007) proposed mixed effect model (locations, years, isogenic varieties)

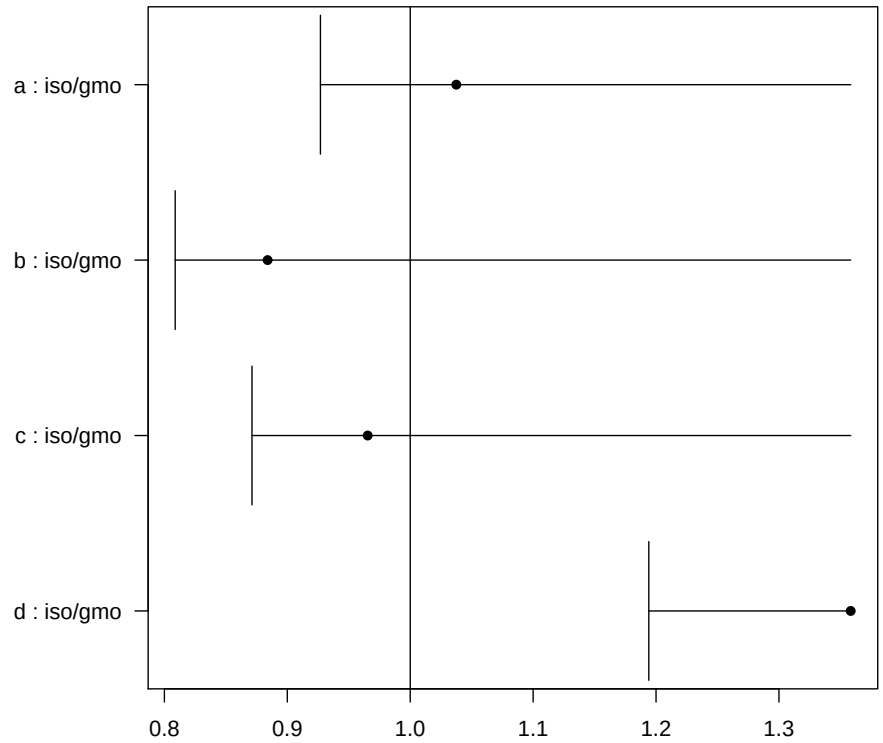
Example for a single endpoint I

- Phytic Acid in oilseed rape seeds. Low is critical!
- Nonparametric two-sample ratio-to-comparator confidence intervals (pairwiseCI)
- If we accept about a 86% decrease as still tolerable, a harmless conclusion is possible; otherwise Phytic acid is harmfull reduced.



Example for a single endpoint II

- 4 Scenarios



- Scenario b: harmfull (GMO inferior vs. isogenic) because $\theta < 75\%$ too small to accept

Example for a single endpoint III

- 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 d: harmless (GMO even superior vs. isogenic) because lower limit even above 1

Proof of safety for multiple endpoints I

- For the ratio to isogenic μ_{GMO}/μ_{iso}.
- 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) because 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)

Proof of safety for multiple endpoints II

- **Approach I:** Claiming local safety by independent analysis of each endpoint
- **Approach II:** Claiming global safety (more appropriate)
  y_1 AND y_2 AND...AND y_k are safe
  This is an IUT, hence each elementary test can be performed at level α (Hoffelder et al. 2015)

- A) Uncorrelated:
  eq $\Rightarrow CI^1_{lower} > \delta_{lower} ANDCI^2_{lower} > \delta_{lower} ANDCI^1_{upper} < \delta_{upper} ANDCI^2_{upper} > \delta_{upper}$
- Is an IUT(IUT).
- Univariate t-distributed $(1 - 2\alpha)$ intervals.
- B) Correlated: Bivariate t-distributed with $t_{2,R,2-sided,(1-2\alpha)}$
- Properties: with increasing p and/or increasing R the intervals become monotonic smaller that the marginal univariate intervals. Hard to accept: multivariate equivalent, but not univariate

