# Scientific Opinion on Statistical considerations for the safety evaluation of GMOs

EFSA Panel on Genetically Modified Organisms (GMO)

**EFSA Publication; Sørensen, Ilona Kryspin**

*Publication date:*
2010

*Document Version*
Publisher's PDF, also known as Version of record

**SCIENTIFIC OPINION**

# Scientific Opinion on

# Statistical considerations for the safety evaluation of GMOs [1]


## EFSA Panel on Genetically Modified Organisms (GMO)[2, 3]

European Food Safety Authority (EFSA), Parma, Italy



## ABSTRACT

This opinion proposes: 1) updated statistical guidelines and possible approaches for the analysis of compositional, agronomic and phenotypic data from field trials carried out for the risk assessment of GM plants and derived foods/feeds; 2) minimum requirements that should be met in the experimental design of field trials, such as the inclusion of commercial varieties, in order to ensure sufficient statistical power and reliable estimation of natural variability. A graphical representation is proposed to allow the comparison of the GMO, its conventional counterpart and the commercial varieties with respect to many variables, taking into account natural variability. It is recommended to quantify natural variability from data on non-GM commercial varieties treated in the same way and in the same experiments as the GM and the conventional counterpart test materials. Only when such estimates are unavailable may they be estimated from databases or literature. Estimated natural variability should be used to specify equivalence limits to test the difference between the GMO and the commercial varieties. Adjustments to these equivalence limits allow a simple graphical representation so that a single pair of confidence limits may be used to display statistically significant differences and to visually assess equivalence. The possible types of outcome of this graphical representation are described and a proposal is made when further evaluation should be performed. In addition to providing specific recommendations for the interpretation of compositional analysis, this opinion highlights some statistical issues of a more challenging nature, such as the simultaneous assessment of many characteristics (i.e. multivariate analysis), which will require further research. The principles proposed in this opinion may be used, in certain cases, for the evaluation of GMOs other than plants.



## KEY WORDS

GMO, equivalence limits, field trials, compositional analysis, mixed model, proof of hazard, proof of safety, confidence interval, difference test, equivalence test.


---

1 On request of EFSA, Question No EFSA-Q-2006-080; originally adopted on 21 April 2009, updated on 2 December 2009. This scientific output, published on 1 February 2010, replaces the earlier version published on 31 July 2009.

2 Panel Members: Hans Christer Andersson, Salvatore Arpaia, Detlef Bartsch, Josep Casacuberta, Howard Davies, Lieve Herman, Patrick Du Jardin, Niels Hendriksen, Sirpa Kärenlampi, Jozsef Kiss, Gijs Kleter, Ilona Kryspin-Sørensen, Harry Kuiper, Ingolf Nes, Nickolas Panopoulos, Joe Perry, Annette Pöting, Joachim Schiemann, Willem Seinen, Jeremy Sweet and Jean-Michel Wal. Opinion is shared by all members of the Panel. 0 Panel member with minority opinion. 0 members of the Panel did not participate in the discussion on the subject referred to above because of potential conflicts of interest identified in accordance with the EFSA policy on declarations of interests. Correspondence: GMO@efsa.europa.eu


3 Acknowledgements: The Panel wishes to thank the members of the Working Group for the preparation of this opinion: Marco Acutis, Ludwig Hothorn, Jim McNicol, Hilko van der Voet and EFSA's staff members Claudia Paoletti and Billy Amzal for the support provided to this EFSA scientific output.

## SUMMARY

The European Food Safety Authority (EFSA) asked its Panel on Genetically Modified Organisms (GMO) to investigate whether more detailed guidance could be provided regarding the performance of field trials and the analysis of data using appropriate statistical models, with the objective of ensuring a more uniform approach and greater transparency in risk assessment of GMOs. In order to carry out this investigation, the GMO Panel has convened a dedicated statistics Working Group who addressed the issue. A draft document was published on EFSA website from 21 July 2008 until 21 September 2008 for a 2-month period of public consultation. At the deadline EFSA had received 98 submissions, from 9 stakeholders. The table of all received comments together with a summarized response to the most relevant ones is published on the EFSA web site: *http://www.efsa.europa.eu*. Following the public consultation, the draft document was revised taking into account all the scientific comments that helped enhancing scientific quality and clarity (opinion adopted on 21 April 2009). Subsequently the adopted opinion was identified to have an incomplete statistical formula used to calculate equivalence limits. Further to adjusting the statistical formula the order of the various sections of the opinion was revised to improve logic flow. The present opinion was amended by the GMO Panel on 2 December 2009.

## TABLE OF CONTENTS

## BACKGROUND

In line with other international guidelines (WHO, 1995; Codex, 2003), the GMO Panel has adopted a strategy for the risk assessment of genetically modified organisms (GMO) based on the comparison of the GMOs and their derived products with the respective appropriate conventional counterpart(s) (Regulation (EC) 1829/2003; EFSA Guidance Document, 2006). The underlying assumption of this comparative assessment approach for GM plants is that traditionally cultivated crops have gained a history of safe use for consumers and animals, and familiarity for the environment. Although general principles for risk assessment are discussed in the Guidance Document (EFSA, 2006), with reference to existing internationally agreed test methods and protocols, detailed protocols for carrying out specific experiments are not provided.

With respect to the comparative assessment of GM plants and derived foods/feeds, the Guidance Document (Section III, D7) describes the criteria for choosing an appropriate conventional counterpart and for performing appropriate field trials, i.e. number of sites, growing seasons, geographical spread, replicates, selection of compounds to be analysed etc. Moreover the Guidance Document recommends using appropriate statistical tools for the design of field trials and the analysis of data, but no clear indication is provided for the definition of appropriate statistical power and the interpretation of experiments' results.

An important issue to consider is how differences in composition, agronomy and phenotype between GM plants and their conventional counterparts should be identified and evaluated with respect to their potential impact on humans, animals and/or the environment. In the context of a GMO safety evaluation, it is desirable to assess observed differences against quantified natural variation. Natural variation is the variability occurring naturally because of differences in the genotypes of plants, effects of environmental factors and the interaction between them. Accurate estimation of natural variation is challenging and requires an extensive knowledge of the existing natural variation in compositional, agronomic and phenotypic parameters of plants.

Another important aspect is the evaluation of the results of animal studies, e.g. 90-day toxicity studies in rodents with whole foods/feed. Such studies are carried out on a case by case basis, as deemed necessary. Observed differences in values of biological test parameters between the GM plant derived food/feed and its (usually near-isogenic) conventional counterpart(s) should be assessed against the natural variation in these parameters. Natural variation may be influenced both by the genetic background of the test animals and the genetic background of the feed crops (which may influence animal endpoints through a changed diet composition), as well as by environmental factors (housing, feeding, test diets etc.). We emphasise that whilst the present opinion makes statistical proposals for both the experimental design and analysis of field trials of GM plants for compositional data, these do not relate to the design of animal studies, for which recent guidance has been issued separately (EFSA GMO Panel Working Group on Animal Feeding Trials, 2008). However, it is the case that the statistical approach outlined here for analysis might also be used for the analysis of data from animal feeding studies with whole GMO foods/feed, where appropriate and on a case-by-case basis, especially if these include commercial varieties with a history of safe use.

The experience of the GMO Panel gained from the evaluation of GMO applications under Directive 2001/18 (EC) and Regulation (EC) 1829/2003 since 2003, shows that applicants use widely differing protocols to carry out field trials and to analyse the collected data or to evaluate data from animal feeding trials. Moreover, different models for statistical analysis of the data have been used (e.g. Oberdoerfer *et al*. 2005, Hammond *et al*. 2006, Hothorn and Oberdoerfer 2006, Herman *et al*. 2007, McNaughton *et al*. 2007). Application of different statistical approaches and models may lead to different conclusions regarding the risk assessment of GM plants and derived foods/feeds.

Therefore EFSA and the GMO Panel were of the opinion that it would be worthwhile investigating whether more detailed guidance could be provided to applicants regarding the use of appropriate statistical models for the analysis of the data from field trials for compositional, agronomic and

phenotypic studies and animal feeding trials, and regarding the design of field trials. In the long term this should lead to a more uniform approach to be taken by applicants and risk assessors, which may contribute to a greater transparency in an accurate risk assessment of GMOs and a faster safety evaluation of GMO applications.

## SCOPE

The scope of this document is the identification of a strategy for better harmonization of approaches for data evaluation in GMO risk assessment and a more precise definition of experimental design requirements for field trials. Specifically, in order to provide guidance on these issues, the EFSA GMO Panel Statistics Working Group pursued the following main objectives:

1. To review statistical methods and possible approaches, including those applied by applicants, which could be appropriate in the framework of the comparative risk assessment of GM plants and derived foods/feeds. To explore univariate data analysis methods suitability with respect to reliability of conclusions, i.e. the probabilities of occurrence of false positives or false negatives. To make an initial assessment of the potential contribution of multivariate methods.

2. To identify possible strategies to incorporate natural variability of test parameters due to genetic and environmental causes. To investigate the suitability and possible application of both the equivalence and the difference testing approaches for the risk assessment of GM plants and derived foods/feeds.

3. To undertake a feasibility study regarding the applicability of proposed statistical tools using suitable data.

## LIMITATIONS

The GMO Panel is of the opinion that newly developed guidelines for statistical approaches in comparative assessment and GMO safety evaluation should be tested on example datasets. A practical example illustrating the proposed methodology on a real-case dataset is provided in this document. The GMO Panel emphasises that future scientific developments and the analysis of further datasets will inevitably lead to refinements in technique. Consequently, this guidance will be reviewed regularly, and the present document represents the first such review.

It was realised from the beginning that the task of developing guidelines for statistical approaches presented two different problem areas: firstly, the development of suitable approaches for single endpoints; and secondly, the development of suitable approaches for simultaneous statistical analysis of a large set of endpoints. It was agreed that the Working Group would first work on statistical approaches for single endpoints, while making an initial assessment of the problems connected with analysing multiple endpoints.

Assessment of statistical approaches for the analysis of data generated for use in environmental risk assessment was not included in the mandate of the self-tasking activity and is therefore not discussed in this opinion.

# 1. General Principles

## 1.1. Introduction

The objective of this opinion is to propose statistical methods and possible approaches regarding the comparative risk assessment of GM plants and derived foods/feeds. Results from any appropriate statistical analysis (e.g. confidence intervals, *p* values, etc.) will need further interpretation with respect to a possible impact on human/animal health, particularly because statistically significant results are not always biologically or toxicologically relevant.

