Proof of safety for a novel food – multiple endpoints

Ludwig A. Hothorn[1]

[1] Leibniz University Hannover, D-30419 Hannover , Germany
E-mail : hothorn@*biostat.uni-hannover.de*

Abstract
Statistical proof of safety for a novel food in a field trial or feeding study can be translated into claiming equivalence. Even for a single primary endpoint such an equivalence test is not simple, but for the common multiple endpoints a challenge exists. The main problem is the non-availability of the many endpoint-specific equivalence-threshold margins. Their estimation from the variability between standard varieties in multiple environments is inappropriate. A marginal ratio-to-standard confidence interval approach for multiple endpoints with post-hoc interpretation is proposed as an alternative.
Keywords: Proof of safety, multiple endpoints, confidence intervals for equivalence

1. Introduction

Approval of a novel food requires both a statistical proof of efficacy for the primary efficacy endpoint and a proof of safety for possible side effects (a priori planned or solicited), e.g. similar abundances of non-target species (e.g. lepidopteran larvae) in field trials of Bt-corn (Gathmann et al. 2006) or similarity of many nutritional components of oilseed rape seeds in a feeding study (Hothorn and Oberdoerfer 2006). From a statistical perspective, similarity is formulated as claiming equivalence. The problem is characterized by i) many multiple endpoints, ii) different scaled endpoints, iii) the aim to claim equivalence for almost all side effects, and iv) unknown endpoint-specific equivalence margins. Already for a single endpoint problems exist, but for multiple endpoints a challenge exists-without a simple solution up to now.

2. Claiming equivalence

2.1 A single endpoint

Rather complex designs are used for the above trials including blocks, locations, and years (random factors). For simplicity here considering a completely randomized one-way layout with the two treatments GMO and near-isogenic variety. The common approach is two-one-sided-tests (TOST) which requires the a priori definition of the equivalence thresholds. For the multiplicative model frequently the [0.8, 1.25] thresholds are used, but they were defined for bioequivalence only (i.e. area-under-the curve for the serum concentration of a drug) by the US Food and Drug Administration. The use of the variability between standard varieties in different environments (and their interaction) (Acutis et al. ,2006 ; van der Voet et al., 2011 ; Vahl and Kang, 2015) is inplausible, however. Imagine a feeding study with the two endpoints body weight and liver transaminase. The first can be measured very precisely (e.g. CV 1%), the second rather unprecisely (e.g. CV 100%). Would this mean that we can only tolerate [0.99 ; 1.01] for body weight, but we must tolerate [0.2 ; 5.0] for the

liver-transaminase as still similar - certainly not ! The crux for multiple endpoints in the above trials are their rather different variances (CV's)- not nessesarily correlated with their tolerability.
TOST can be re-formulated as inclusion into a two-sided (1-2alpha) interval. Therefore, the estimated confidence limits can be used for a post-hoc interpretation of endpoint-specific tolerability- without a priori definition of the thresholds. However, because of the strong dependence between [power, alpha, variance, margin and sample sizes], endpoint-specific different conditions occur in a trial with the inherently same sample size for all multiple endpoints. This problem can not be solved (and was not solved for multiple tumors in animal long-term carinogenicity assays), but power calculation is available for some tests. By means of endpoint-specific power approach, at least the consequences of the different conditions per endpoint can be quantified.
The next problem is the definition of an appropriate effect size. Whereas for claiming superiority the difference of expected means seems the quasi-standard, for the equivalence intervals the assumption of log-normal distribution is common ; but see the serious counter-arguments (Schaarschmidt, 2013). Instead of unstandardized, standardized effect sizes can be used (Wellek, 1993) with their relationship to Cohen's effect size and the win probability (allowing an individual definition of equivalence).
An alternative is the non-parametric ratio-to-standard intervals proposed in the talk and available in the R library(pairwiseCI) (Schaarschmidt, 2016).

2.2 Multiple endpoints

Using the intersection-union test principle (IUT), a global equivalence test for multiple endpoints can be formulated. This works also with two-sided simultaneous confidence intervals. However, with increasing number of endpoints this approach is extreme conservative and ignores the (high) correlations between the endpoints (Hoffelder et al., 2015). A stepwise version can be used for almost all endpoints which represents and UIT-IUT approach (Hasler and Hothorn, 2013) which is available in the R-library (MultEq). Considering the large number of endpoints, their un-designed nature, their rather different precisions, the screening character of the trials, marginal (1-2alpha) confidence intervals for ratio-to-standard with post-hoc interpretation of their endpoint-specific tolerability are proposed in the talk. This approach will be demonstrated by real data for a feeding study using R libraries.

References

Acutis, M, Hothorn,L McNicol J van der Voet, H (2009) Statistical considerations for the safety evaluation of GMOs. *EFSA Journal* 1250: 1

Gathmann, A; Wirooks, L; Hothorn, LA; Bartsch, Schuphan, I. (2006) Impact of Bt maize pollen (MON810) on lepidopteran larvae living on accompanying weeds. *Molecular Ecology* 15: 2677-2685.

Hasler M, Hothorn LA (2013). Simultaneous confidence intervals on multivariate non-inferiority. *Statistics in Medicine* 32(10), 1720-1729.

Hoffelder, Th., Gössl R. and Wellek S. (2015) Multivariate Equivalence Tests for Use in Pharmaceutical Development, *J. Biopharm. Stat.*, 25:3, 417-437.

Hothorn, LA and Hasler, M. (2008). Proof of hazard and proof of safety in toxicological studies using simultaneous confidence intervals for differences and ratios to control. *J. Biopharm Stat.* 18: 915-933.

Hothorn, LA and Oberdörfer, R (2006). Statistical analysis used in the nutritional assessment of novel food using the proof of safety. *Regul Toxicol. Pharmacol* 44 : 125-135

Schaarschmidt F (2013). Simultaneous confidence intervals for multiple comparisons among expected values of log-normal variables. *Comput. Statistics and Data Analysis*, 58:265-275.

Vahl CI and Kang Q. (2015) Equivalence criteria for the safety evaluation of a genetically modified crop: a statistical perspective *J Agri Sci* doi:10.1017/S0021859615000271

Wellek S (1993) Basing the analysis of comparative bioavailability trials on an individualized statistical definition of equivalence. *Biom J.* 35:1; 47-55.

Van der Voet, H., Perry et al. (2011). A statistical assessment of differences and equivalences between genetically modified and reference plant varieties. *BMC Biotechnology* 11, 15.