Proof of safety for multiple endpoints III

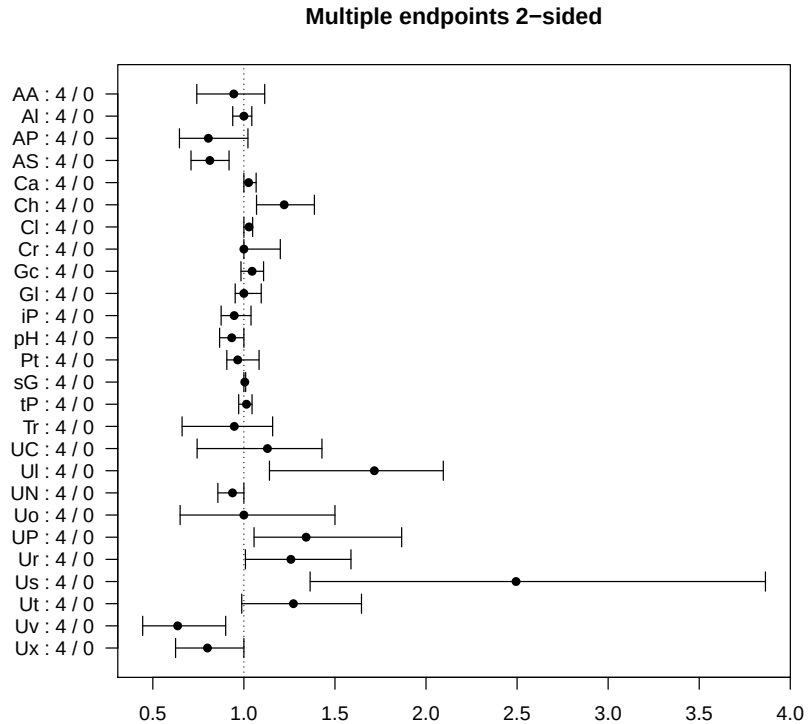
- Alternative UIT(IUT) (Hasler 2013)
- The outcome of global safety of hundreds of different endpoints is not likely in real field trials:
  i) from a practical point of view,
  ii) from the characteristic of the IUT: with increasing k the IUT becomes seriously conservative, remember $k > 100$

Proof of safety for subsets of multiple endpoints I

- Stepdown approach (Quan et al., 2001) approach for three clinical endpoints, according to Hasler and Hothorn (2007):
  i) **In a first step** , calculate the $(1-\alpha)$ upper 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 $(1-\alpha/(j+1))$ upper 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 $(1-\alpha/k)$.

Proof of safety for subsets of multiple endpoints II

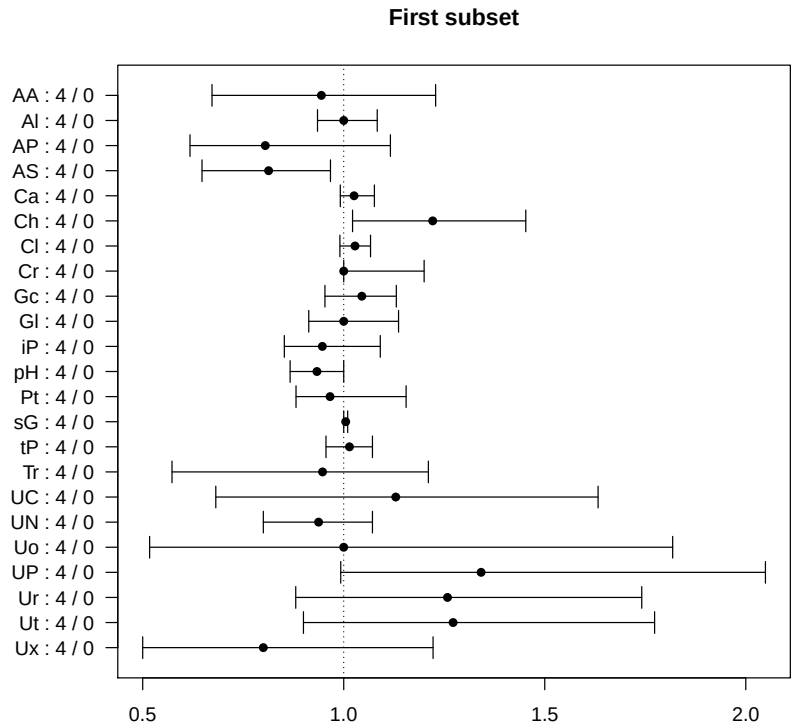
- An example: High dose and control for females data from a 90 days feeding study (EFSA stats working group, 2007)



Proof of safety for subsets of multiple endpoints III

- 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

Proof of safety for subsets of multiple endpoints IV



- These 21 endpoints are equivalent

Conclusions I

- Proof of safety is a serious challenge for novel food trials with multiple endpoints
- Neither a relevance-related endpoint-specific choice of δ_i is available, nor a power approach (to determine n_i).
  Therefore: a global (or partial) **test** on safety is unrealistic
- Despite of all problems: the non-significance of a common t-test as a criterion for harmlessness
  for each individual endpoint should be avoided at all

Conclusions II

- Marginal two-sided $(1 - 2\alpha)$ confidence intervals for ratio-to-comparator (corrected for variance heterogeneity) and their post-hoc endpoint-specific interpretation can be recommended

- A stepwise approach is possible to identify the majority of endpoints as safe, only some as inferior, but δ_i needed

- Requirement for certain n_i (not as small as $n_i = 4$)

- Related problems exist in long-term toxicity studies and safety assessment in randomized clinical trials
- Written in `knitr`