In the comparative risk assessment a GMO is compared to an appropriate conventional counterpart or control organism/material. The comparison begins by measuring a number of specific agronomic, phenotypic, and compositional characteristics of the GM plant and/or derived foods/feed and of its non-GM counterpart. The main purpose of the comparative assessment is to demonstrate whether the GM plant and/or derived food/feed is different from its appropriate conventional counterpart and/or equivalent to commercial varieties, apart from the inserted trait(s).

Equivalence is in this context defined as the absence of differences other than ordinary biological variation and other than the expected differences due to intended modifications. For each chosen endpoint, or for groups of endpoints, limiting values for which the difference is acceptable, must be determined. These are known as equivalence limits. Statistical methods can be used to assess the observed differences against the natural variability observed between commercial varieties.

Considering each single measured characteristic (endpoint) three different assessments of GMOs may be of interest:

1. The GMO may be shown to be different from the conventional counterpart (proof of difference). A difference may constitute a hazard (potential risk) which should be subject to further safety evaluation (for this reason it is sometimes referred to as proof of hazard).

2. Theoretically and in principle, it is possible to establish equivalence limits prior to the comparative risk assessment. Then, the GMO may be shown to be within these equivalence limits (proof of equivalence). Equivalence limits may be derived in absolute terms, or as relative deviations from the conventional counterpart, or as relative deviations from the overall mean of commercial varieties. Established equivalence of a GMO has been interpreted as relevant for subsequent toxicological risk assessments.

3. However, in practice, it must be emphasised that equivalence limits have almost never been established. Therefore, commercial varieties are to be included in the experiments, to allow a direct comparison of the GMO with the commercial varieties. This may be seen as a test of the difference between GMO and commercial references, but it should also be recognised that the commercial varieties in the experiment allow the estimation of equivalence limits, which are subsequently used for assessing the equivalence of the GMO.

In the future, as databases for compositional data are extended and experience gained from further field trials, it may be come possible to establish equivalence limits as in point 2 above. Then, it should be noted that the approach followed for their calculation will affect the final interpretation of the results. If equivalence limits have been established as relative deviations from the conventional counterpart, the outcome of the equivalence test will establish equivalence between the GMO and its counterpart. If on the other hand, equivalence limits have been established as relative deviation from the mean of commercial varieties, the outcome of the equivalence test will establish equivalence between the GMO and the set of commercial varieties. To cover the possibility of both cases 2 and 3 above, in text where it is unspecified how equivalence limits have been established (whether the conventional counterpart or the set of commercial varieties is meant), the word ‘reference’ will be

used in this document to discuss matters that are applicable to both situations. Also, in the text below the term 'test materials' is used to denote: the GM crop; its conventional counterpart; the set of commercial varieties; and any additional test materials used because of the need for additional comparators.

Statistical methodology should not be focussed exclusively on either differences (1) or equivalences (2/3), but should provide a richer framework within which the conclusions of both types of assessment are allowed. Both approaches are complementary: statistically significant differences may point at biological changes caused by the genetic modification, but may not be relevant from the viewpoint of food safety. On the other hand, equivalence assessments may identify differences that are potentially larger than normal natural variation, but such cases may or may not be cases where there is an indication for true biological change caused by the genetic modification. A procedure combining both approaches can only aid the subsequent toxicological assessment following risk characterization of the statistical results.

Section 4.1 lists the possible types of outcome if both approaches are considered simultaneously. Briefly, there will be categories with a clear conclusion on equivalence, and categories where the statistics do not lead to an unambiguous result. This possibility of 'grey' outcomes between 'black' and 'white' outcomes is characteristic of any statistical approach, and such an outcome is an indicator of scientific uncertainty rather than failure of the method.

Stringent use of the concept of equivalence would require the necessity of proving equivalence for all endpoints simultaneously (global equivalence). Such a proof of global equivalence turns out to be technically difficult to undertake. While the provision of methodology for a global equivalence assessment might prove of some use in the long-term, the mandate of the Working Group specifies that only an initial assessment of such methodology is made. In this opinion the focus is on statistical methods applied to single endpoints.

As stated above, very few, if any equivalence limits for measurable endpoints have been established within the scientific literature. Therefore the statistical approach should be sufficiently flexible to address such situations, as will be discussed further in Section 3.3. For example, equivalence limits may be estimated from concurrent data on commercial varieties or other available information may be used where appropriate in the future. When there is serious uncertainty about appropriate equivalence limits, it may be useful to present results for several possible values of the equivalence limits.

### 1.2. Error types and statistical power

Equivalence testing contrasts with much of other biological experimentation: in the former the risk assessor seeks assurance that a hypothesis of equality of GMO and its conventional counterpart is approximately true, although strict equality can never be proven. By contrast, most biological experiments are designed to reveal and quantify differences between varieties and controls. In any test of a null hypothesis there are two possible types of errors, which are mutually exclusive. A so-called 'Type I' error occurs if the null hypothesis is erroneously rejected when it is actually true. A 'Type II' error occurs when the null hypothesis is not rejected even though it is actually untrue. In a traditional proof-of-difference approach the null hypothesis is taken to be equality of GMO and non-GM control, therefore not finding a true difference is a 'Type II' error. In a proof-of-equivalence approach the null hypothesis specifies the existence of a difference of a given magnitude, and concluding equivalence which actually does not exist is a 'Type I' error.

It is relatively simple for scientists to set the Type I error rate for an experiment, but it is much more difficult to estimate the Type II error rate accurately, l*et al*one set it to a desired value. Traditionally, in many experimental disciplines the Type I error rate, $\alpha$, sometimes called the size of the test, is set to $\alpha = 0.05$. Such tests, at the so-called '5% level' are conventionally considered as acceptable in risk assessment. However, if in safety testing we retain the traditional null hypothesis of zero difference, the Type II error (i.e. accepting that GMO and conventional counterpart yield equal responses when

there is in fact a difference) is the most serious and relevant one (e.g., Hill and Sendashonga, 2002). Clearly, poorly designed experiments, or those with inadequate replication, even though using a 5% Type I error rate, have such large Type II error that they lack the ability to discriminate between the GMO and its conventional counterpart. Ignoring Type II errors might lead to an erroneous indication of safety, while in reality the experiment simply was not sensitive enough to detect adverse effects. The complement of the probability of Type II error is termed 'statistical power'. Statistical power is therefore the probability of detecting a difference between GMO and its conventional counterpart, when there is a real difference of a certain size to detect; it is often quoted as a percentage. The risk assessor must ensure that an evaluation has sufficient power to provide reasonable evidence of equivalence. A level of 80% is usually considered to be the acceptable minimum degree of statistical power and optimal experimental design should be directed to attain this level. Statistical power depends, amongst other things, upon the chosen experimental design, the magnitude of the variety difference, the baseline variability of the experimental units, the size of the test and the replication of the experiment. In general, other things being equal, a decrease in $\alpha$ will generally lead to a decrease of power.

A power analysis, executed when the study is being planned and prior to its start may be used to estimate power, to choose appropriate replication and to give confidence that the experiment will detect any significant effect that is present. For example, Perry *et al*. (2003) reported a power analysis performed prior to an experiment to assess the risk of indirect effects on farmland wildlife of genetically modified herbicide-tolerant management systems of weed control, compared to current conventional farming.

A common approach to deal with Type II error in proof-of-difference tests, but one of dubious validity, is the calculation of statistical power from the experimental data obtained (so-called retrospective or *post-hoc* power analysis). In this approach an applicant may seek to compensate for a possible lack of power in a relatively poorly replicated experiment by adjusting the size of the experiment (the Type I error rate), which uniquely determines the retrospective power of the experiment. Problems associated with such a strategy were identified, for example by Schuirmann (1987), Hoenig and Heisley (2001) and by Walters (2008). Tempelman (2004) pointed out how a poorly executed experiment would be rewarded a greater chance of concluding equivalence. It must be emphasised that one of the approaches proposed in this document, which specifies explicit equivalence limits and then employs two types of hypothesis test, overcomes the problems mentioned above.

Notwithstanding the problems of retrospective power analyses, it can still be useful to reassess studies for which a prospective power analysis was done, to check model assumptions and parameters estimated *a priori*. For example, Clark *et al*. (2005, 2007) assessed the results and power analysis of the UK Farm Scale Evaluations.

### 1.3. Decision analysis, tests and confidence intervals

The result of a risk assessment should be a risk characterization to be used by risk managers for decision making. From a statistical point of view there are several approaches that can be taken to decision problems. In the most general form a decision theoretic approach can be followed (see e.g. Lindley 1998). A more classical approach is hypothesis testing.

The decision analysis approach of Lindley (1998) requires the specification of relative losses connected with the two types of erroneous decisions. In case of application to GMO risk assessment, the decision analysis approach would require answering the following question: what is the relative loss when approving a specified use of a GMO that is not equivalent, in comparison to the loss when prohibiting the use of a GMO that in fact is equivalent? Although the GMO Panel considers this interesting from a statistical point of view, this is out of the scope of the risk assessment process.

Hypothesis tests can be performed in isolation (purely as a process of rejecting or not rejecting a null hypothesis) or they can be performed after the construction of confidence intervals. It is well-known

that there is usually symmetry between hypothesis testing and the construction of confidence intervals. In fact, a 95 % confidence region is the set of null hypothesis values that would not be rejected by a 95 % confidence test using the same data. The use of point estimates and associated confidence intervals has been advocated earlier (e.g. Gardner and Altman 1986, Kieser and Hauschke 2005, Newman 2008).

There are several advantages connected to the use of confidence intervals for testing hypotheses:

1. The result is not only a yes/no decision about rejecting the null hypothesis, but it gives a more detailed description of the magnitude of the difference between the GMO and its conventional counterpart as well as the uncertainty about this difference.

2. When two different hypotheses have to be tested (as is the case when both the proof-of-difference and the proof-of-equivalence tests are done) then only one confidence interval needs to be constructed.

3. It is possible to prepare graphical overviews, which is especially useful when there are multiple endpoints to be tested.

4. Confidence intervals can be constructed even in the absence of clearly defined null hypothesis values (e.g. in the absence of equivalence limits).

For these reasons the GMO Panel proposes the use of confidence intervals as a standard instrument for the testing of differences as well as equivalence. Of course, because of the fundamental equivalence between confidence intervals and tests, the results can be supplemented with test results (e.g. in the form of *p* values) when this is considered useful.

Conventional confidence intervals are two-sided, meaning that they have a lower and an upper limit. The concern may be a potential difference between a GMO and its conventional counterpart in one direction only (either an increase or a decrease). In such cases a one-sided confidence limit is more appropriate, having more power at the same confidence level.

When a multiplicative scale is appropriate, the GMO Panel proposes confidence intervals for the ratio of the GMO to its conventional counterpart and as long as the data are in reasonable agreement with the necessary conditions for statistical analysis of that ratio (see Section 3.1 for further discussion). The advantage of considering ratios is that often endpoints vary naturally on a multiplicative, rather than additive scale. On such a dimensionless scale, treatment effects are expressed in terms of proportional or percentage change; these can be easily compared over multiple endpoints. Several approaches for relative confidence intervals are available, depending on the assumed characteristics of the endpoint: i) lognormal distribution, ii) normal distribution and iii) any continuous distribution. For counts and proportions analogous relative confidence intervals are available as well (e.g. for risk ratios or odds ratios).

Statistical methods are described in more detail in Section 3.

## 2. Proposals concerning field trial design

### 2.1. Experimental design

Field experiments are to be replicated at multiple sites. At each site a field trial is to be conducted with the varieties randomized over plots in multiple blocks (or replications). The statistical analysis of data from the experiments for comparative risk assessment is mainly concerned with studying the average difference and the average equivalence (see Section 4.2) over sites. Here, the term ‘average equivalence’ is adopted in the sense used in the drug testing literature (e.g. Wellek, 2002).

Nevertheless, applicants should allow for the possibility of checking for possible site-specific effects, i.e. genotype by site interactions. If genotype x site interactions are identified, then it is important that each individual site trial is sufficiently well-replicated to allow a credible site-specific analysis at each of the sites. Therefore the requirements for the levels of replication are based on power considerations for single field trials (*per* site).

A good discussion of field experimental design, of some relevance for risk assessment of GMOs, can be found in Anon (2007). This document provides advice for the choice of adequate levels of replication in field trials, for appropriate forms of compositional, phenotypic and of agronomic analysis.

The experimental design problems encountered in comparative assessment are partly similar to those encountered in other studies (Basford & Cooper 1998, Spilke *et al.* 2005). GM crops usually have also to fulfill the requirements for variety registration before being allowed onto the market for field cultivation. In the EU, as in a number of other countries, crop field testing for variety registration comprises two aspects. First, the new variety should be "novel" and assessed and declared as suitably levels of "distinctness, uniformity, and stability" compared to varieties of common knowledge. These requirements comply with the general formats and crop-specific guidelines published by the international plant variety protection agency UPOV (http://www.upov.int). UPOV recommends the use of randomized block designs for most field trials with each block containing separate plots with different varieties, over several seasons and locations. In addition, from an agronomic perspective, the crop should have "value for cultivation and use", *i.e.* it should have an advantage in terms of yield and/or quality over currently used varieties Variety registration in the EU is controlled on a crop by crop basis with respect to numbers of year of testing, locations and the design and replication of the trial (see details from the EU Community Plant Variety Office, http://www.cpvo.europa.eu).

Many agronomic and phenotypic endpoints such as plant height, kernel weight and leaf colour may be studied within the same experimental design as for the study of compositional endpoints. Occasionally, some endpoints, such as abiotic stress for which conditions have to be controlled, may require separate experiments. Even in these cases, the same principles for design of compositional studies outlined here should be followed. For example, designs for agronomic and phenotypic endpoints should include commercial varieties to allow equivalence testing.

In an extensive experiment to assess the impact of GM crops on UK wildlife, the problems described by Perry *et al.* (2003) included: (i) the need to decide on the size and location of experimental unit; (ii) the need to choose plots, fields and farms that were representative of the regions for which inferences would be made; (iii) the need to avoid selection bias in randomization of varieties to experimental units; (iv) the need for an auditable procedure to ensure neither the experimenter, recorder or biometrician could influence the randomization; (v) the need for infrastructural underpinning of analysis including database management, data verification, punching, storage, integrity and extraction; (vi) the desirability of a common approach to analysis using automated software where appropriate, especially where the number of variables tested is large. However, this experiment was principally focussed on assessing environmental risk for non-target organisms, so is only weakly applicable to food safety issues.

## 2.2. Power of field experiments

Perry *et al*. (2003) conducted a power analysis to inform replication levels in a study involving GM crops, although their study was concerned with environmental and not food-feed risk assessment. They aimed to provide about 80% power for each analysis for the most important response variables. In their study, Perry *et al.* noted that: (i) it was difficult for biologists to answer in quantitative terms the question: 'what degree of variety difference do you consider important?'; (ii) since power is a continuum that varies gradually with sample size, there is no single threshold level of replication below which an experiment is too poorly resourced to be worth conducting and above which it is satisfactory; (iii) in experiments with many response variables power must vary between them and

cannot be optimised separately for each; and (iv) for variables likely to have non-normal distributions, power estimates might require special calculations. It proved possible to estimate quite accurately the dependence of power on some of the variables mentioned in the paragraph above, namely the magnitude of the difference between varieties, the baseline variability of the experimental units and the replication of the experiment. However, Perry *et al.* (2003) demonstrated that power was model dependent, emphasising the need for model checking. Also, they estimated to what degree power was dependent on the magnitude of the variables measured, emphasising the need to achieve adequate samples.

For risk assessment of GM crops we require good experimental designs to perform compositional, agronomic and phenotypic analysis. The primary difficulty listed above, the lack of ability to specify the difference between varieties required to be detected, is just as problematic for field trials for compositional, agronomic and phenotypic analysis. Furthermore, since experience has shown that generalized guidance on the need to consider power rarely has the effect desired, it is necessary to replace this with specific recommendations concerning the design of the field trials and minimum amounts of replication. The motivation behind the specific replication called for is discussed in the next section.

## 2.3. Choice of levels of replication

### 2.3.1. Number of replications per site

The choice of levels of replication for a field trial should ideally be based on a full power analysis, conducted prior to finalising the design. Otherwise, it may be possible to reach a decision based upon the related requirement that confidence intervals on differences between varieties should be no more than some predetermined width. Failing this, for reasons discussed above, this section considers the relationship between the number of varieties and the degree of replication in relation to the resulting degrees of freedom for error in a simple single-site analysis for a test of difference where all factors are assumed to have fixed effects. This is a simplification that ignores two important issues. The first is that we recommend a test of equivalence as well as a test of difference, and the second is that we recommend the use of random effects to model commercial varieties and possibly also environmental factors. Therefore we recommend the use of mixed models. However, some simplification is unavoidable, since neither sufficient data nor theoretical studies are available to allow us to make recommendations that are statistically optimal in the strict sense of the term.

The approach is based on the idea that the number of degrees of freedom for error may provide a reasonable criterion for the choice of the number of replications per site. For a useful statistical analysis to be made, the number of residual degrees of freedom (df) must be sufficiently large. For example, in an experiment with 8 varieties and 4 replicates with a randomized block design, there are 21 residual df. These are calculated as: total df (32 - 1 = 31) minus variety df (8 - 1 = 7) minus blocks df (4 - 1 = 3), i.e. 31 - 7 - 3 = 21. With only 4 varieties and 4 replicates, the corresponding figure is 15 - 3 - 3 = 9, generally considered of marginal use. Residual df should be increased by increasing the replication; often this will entail using extra blocks in a randomized block experiment.

The number of desirable residual df depends on the questions asked, the form of the data, the degree of precision (power) required of the trial and other contingencies. For example, for components where many values below a certain level are only reported as “less-than”, it can be expected that the estimated residual df from the experimental design will be too low, and in general more replication will be necessary (as well as an adapted method of statistical analysis). Furthermore, whilst it may be the case that for many endpoints, typical values of CV% for field trials of 2-12% may be achieved, for some endpoints such as secondary metabolites, CV% may be much larger. Expert statistical advice should be sought if in doubt. However, in very general terms, experience with trials on efficacy evaluation over many years has shown that it is inadvisable to lay out trials with less than 15 residual df. More degrees of freedom are usually required for a relatively highly-variable endpoint such as a

count of the abundance of an organism the distribution of which may be highly skew, if power is to be similar over all endpoints.

It is stressed that optimal designs for mixed models are still an open problem, and whereas general guidance is given now, the precise optimal design for each particular situation may only emerge as a result of future research in this area, on a case-by-case basis. In general, an investment in more sites and/or replication within sites generally improves any given design.

The choice of the experimental design has an influence on the number of residual df. The fully randomized design gives the maximum number of df. The randomized block design uses some of these df to allow for the heterogeneity of the environment (such as that along one gradient); the Latin square design uses still more, to account for heterogeneity along two gradients. The split-plot design uses df to allow for the possible sources of more than one component of variation. Incomplete block designs are used when the number of varieties in a block is so great that homogeneity of plot variance may be compromised. The experimenter must try to leave the maximum number of df to estimate the residual variation, whilst choosing an optimal design to minimize that variation, by allowing for all the known sources of heterogeneity. Whatever the design, the concept of randomisation is crucial to ensure a proper basis for the estimation of variability. In particular the commercial varieties should be randomized in the same way as are the GM plant and its comparator(s).

In general, there may be results from previous experiments to indicate the likely variability of observations. If such data exist, it is possible to make some judgement as to the design and size of experiment needed to give the required power. Various computer-based or graphical systems are available to assist in determining the number of replicates needed; these use the magnitude of the difference required to be estimated, or the level of significance required for that difference, and the precision expected.

### 2.3.2. Number of commercial varieties

Information on variability between commercial varieties is clearly very important in the setting of equivalence limits. It is good statistical practice to include commercial varieties fully randomized within each of the set of field trials, in addition to the GM plant, its conventional counterpart, and, eventually, additional comparator(s). Again, we stress that there is insufficient information on which to base a determination of the statistically optimal number of commercial varieties per trial (i.e. per site in a multi-site experiment). However, for a good estimate of variability between varieties we consider that data should be gathered from at least three varieties from each trial. Further, since varieties are intended to be representative of the sites at which they are grown, and since sites within the trials are intended to represent the full range of receiving environments, it is likely that different varieties will be used in different trials (i.e. at different sites), and that the range of varieties across the set of trial sites will be larger than at any individual trial site. Six varieties overall should provide a pragmatic minimum basis to estimate variability that will aid the setting of equivalence limits.

### 2.3.3. Number of sites

Environmental variation is manifest on two scales: site-to-site and year-to-year. Many years are required to capture adequately the full range of the year-to-year variation. Since the primary concern is not environmental variation per se, but whether potential differences between the test materials vary across environmental conditions, the approach recommended here defines a minimum number of sites for replication of the field trials, but allows flexibility in the number of years over which those trials are conducted. In the case that sites cover a very restricted geographic range, then replication of trials over more than one year is required.

Similar pragmatic considerations as described above for the number of commercial varieties have been used to recommend a minimum number of sites for the set of trials.

Each field trial must be replicated at a minimum of eight sites, chosen to be representative of the range of likely receiving environments where the crop will be grown. The trials may be conducted in a single year, or spread over multiple years. The commercial varieties may vary between sites, but unless there is explicit justification there must be at least six different commercial varieties used over the entire set of trials.

Experiments may have to be replicated through time because the effects of varieties may alter with, for example, seasonal temperature, photoperiodic effects, etc. Temporally, sample units may be autocorrelated if placed too close together in time. Then, the information in successive samples is less than that in two separate samples; an example might be the increase in insect damage by some pest. Repeated measures analysis may be used to analyze such autocorrelated responses, sometimes taken on the same individual, within a classical analysis of variance framework.

### 2.4. Experiments with multiple GM crops

When it is desirable to assess several different GM plants for one crop species (e.g. *Zea mays*) the production of material for the comparative assessment of these different GM crops may be produced simultaneously at the same sites and within the same field trial by the placing of the different GM plants and their appropriate comparator(s) in the same randomized block.

In order to provide clear recommendations that will lead to robust experiments in the majority of cases, some simplifications are required that are not strictly necessary from the point of view of statistical design theory. For example, for simplicity of the experimental design it could be recommended that two conditions be met: (i) each of the appropriate conventional counterpart(s) must always occur together with its particular GM crop in the same block; (ii) all the different GM crops and their conventional counterpart(s) and all the commercial varieties used to test equivalence with those GM crops must be fully randomized within each block.

As an example, suppose at a particular site, GM1, GM2 and GM3 denote three different GM maize crops; NIC1, NIC2 and NIC3 denote their appropriate respective conventional counterparts; and that CV1, CV2, CV3 and CV4 denote four commercial varieties to be used for the estimation of equivalence limits and equivalence testing of the three GM crops. Then, assuming the minimum number of four randomized blocks is used, one example of the randomized allocation of plants to plots within blocks could be:

| Block | Plot | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | GM2 | CV2 | CV1 | GM3 | NIC3 | NIC1 | CV3 | GM1 | NIC2 | CV4 |
| 2 | CV2 | GM2 | CV3 | NIC3 | NIC2 | GM1 | NIC1 | CV4 | CV1 | GM3 |
| 3 | NIC1 | NIC3 | GM1 | CV1 | GM3 | NIC2 | CV2 | CV4 | CV3 | GM2 |
| 4 | GM3 | GM2 | CV1 | NIC1 | CV2 | NIC2 | NIC3 | CV3 | CV4 | GM1 |

It is recognised that arguments concerning degrees of freedom would show that, for the purposes of statistical analysis, the most efficient approach would be to use all this information simultaneously, as this would help to reduce baseline residual variance. However, the EFSA GMO Panel is of the opinion that, since it is necessary to maintain transparency and verifiability, currently the GM crops should all be assessed separately. Hence, for GM1, only plots 2,3,6,7,8,10 in block 1 should enter the analysis; for GM2, only plots 1,2,3,7,9,10 in block 1, should enter the analysis, etc.

If, and only if the number of plots per block required for such a trial were to exceed 16, then a partially balanced incomplete block design may be used, if desired, to reduce the number of plots per block, by excluding some of the GM crops and their appropriate conventional counterpart(s) from each block. Again, for simplicity of the experimental design, it could be recommended that two conditions be met: (i) each of the appropriate conventional counterpart(s) must always occur together with its particular GM crop in the same block; (ii) all of the commercial varieties must appear in each of the incomplete blocks and be fully randomized with the GM crops and their conventional counterparts.

For example, a trial at a site with 5 commercial varieties, each to be tested for equivalence against 6 different GM crops, each of which had its conventional counterpart, would require a minimum of 4 randomized blocks each with 17 plots per block. These could be replaced, if desired, by 6 incomplete randomized blocks each of 13 plots per block, each comprising the 5 commercial varieties plus 4 of the 6 GM crops, each with its conventional counterpart. As already stated above for the case of a single GM crop assessment, it should be stressed that when several different genetically modified crops are used simultaneously at the same site in this way, all of the crops involved and all of the commercial varieties in the trial must be appropriate for that site, and the requirement of a minimum of 4 replicates per site and of 8 sites in total is unchanged.

An additional possibility is to adopt a linked structure, where some (but not all) of the commercial varieties would be included as usual in the same set of randomized and replicated field trials with one GM and its comparator(s), and then (some of) these commercial varieties may also be used in another set of trials and with perhaps still more commercial varieties, so that the incidence of treatments might be as follows:

| field trials set | Varieties in the set | Commercial Varieties |
|---|---|---|
| 1 | GMO1 comparator1 | 1 2 3 |
| 2 | GMO2 comparator2 | 2 3 4 5 |
| 3 | GMO3 comparator3 | 1 5 6 7 |
| 4 | GMO4 comparator4 | 6 7 8 |

Then the linkages between the commercial varieties over the field trial sets would allow the recovery over inter-set information yielding a more efficient estimate of between commercial variety variance, corrected for differences between the sets. Of course, within each set of field trials there must be consistency with the requirements given earlier.

### 2.5. Experiments with multiple comparators

It may be required to use more than one test-material for comparison (i.e. combination of genetic line and treatment), as for example when herbicide tolerant systems are assessed. Then, three test materials are compared: the GM plant exposed to the intended herbicide, the conventional counterpart treated with conventional herbicide management regimes and the GM plant treated with the same conventional herbicide(s). Such comparison allows the assessment of whether the expected agricultural practices influence the expression of the studied endpoints. Such extra comparators should be fully randomised and replicated, as are the other test materials.

# 3. Statistical approaches

## 3.1. Introduction: choice of model and processing of data

Measurements are made on several scales (continuous, ordinal, quantal, binary, count, multinomial). A statistical model appropriate for the scale used should be chosen. In this opinion we focus on measurements made on a continuous scale, which is appropriate for most compositional, agronomic and phenotypic variables in field studies, and chemical analyses in blood and urine traits measured in animal studies. For measurements made on other scales it is often possible to devise similar statistical approaches as described here.

It is often appropriate to transform data before standard statistical methods are used. For example, many biological effects are manifest on a multiplicative scale rather than on an additive scale. Differences are commonly expressed as a percent change, i.e. as relative differences (ratios) rather than absolute differences. However, most statistical models are additive models, they are used to estimate or test absolute differences. A good choice of a scale for statistical modelling is therefore important. A logarithmic transformation of the data may be appropriate because of the basic property that it transforms a multiplicative model into an additive model, and thus relative differences into absolute differences $\log(A/B) = \log(A) - \log(B)$. Only when reporting results (graphs, tables) these should be back-transformed to the original scale.

Another common phenomenon is inequality of variation (heteroscedasticity), whereas many statistical models assume equal variance (homoscedasticity) among groups of observations. Often the standard deviation increases with the mean, but the coefficient of variation is approximately constant. In these cases a logarithmic transformation is appropriate because the transformed data will become homoscedastic.

Continuous parameters in field trials and animal studies often have a skew distribution, whereas many simple statistical models need the assumption of a symmetric distribution. When the data are reasonably well described by a lognormal distribution, as it seems to be often the case with compositional data, a logarithmic transformation is appropriate to obtain an approximately normal distribution.

Whereas there are many cases in which the logarithmic transformation is an appropriate pre-processing of continuous data, there may be situations where it is inappropriate, and it should never be applied without thought. For example, when values are zero, the logarithmic transformation cannot be applied. Also the assumption of a constant coefficient of variation typically breaks down at very low measurement values, and the log-transformed data may show more variability than at higher levels. In general the appropriateness of the chosen statistical model should be checked, at least by graphical techniques, such as plots of residuals against fitted values.

There may therefore be occasions where the use of normal distribution based models on log-transformed data is not appropriate, either because the data are of a fundamentally different nature (quantal data, ordinal data, counts), or there are outliers, or because assumptions are not fulfilled, e.g. the assumption of lognormality may not hold. Given enough data the assumption of normality may be checked using standard normality tests (e.g. Shapiro-Wilk test, D'Agostino test), but the amount of data available in practical cases is usually too small for this. Another problem may be that even after logarithmic transformation the variances are not homogeneous, and also for this case tests are available (e.g. Levene test). Whereas the relatively simple models proposed in this opinion may not suffice in such situations, it is stressed that the general principles remain valid.

Outlying observations can distort statistical analyses. Applicants should investigate whether this might be a problem. In general graphical approaches are advised, e.g. by looking at residual plots. Rejection of outliers is only allowed when there are biological/technical reasons. Outliers should always be identified. Typically outlier tests play a minor role: their power is limited at the small sample sizes

which are typically available. Outlier tests should never be applied for automatic outlier removal. When outliers have been found in the data, in general it is required to provide analyses based on the data with and without outliers. Finally, in risk assessment it may occur that results which seem to be outliers are in reality the effects of rare but very real anomalous toxicological reactions. An outlying observation may thus be the only important point in the data set, and toxicological rather than statistical expertise is needed to judge this.

There may be more complex reasons why the data fail to have a simple distribution. For example the dataset may be a mixture of responding and not responding animals (as in the tolerance model of toxicology). In such cases a simple statistical approach may not be feasible, and more complex methods may be needed.

### 3.2. Comparative assessment when equivalence limits are available

In this section we treat the situation that would occur if equivalence limits were already available. It is reiterated that for almost endpoints this is not currently the case. For the estimation of equivalence limits, see Section 3.3.

#### 3.2.1. Equivalence limits

In order to test equivalence in a statistically rigorous manner it is necessary to specify for each tested variable a maximum acceptable difference, set either as the difference $\theta$ between the GMO and its conventional counterpart, or as the difference $\theta'$ between the GMO and the mean of commercial reference varieties. As emphasised above, consideration of quantities related to $\theta$ in this section refer to future scenarios where equivalence limits are assigned set values; consideration of quantities related to $\theta'$ are covered in Section 3.3, where the estimation of equivalence limits is discussed. It is possible that $\theta$ will be a value on a transformed scale. For a logarithmic transformation $\theta$ corresponds therefore to a maximum acceptable percent change. In principle the limits on the difference can be different in the positive and the negative direction; these are termed, respectively, the 'upper equivalence limit', $\theta_U$, and the 'lower equivalence limit', $\theta_L$.

Customarily, for example in pharmaceutical applications, the expression of equivalence limits (maximum acceptable difference) is done in a relative way, as a percentage (e.g. 20 % difference) or as a multiplication factor (e.g. 1.25). Note that these two ways of specification are fundamentally different, because the use of a multiplication factor translates into asymmetrical percentages. For example, the multiplication factor 1.25 = 5/4 corresponds to +25 % or -20%, and a multiplication factor 2 corresponds to +100% or -50%. When comparisons between GMOs and comparators are made by forming a ratio of the respective values, then this corresponds to a difference, $\Delta$, after logarithmic transformation, and a multiplication factor (e.g. 1.25) transforms to symmetrical limits $\theta_U = \ln(1.25) = 0.223$ and $\theta_L = \ln(1/1.25) = -0.223$. On the other hand a specification of ±20% would correspond to asymmetrical limits $\theta_U = \ln(1.20) = 0.182$ and $\theta_L = \ln(0.80) = -0.223$.

The use of a logarithmic scale is in correspondence with the fact that, most often, limits for continuous variables will be available as relative changes of the GMO with respect to its conventional counterpart. Such relative differences on the original scale (e.g. the GMO mean should be between -20 % and + 25 % of the conventional counterpart mean) correspond to absolute differences on the logarithmic scale. A further advantage of relative effects is the comparability of the confidence intervals of multiple endpoints.

In the field of GMO risk assessment, Hothorn and Oberdoerfer (2006) and Oberdoerfer *et al*. (2005) have chosen to apply equivalence limits of ±20% (range for the GMO mean of 80% to 120% of the

comparator mean), referring to FDA (1997) and Nordic Council (2000). Actually, FDA (2001) mentions, for the case of area under the curve (AUC) of serum content of generic drugs, usual limits based on a factor 1.25, which leads to a range from 80% to 125% of the reference value. These limits are based on analysis of bioavailability studies with drugs administered to humans. This interval is also prescribed as a standard for certain pharmacokinetic parameters in drug testing by EMEA (2001). It is difficult to find further justification for this choice; it is standard only in pharmaceutical applications. It would seem impossible to state that such values would also be optimal for, say, the composition of raw agricultural commodities or for results from animal studies. Moreover, in pharmaceutical research comparisons are made within patients (e.g. using cross-over designs), whereas in field trials comparisons are made within combinations of sites and years, and in animal studies comparisons are made between different groups of animals. It is not at all obvious that this would lead to a similar variation in general. Further investigations for the definition of suitable ranges are needed in this area.

### 3.2.2. Single endpoints in simple two-group designs

For simplicity we first sketch the proposed approach for the simplest situation, where measurements on the GMO and its conventional counterpart are available from two unstructured groups. Data from animal feeding studies may give data of this type. In the next section we discuss more complicated designs such as are usual for field trials.

When testing for differences (proof of difference approach) the null hypothesis and alternative hypothesis are:

$$H_{0:}\ \Delta = 0 \qquad \text{vs.} \qquad H_1:\ \Delta \neq 0$$

or, in words, the null hypothesis is "no difference between the GMO and its conventional counterpart" against the alternative hypothesis: "difference between the GMO and its conventional counterpart". Note that this two-sided test (both increased and decreased endpoints should be detected) is the most common case, but if it is *a priori* known that differences can only be in one direction, then it can be easily adapted to one-sided versions (to detect only increases or decreases).

A statistically significant test result identifies a difference, whether it is practically important or not. For each test with significance level 1 - α (e.g. 95 %), there is a limited Type I error probability (α, the size of the test) that a significant result is obtained (i.e. a difference is found) whereas no difference exists in reality. However, these tests do not restrict the Type II error probability (β) of finding no significance whereas in reality there is a difference. So the absence of significant results is not a proof for equivalence of the GMO and the conventional counterpart, or ''absence of evidence is not evidence of absence'' (Altman and Bland, 1995 and 2004).

When testing for equivalence (proof of equivalence approach) the null and alternative hypotheses are:

$$H_0:\ \Delta \leq \theta_L \text{ or } \Delta \geq \theta_U \qquad \text{vs.} \qquad H_1:\ \theta_L < \Delta < \theta_U$$

Or, in words, the null hypothesis is "the difference between the GMO and its reference is at least a certain minimum size" (i.e. there is non-equivalence) against the alternative hypothesis: "there is no or at most a small difference between the GMO and its reference". In this testing procedure we need a significant result (rejection of the null hypothesis) in order to conclude that the GMO and the reference are equivalent. Thus, there is a limited Type I error (α) that equivalence is concluded whereas a difference larger than this limit value, of a certain minimum size, exists in reality. This way of testing equivalence is used in pharmaceutical applications (FDA, 2001; EMEA, 2001).

Both the difference test and the equivalence test can be implemented using the calculation of confidence intervals. The reasons why this is preferable were discussed in Section 1.3.

In the case of difference testing a (1 - α) confidence interval on $\Delta$ can be calculated, and the null hypothesis will be rejected when the complete interval does not include 0.

In the case of equivalence testing the approach, also called the two one-sided tests (TOST) approach (Schuirmann, 1987), can be performed by computing a (1 - 2α) confidence interval on $\Delta$, and rejecting the null hypothesis when the complete interval falls between the equivalence limits. In equivalence studies the choice of a 90% confidence interval is customary (FDA, 2001; EMEA, 2001) as it corresponds with the customary 95% level for statistical testing. However, it should be stressed that preference for levels of confidence is not a statistical decision, rather one to be made by risk managers. The choice made in this opinion is only made for reasons of simplicity.

Rather than calculating confidence intervals separately with different confidence levels for the difference and the equivalence tests, the GMO Panel proposes to calculate by default two-sided 90% confidence intervals. This implies that each (two-sided) difference test will have a 90% confidence level, and each equivalence test a 95% confidence level. If it has been decided *a priori* that only deviations in one direction are of importance, then one-sided difference tests are appropriate. The confidence level of the procedure where only one of the limits of the two-sided 90 % confidence interval is inspected is also 95%.

Assuming a simple two-group design of the experiment and a lognormal distribution for the observations in each group, a symmetric two-sided 90 % confidence interval is calculated as:

$$(\bar{y}_1 - \bar{y}_0) \pm \{[t(df;0.95)]\, s\, (n_1^{-1} + n_0^{-1})\}$$

Where:

$y$ refers to the natural logarithm of the original observations;

$\bar{y}_1$ is the average value of $y$ in group 1 (the GMO);

$\bar{y}_0$ is the average value of $y$ in group 0 (the chosen conventional counterpart);

$n_1$ and $n_0$ are the number of observations in group 1 and 0, respectively;

$s$ is the pooled within-group standard deviation of $y$; it is calculated from the group standard deviations $s_1$ and $s_0$ as $s = \sqrt{\frac{df_1 s_1^2 + df_0 s_0^2}{df_1 + df_0}}$, where $df_j = n_j - 1$;

$df_j$ are the appropriate degrees of freedom for the $j$th group; (note that in designs with more groups other groups can also be used in this pooling)

$t(df;0.95)$ is the 95 % point of Student's $t$ distribution with $df$ degrees of freedom.

The calculated confidence interval can be plotted together with the value 0 (for difference testing) and the equivalence limits $\theta_L, \theta_U$. Such a plot will immediately reveal whether the GMO is significantly different from the conventional counterpart (at the 90 % confidence level), and/or equivalence can be claimed or denied (at the 95 % confidence level).

When it is considered useful to have results also in the form of $p$ values from statistical significance tests, then these can be easily calculated, without loss of generality, as:

$$p = \Pr\left[ t_{df} > \frac{|\bar{y}_1 - \bar{y}_0|}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} \right]$$ for a two-sided significance test of the difference,

$$p = \Pr\left[ t_{df} > \frac{(\bar{y}_1 - \bar{y}_0) - \theta_U}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} \right]$$ for a significance test of non-equivalence,

Where the latter test is appropriate when $\bar{y}_1 \geq \bar{y}_0$, and where a similar test for the lower limits should be used otherwise. Note that a significance test of equivalence, (that is a test with a null hypothesis of equivalence rather than the usual null hypothesis of non-equivalence) would be available, if required, by simply reversing the inequality sign in the equation above.

In cases where the assumption of lognormality is considered invalid other approaches are needed. If the measurements themselves have a normal distribution, then a two-sided confidence interval (or a one-sided limit) for the ratio of GMO to its conventional counterpart can be estimated solving a quadratic equation according to Fieller (1954). Related simultaneous intervals and limits for comparisons of several varieties relative to a control are available according to Dilba *et al*. (2004). Modified versions for the case of variance heterogeneity are available as well.

If neither lognormality nor normality can be assumed for the endpoints, a non-parametric Hodges-Lehmann-type confidence interval for the ratio of medians of continuous endpoints is available according to Hothorn and Munzel (2002).

Three approaches for the ratio of means (assuming a lognormal distribution according to Chen and Zhou (2006), assuming a normal distribution, assuming any continuous distribution) are available in the R library pairwiseCI (*http://cran.r-project.org/web/packages/pairwiseCI/*) together with the modification for variance heterogeneity for the case of normal distribution according to Tamhane and Logan (2004). Simultaneous confidence intervals and limits can be estimated by means of the R library mratios (*http://cran.r-project.org/web/packages/mratios/*).

Usually multiple endpoints are to be tested. If the assumptions hold, then the procedures of this section have the correct statistical properties for single endpoints. It is advised to apply the procedure to a series of endpoints, and plot the results for many endpoints together in one graph (or a few graphs). However, for the complete simultaneous assessment the overall confidence level will then be lower, i.e. the probability of a type I error (finding at least one difference where none exists) will be higher than the nominal value (10 %) in a proof of difference. See Section 3.2.4 for further discussion.

### 3.2.3. Single endpoints in more complex experimental designs

For more complex experimental designs and/or other assumptions regarding the statistical distribution it will often be possible to calculate similar confidence intervals as given in the previous section, though by the application of perhaps more advanced statistical methods. Common techniques are analysis of variance (ANOVA) with fixed and/or random effect models. The Residual Maximum Likelihood (REML) method is another well-known algorithm for fitting these models. In principle the formulae in the previous section that dealt with single endpoints in simple two-group designs might apply also to these more complex situations if *s* is replaced by the appropriate residual standard deviation and *df* by the appropriate degrees of freedom. It is recommended to pool estimates of residual variation over all treatments in the experiment. For example, when commercial varieties are

included in the design for the estimation of equivalence limits (see Section 3), the residual variation for these varieties can also be included in the estimate of residual variation to be used for the two tests of difference and equivalence. For unbalanced designs it is recommended to calculate degrees of freedom by the Kenward-Roger method (Kenward & Roger, 1997; Spilke *et al.* 2005; and see example in Section 5).

Commonly in field trials, the GMO is compared with the appropriate conventional counterpart and other commercial varieties at different sites, and possibly over repeated years, using a completely randomized block design on the particular field. The number of plots, the experimental units to which treatments are randomized, has previously often been rather small at each site, typically between 3 and 8 (see Sections 2.2 and 2.3 above). With such small sample sizes, the power of a fixed effect ANOVA model including interactions between the fixed factors variety, site, (possibly year) would be inappropriately small. Particularly, the power of a per-site (per-year) evaluation is commonly so small, that such an evaluation is not adequate for claiming equivalence.

A mixed effect model can be used for the analysis of the complete data set (all sites and/or years) where the factors site and, if present, year are assumed to be either random or fixed, depending on the details of the experimental design. Here, and in the example in Section 5, we will assume random site and year effects. The power of the comparison between the GMO variety and its conventional counterpart (and the other commercial varieties) depends in a complex manner on the number of plots, the number of sites and the number of years. The between-site, between-replication, between-plot and possibly the between-year variability will be estimated as related variance components. The primary objective for an average difference/equivalence approach is neither the identification of possible interactions nor per-site (per-year) evaluation. Instead, overall (for all sites, plots, years) confidence limits are estimated, allowing an overall claim of equivalence. However, to aid the identification of unintended effects that might otherwise be missed in an overall analysis it is required that applicants should provide a table or graph, giving, for each (transformed) endpoint, the means and standard errors of means of the GM and conventional counterpart(s) for each site.

More specific recommendations for analysis are given in Section 3.3.3.

### 3.2.4. Multiple endpoints

In an agronomic, phenotypic or compositional analysis or in an animal study there are usually many analysed endpoints, each representing a different biological and/or chemical characteristic. For a comparative risk assessment of GM plants and/or derived foods/feeds it is then necessary to integrate the statistical findings on all the endpoints of interest. As mentioned briefly in Sections 1.3 and 3.2.2, this can be done in an informal way, or more formal statistical approaches can be applied. The GMO Panel notes that such formal statistical approaches are still very much under development. Possible approaches can be based on an integration of statistical procedures for single endpoints (multiple comparison approach) or on the application of statistical methods for multivariate data (multivariate analysis approach).

#### 3.2.4.1. Possible approaches for multiple comparisons

When it is required to establish equivalence for each individual endpoint, global claims of no differences at all, or of equivalence for all endpoints become very difficult. The probability of obtaining significant differences by chance alone may become large. The same is true for the probability that at least one of the endpoints cannot be shown to be equivalent. This increasing difficulty to provide answers by statistical means is a direct consequence of the implicitly increasing vagueness of the questions being asked when many endpoints are considered. Better definition of equivalence limits at the beginning of the process can help obviating the problem by limiting the number of significant deviations of endpoints which may be considered, in the end, biologically or toxicologically irrelevant.

Problems of multiplicity are ignored in many statistical reports on GMO comparative evaluation. For example in Oberdoerfer *et al*. (2005) numerous non-equivalences are found (and acknowledged), e.g. for calcium, iron, vitamin B1, pantothenic acid, and vitamin E; still the authors claim an overall compositional equivalence. Examples like this illustrate the need for a more scientific and objective approach.

The overall approach favoured in this document for the display of the statistical analysis of a given data set is the simultaneous plotting of single endpoint confidence intervals for the comparison of a GMO and its conventional counterpart (see Section 4), together with lines representing the no difference situation and all the respective equivalence limits.

A first approach to the interpretation of such a display might be an analysis of the number of significant results obtained in comparison with what is expected under certain assumptions. Obviously, when testing a large number ($p$) of endpoints each at level $\alpha$, then, ignoring certain complications outlined below, $\alpha \cdot p$ tests can be expected to give a significant result by chance alone. For example, with $p = 500$ and $\alpha = 0.10$, fifty spurious significances are expected. And due to random variation this number is expected to be even larger in half of all cases.

There are at least two reasons why finding more than $\alpha \cdot p$ significant test results should not be unexpected under more realistic assumptions of GMO comparative assessment. First, endpoints typically are correlated. Allowing for correlation leaves the expected proportion of test results that are significant by chance unchanged (at 5 or 10 %), but shows that deviations from this expected proportion are more likely than under the assumption of completely uncorrelated variables. Effectively, the number of endpoints is less than the nominal number $p$. To assess whether the actually observed number of significant test results can be due to random variation alone it is worthwhile to estimate by simulation how likely it is to observe this many significant results under the assumption that GMO and conventional counterpart means are exactly the same, but with a correlation structure as estimated from the data. See EFSA (2007) for an example of this simulation approach.

Secondly, there is a discrepancy between the assumption of strict equality of the GMO and conventional counterpart means that is used as a null hypothesis in the statistical test of difference, and the idea of the existence of natural variation between any pair of varieties. When it is generally accepted that there is natural variation between lines, then it is also reasonable to expect some variation between the GMO and the comparator varieties. Again, simulation can be used to estimate how likely it is to obtain the actually observed number of significant results under the assumption that GMO and conventional counterpart means might in fact be slightly different, given a distribution of acceptable differences. In these simulations the degree of acceptable difference should be specified, and that can for example be taken equal to the observed variation between the means of the commercial reference varieties. In general, this procedure then estimates the distribution of the number of significant differences that would be obtained with difference tests between two randomly chosen commercial varieties. This is considered a useful point of reference for judging the actually obtained number of significant differences between GMO and conventional counterpart. An example of this simulation approach to evaluate the number of observed significant differences can also be found in EFSA (2007), and in Section 4 of the current opinion.

Formal approaches to multiple hypothesis testing usually consider the difference-testing case (see e.g. reviews in Shaffer, 1995 or Dudoit *et al*., 2003 the latter in the context of microarray experiments). Much less attention has been given to the equivalence-testing case (e.g. Berger, 1982; Bofinger and Bofinger, 1995; Berger and Hsu, 1996; Wang *et al*., 1999; Quan *et al*., 2001; Romano, 2005). In addition, most of this work is theoretical and not yet adaptable to cases where practical analysis is required.

Given known limits and comparable equivalence tests, global equivalence can be claimed when each individual test decides on equivalence, each at level $\alpha$. This multiple testing approach with individual level $\alpha$ tests is a consequence of the intersection-union (IU) test principle (see details Berger, 1982).

However, this global test is rather conservative and ignores the correlation between the endpoints. Up to now, no IU-test taking the correlation between endpoints into account is available. However, it is not likely that in a real study with many endpoints a claim on global equivalence is possible, at least with sensitive equivalence limits.

It has been suggested recently in the literature on practical GMO risk assessment that *p*-value adjustment using the concept of the false discovery rate (FDR) would be useful to account for the numerous comparisons and to minimize the number of declared significances (Herman *et al*., 2007; McNaughton *et al*., 2007). The false discovery rate is the expected proportion of false positive tests among all rejected hypotheses. It was introduced by Benjamini & Hochberg (1995), and many modifications have been suggested, most notably by Storey (2002). It has gained popularity for assessing significance in genomic studies with thousands of features (see e.g. Storey and Tibshirani, 2003; Dudoit *et al*., 2003; Pawitan *et al*., 2005). However, in such studies one is typically interested in the quality of the inference in the subset of variables which were found to be significantly different following a certain test procedure. This means that FDR as usually applied (i.e. in a context of difference testing) is a property of the subset of endpoints for which a significant difference has been found. It does not address the endpoints for which no significance has been found and therefore FDR applied to difference testing does not seem sufficient as a measure in GMO risk assessment. It could be of interest to adapt the FDR concept for equivalence testing, i.e. for a situation where hypotheses are reversed, but the GMO Panel is not aware that this has yet been done.

#### 3.2.4.2. Possible approaches for multivariate analysis

Formulating hypotheses in multivariate space is standard for difference testing. However, there is little experience with multivariate tests of equivalence (see e.g. Brown *et al* 1995, Munk and Pflüger 1999, Enot and Draper 2007). In future work such approaches could be further investigated along the following lines:

1. Modelling of biological variation - The ordinary biological variation between reference varieties may be captured by studying the multivariate dataset of variety mean values with e.g. principal component analysis (PCA). Statistical models considering an appropriate low-dimensional subspace where most biological variation occurs may be defined. The importance of within-variety biological variation could be investigated and possible ways to model it.
2. Modelling of equivalence - Boundaries of biologically/toxicologically relevant differences can be defined for example as *p*% confidence or tolerance bounds in the multivariate space. The acceptable region in the simplest case will be an ellipsoid.
3. Equivalence testing - Multivariate equivalence tests can be performed for the GMO variety by using the within-variety variation in a test with null hypothesis that the GMO is on the boundary of the acceptable region and the alternative hypothesis that it is inside the acceptable region. As in the univariate case the tests may be implemented using (multivariate) confidence sets.

### 3.3. Estimation of equivalence limits

#### 3.3.1. Which data can be used?

Often the information on the natural variation of levels of relevant crop constituents is rather limited (Kuiper *et al.*, 2002). There are several ways in which data on natural variation may be available.

1. In addition to the GMO and its conventional counterpart, the trials performed include several commercial crop varieties, which must represent non-GM varieties with a proven history of safe use, and these should be fully randomised as integral parts of the experiment.

2. Data on such commercial crop varieties may be available from other experiments, databases or in the literature.

In the opinion of the GMO Panel, the first of these is mandatory in principle, whereas the other options may be alternatives only for those rare cases that might be met in the future, where strong justification can be given why the first option was impossible. The first type of information is required, because the comparison between genotypes is made under strict experimental procedure designed to eliminate confounding effects (see Section 2). As additional evidence for the risk characterization phase, in which the statistical results of field trials are put into the context of biological or toxicological relevance, the possibility to use other experiments or historical data to provide alternative estimates of equivalence limits should also be considered. This possibility and checks on data quality are further discussed in Section 3.3.3.

### 3.3.2. Use of concurrent data to estimate equivalence limits, and the testing of difference and equivalence

When commercial varieties are included in the same experiment where the GMO is tested against the conventional counterpart(s) then data on commercial varieties are obtained in identical conditions to that of the GM and its conventional counterpart. This has the major advantage of eliminating uncontrollable confounding effects. The additional number of plots required is minimal because the commercial varieties can be grown on some of the plots that would otherwise have to be allocated to either the GM or its conventional counterpart (see Section 2).

The total variability of each endpoint observed in the field trials can then be estimated and partitioned using appropriate statistical models in order to derive two sets of confidence limits and to set a lower and upper equivalence limit based on the variability observed among the commercial varieties. One set of confidence limits is used in the test of difference; the other set and the equivalence limits are used in the test of equivalence. The recommendation is that the both tests, with null hypotheses of difference and of non-equivalence, respectively, are always performed.

It is sensible to derive equivalence limits by considering how the commercial varieties compare to the GMO. Established equivalence of the GMO and commercial varieties has often been interpreted as relevant for subsequent toxicological risk assessments. If on the other hand the GMO differs from the commercial varieties then this may be a reason for concern, and the result should be placed in context and interpreted within a risk assessment framework. It is recommended to apply a linear mixed statistical model, fitted to (possibly transformed) data, in order to derive an estimate of variation between commercial genotypes. One model (denoted model 1) should be used for calculation of the confidence limits for both tests (difference and equivalence); a slightly different model (model 2) should be used to estimate the equivalence limits to be used in the equivalence test.

Denote by $I$ an indicator variable (uncentered in the mixed model) such that $I$=1 for a field plot having any of the commercial varieties, and $I$=0 otherwise. Then the random factors for model 1 should include, but not necessarily be restricted to, those representing the variation: (i) between the test materials (a set that includes the GM crop, its conventional counterpart, each of the commercial varieties and any additional comparators); (ii) in the interaction between the test materials and $I$; (iii) between sites; and (iv) between blocks within sites. Model 2 should be identical to model 1 except that the random factor representing the interaction between the test materials and $I$ is omitted.

The fixed factor for both models should have as many levels as there are test materials and represent the contrasts between the means of the test materials. Here, the set of commercial varieties should be considered as a single level of the fixed factor. For the difference and equivalence tests, the component of the fixed factor of interest is the single degree-of-freedom contrast between, respectively, the GM crop and its conventional counterpart, and the GM crop and the set of commercial varieties.

Both the difference test and the equivalence test are implemented using the well-known correspondence between hypothesis testing and the construction of confidence limits. In the case of

equivalence testing the approach used should follow the two one-sided tests (TOST) methodology (e.g. Schuirmann, 1987) by rejecting the null hypothesis of non-equivalence when the both confidence limits fall between the equivalence limits. The choice of 90% confidence limits corresponds to the customary 95% level for statistical testing of equivalence.

For each endpoint, calculation of the confidence limits, estimation of equivalence limits and associated statistical tests should be performed as described below, using the following notation. Sample means are denoted by *m*, with subscripts G, C and R for the GM crop, its conventional counterpart and the set of commercial (reference) varieties, respectively. The variability encompassed in the standard error of the difference between the means of any two test materials, X and Y, calculated using model *i* (*i* = 1,2), is denoted *sed*(XY;*i*). The 100*a*% point of Student's *t* distribution is denoted as *t*(*df*;*i*;*a*), where *i* denotes the model used and *df* is the appropriate number of degrees of freedom which is recommended to be calculated by the Kenward-Roger method. The least significant difference between the means of any two test materials, X and Y, using model *i,* should be calculated as the product of *t*(*df*;*i;a*) and *sed*(XY;*i*), and is denoted *lsd*(XY;*i*;*a*).

For the difference test, the two-sided 90% confidence limits should be calculated about $m_G$, as $m_G \pm lsd$(GC;1;95); the null hypothesis of equality between $m_G$ and $m_C$ should be rejected and the test deemed statistically significant if $m_C$ falls outside these limits. For the equivalence test, the two-sided 95% equivalence limits should be estimated as $m_R \pm lsd$(GR;2;97.5) and two-sided 90% confidence limits should be calculated about $m_G$, as $m_G \pm lsd$(GR;1;95); the null hypothesis of non-equivalence should be rejected and the test deemed statistically significant if and only if the confidence limits lie entirely inside the equivalence limits.

For reporting purposes, it is recommended that full details be given for each endpoint analysed, listing: (a) the assumptions underlying the analysis, (b) full specification of the mixed models chosen, including fixed and random effects, (c) results of any test of interaction between the test materials and sites, (d) fixed effects, together with the appropriate estimated residual variation with which it is compared, and variance components for the random factors, (e) estimated degrees of freedom, (f) any other relevant statistics. The likely impact of other growing conditions not tested in the trial should also be discussed.

The recommended graphical approach to the display of the results of this analysis is outlined in Section 4.1, which describes a convenient method to assess simultaneously the results of both the difference and equivalence test. An example of the mixed model analysis is given in Section 5. This approach assumes that the available commercial varieties represent the population of varieties with a history of safe use with regard to the endpoints of interest. Usually there will be no formal mechanism of variety selection, and therefore it will be up to the scientist performing the study to justify this assumption. Obviously, there will be more confidence in the procedure when the number of commercial varieties is large. With less than six commercial varieties alternative methods should be considered, for example, methods assuming the same variability within groups of endpoints. This again requires a series of *a priori* decisions to be made by the scientists responsible. Limits calculated in this or any other manner from available data should be scrutinized to check that they represent acceptable equivalence limits. A purely statistical approach can make little progress towards reviewing acceptability, which should always have a proper biological/toxicological basis.

It is a consequence of the simplified graphical display that confidence limits for the difference test were chosen as 90%, yielding a 10% size for the difference test, in which 1 in 10 of such tests is expected to yield a significant result by chance alone. Despite the expected proportion of spurious significant differences, it is recommended that the responsible scientist should catalogue, report and discuss all significant differences observed between the GM crop, its conventional counterpart and, where applicable, any other test material, focusing on their biological relevance through risk characterisation.

### 3.3.3. Use of literature or databases to set equivalence limits

There may be rare cases where it is impossible to assess the natural variation from data on commercial varieties in the same experiment, either because such an experiment would be impossible or unreasonable to perform, or because such varieties have for unforeseen reasons not yielded satisfactory data from the experiment. If very strong and explicit justification can be given for not performing an experiment with commercial varieties then the use of external data on natural variation might be considered.

In order of preference such external data may be:

1. Data on such commercial crop varieties may be available from other experiments.
2. Historical data on parameter values connected to safe use of the crop may be available in databases collected by organisations with a statutory or academic regulatory or risk assessment function, such as ILSI.
3. Historical data on parameter values connected to safe use of the crop may be available in the public literature or in research reports, and in particularly from meta-analyses comprising several such sources.

With such external data it is extremely important to check the following points:

1. Is the measured variable the same (commensurability)?
2. Are the data representative of the environmental and genotypic variation (over space, time, varieties, etc.)?
3. Are the experimental or sampling conditions, under which the data were obtained, sufficiently known in order to estimate the natural variation relevant for the GMO to control comparison in the current experiment?

Exactly how such external data can be used to set equivalence limits must be evaluated on a case-by-case basis, and should for example include accounting for inter-study variability, weighting of different estimates according to sample sizes, discounting of data based on data quality considerations, etc.

In general it may be expected that natural variation estimated from external data will not only describe genotypic variation, but also environmental variation. Therefore limits obtained from literature data can be expected to be wider than limits obtained from concurrent data. Allowance must be made for this.

In the rare cases where external data are used to set such fixed equivalence limits, the procedure for analysis is as outlined in Section 3.2.2.

### 3.3.4. Comparative assessment when there are no known equivalence limits

When equivalence limits are not known and no data to estimate such limits are available, then a comparative safety evaluation may have to be based on subjective evaluations of equivalence limits.

Statistical methods can help in presenting the available information optimally. Confidence intervals for the difference between the GMO and its conventional counterpart can be plotted in the same way as described below in Section 4.1. Using plots showing confidence intervals for all endpoints simultaneously the risk assessor may be able to define *ad hoc* limits, considering the observed pattern and available biological knowledge. However, such limits need independent confirmation. In particular it should be stressed that their use for a statistical equivalence assessment is only valid for future experiments. They cannot be used for any statistical interpretation in the assessment in which they have been estimated.

To summarize, if there is no possibility to set or estimate equivalence limits then the ability of statistical methodology to contribute to risk assessment is very limited, other than offering ways to summarize and describe data.

## 4. Interpretation of analysis

### 4.1. Types of possible outcomes of the comparison between the test materials

The recommended method of considering the results of the two tests outlined in Section 3.3.2 above, is to display as many of the analysed endpoints as is feasible, simultaneously, on the same graph. More than one graph is required if endpoints are analysed on different scales, and/or if some are transformed and others not. After the appropriate transformation, simultaneous display is facilitated by shifting all relevant values for each particular endpoint to a scale that has $m_C$, the mean of the conventional counterpart for that endpoint, as its baseline zero value. Therefore, on this new scale, the values of the means of the GM crop, its conventional counterpart and the set of commercial varieties become, respectively: $m_G - m_C$, 0, $m_R - m_C$. Note that a difference of 0 on an additive scale corresponds to a ratio of 1 on a multiplicative scale. Hence, in principle, for a multiplicative scale, both the mean of the GM crop and the equivalence limits can be displayed as ratios to the conventional counterpart (but see below for certain adjustments required to achieve a valid practical outcome).

After shifting all relevant values to the new zero baseline, the confidence limits for the difference test on this new scale become: $m_G - m_C \pm lsd(GC;1;95)$, the equivalence limits $m_R - m_C \pm lsd(GR;2;97.5)$, and the confidence limits for the equivalence test $m_G - m_C \pm lsd(GR;1;95)$. Note that the equivalence limits, chosen to be symmetrical around the centre of the distribution of commercial varieties, are typically asymmetrical (before and after adjustment) on this new scale.

To facilitate visual interpretation, instead of using the two sets of confidence limits in the graphs, it is recommended for convenience that only one be displayed, that for the difference test. Without some adjustment, the confidence limits for the difference test would not give a valid visual representation for the equivalence test on the graph. This problem is overcome by making an adjustment to the displayed equivalence limits. After this adjustment the displayed confidence limits for the difference test may be used as a basis also for the visual representation of the equivalence test. In this way, one confidence limit may serve visually for assessing the outcome of both tests simultaneously. The adjustment of the equivalence limits consists of two steps: (1) scaling the basic equivalence limits, so that the confidence limits required for the difference and equivalence tests have the same width; and (2) an appropriate shift to facilitate display of the adjusted limits, together with $m_G$, on the scale that has $m_C$ as its baseline zero value. The adjusted equivalence limits for visual display should be calculated by the formula:

$$(m_G - m_C) + \{[(m_R - m_G) \pm lsd(GR;2;97.5)]\, lsd(GC;1;95) / lsd(GR;1;95)\}$$

It is recommended that the graph should show the line of zero difference between the GM and its conventional counterpart and, for each endpoint: the lower and upper adjusted equivalence limits, the mean difference between the GM and its conventional counterpart, and the confidence limits for this difference (see the set of possible example outcomes for a single endpoint in Figure 1, below). When, in addition to the conventional counterpart, another test material is used as comparator, the mean difference between the GM and that comparator, its confidence limits and its adjusted equivalence limits should be displayed on the same graph referred to above, for all such additional comparators, by referring this to the same zero baseline as defined by the conventional counterpart. The horizontal axis should be labelled with values that specify the change on the natural scale. In the case of a multiplicative scale and/or a logarithmic transformation, changes of 2x and ½x will appear equally spaced on either side of the line of zero difference. It is assumed here that the line of no difference is in between the equivalence limits. If that is not the case then the conventional counterpart is itself non-

equivalent to the commercial varieties, and a separate, non-statistical discussion should consider the place and relative importance of difference and equivalence testing in the risk assessment.

Regarding proof of difference, each outcome from the graph should be categorised as follows and the respective appropriate conclusion should be drawn:

- Outcome types 1, 3 and 5: the confidence interval bar overlaps with the line of no-difference. The null hypothesis of no difference cannot be rejected and the appropriate conclusion is that there is insufficient evidence that the GM crop and its conventional counterpart differ.
- Outcome types 2, 4, 6 and 7: the confidence interval bar does not overlap with the line of no-difference. The null hypothesis of no difference must be rejected and the appropriate conclusion is that the GM crop is different from its conventional counterpart.

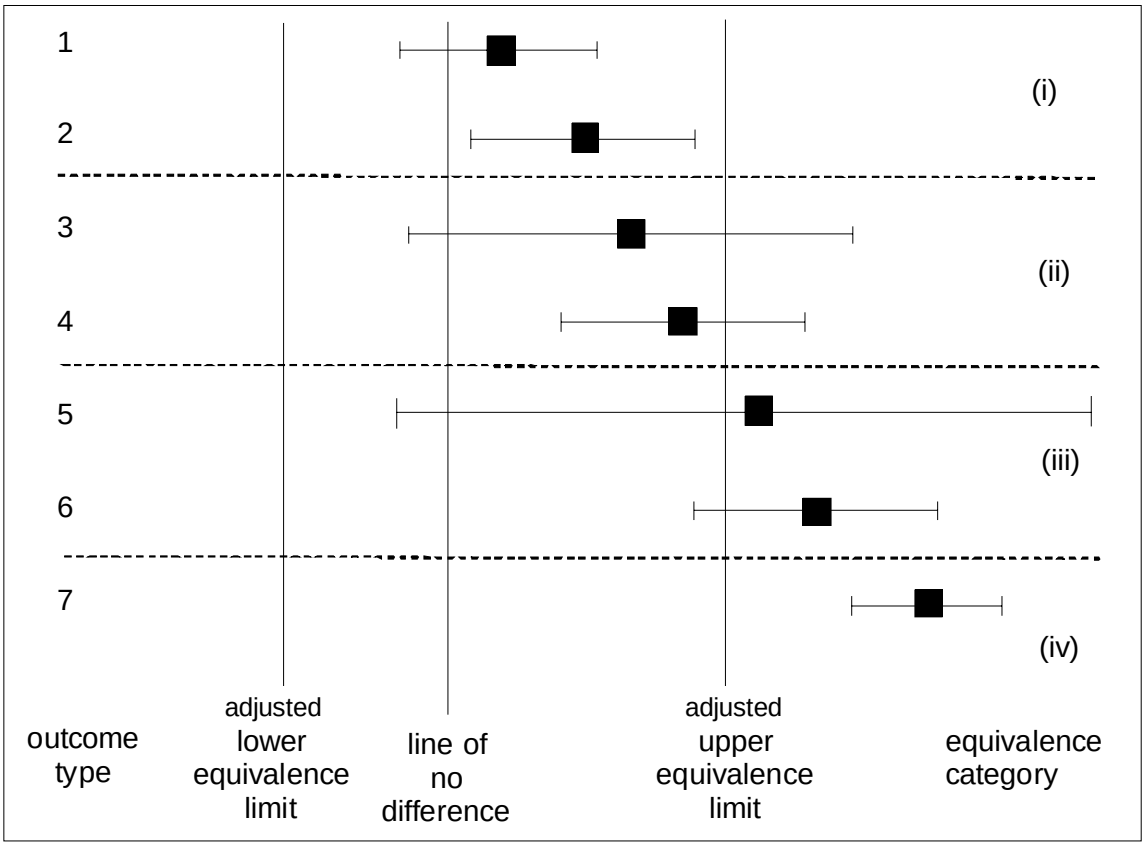


Figure 1. Simplified version of a graph for comparative assessment showing the 7 outcome types possible for a single endpoint. After adjustment of the equivalence limits, a single confidence limit (for the difference) serves visually for assessing the outcome of both tests (difference and equivalence). Here, only the upper adjusted equivalence limit is considered. Shown are: the mean of the GM crop on an appropriate scale (square), the confidence limits (whiskers) for the difference between the GM crop and its conventional couterpart (bar shows confidence interval), a vertical line indicating zero difference (for proof of difference), and vertical lines indicating adjusted equivalence limits (for proof of equivalence). For outcome types 1, 3 and 5 the null hypothesis of no difference cannot be rejected: for outcomes 2, 4, 6 and 7 the GM crop is different from its conventional counterpart. Regarding interpretation of equivalence, four categories (i) - (iv) are identified: in category (i) the null hypothesis of non-equivalence is rejected in favour of equivalence; in categories (ii), (iii) and (iv) non-equivalence cannot be rejected. See text for what appropriate conclusions may be drawn.

Regarding proof of equivalence, each outcome from the graph should be categorised as follows, and the respective appropriate conclusion should be drawn:

- Outcome types 1 and 2 (category (i), Figure 1): both confidence limits lie between the adjusted equivalence limits and the null hypothesis of non-equivalence is rejected. The appropriate conclusion is that the GM is equivalent to the set of commercial varieties.

- Outcome types 3 and 4 (category (ii), Figure 1): the mean of the GM crop lies between the adjusted equivalence limits, but the confidence interval bar overlaps at least one of the adjusted equivalence limits on the graph. Non-equivalence cannot be rejected, but the appropriate conclusion is that equivalence between the GM and the set of commercial varieties is more likely than not. Further evaluation may be required.

- Outcome types 5 and 6 (category (iii), Figure 1): the mean of the GM crop lies outside the adjusted equivalence limits, but the confidence interval bar overlaps with at least one of the adjusted equivalence limits. Non-equivalence cannot be rejected and the appropriate conclusion is that equivalence between the GM and the set of commercial varieties is less likely than not. Further evaluation is required.

- Outcome type 7 (category (iv), Figure 1): both confidence limits lie outside the adjusted equivalence limits. The appropriate conclusion is that the evidence analysed here demonstrates non-equivalence between the GM and the set of commercial varieties. Further evaluation is required.

Note that formally in statistics the null hypothesis is never 'accepted' and that instead the formal conclusion for outcomes 3-7 is that the null hypothesis 'cannot be rejected'. However, regarding outcome 7, for all practical purposes the reasonable conclusion of non-equivalence should prevail for subsequent risk characterization. The direct interpretation of the outcome types 3-6 with respect to GMO risk assessment may be more difficult and may need further safety evaluation, possibly using alternative statistical methods. For example, if differences, even if not statistically significant, were consistent over multiple situations, this could indicate the occurrence of unintended effects. Outcome types 1 or 2 may easily be obtained for characteristics that are stable and precisely measured within each genotype, but that have a large natural variation among commercial genotypes. Outcome types 3 or 5 may easily result when the measurement precision or within-genotype stability is low in comparison to the natural variation.

Following the above interpretation, risk characterization should be used by applicants to specify what further evaluation is needed, based on considerations linked to patterns of observed results and on biological/toxicological relevance.

When there is more than one test-material for comparison (i.e. combination of genetic line and treatment), as for example when herbicide tolerant systems are assessed, the mean difference and its confidence interval for all test-materials should be displayed on one graph, referring all of these, as described above, to the same zero line defined by the conventional counterpart.

### 4.2. Average and site-specific comparisons

Field experiments are to be replicated at multiple sites (see Section 2) and, as specified, at each site a field trial should be conducted with the varieties randomised over plots in multiple blocks (or replications). The statistical analysis of data from the experiments for comparative risk assessment is mainly concerned with studying the average difference and the average equivalence over sites.

Nevertheless, applicants should check for possible site-specific effects, i.e. genotype by site interactions. If genotype x site interactions are identified, then it is important that each individual site trial is sufficiently well-replicated to allow a credible site-specific analysis at each of the sites. In particular, it is recommended that where a difference or lack of equivalence is found for an endpoint, further analysis should be done to assess whether there are interactions between any of the test materials and site. However, it is not absolutely necessary that this further analysis uses the relatively

complex statistical mixed model approach outlined above; a simple, standard ANOVA approach may be all that is required. Applicants should in any case provide a table or graph (see Section 3.2.3), giving, for each (transformed) endpoint, the means and standard errors of means of the GM and comparator(s) for each site; this may be used to aid the identification and discussion of interactions.

### 4.3. More complex situations

Data may be available on endpoints having continuous values (e.g. plant composition or animal blood parameters), discrete values (e.g. counts), or ordinal values (e.g. histological observations). It may or may not be the case that a simple statistical distribution can be assumed to govern the variation of the endpoints. There may or may not be a serious possibility of outliers in the data. In this guidance, which is of a fundamental nature, the focus is on easily understood cases, especially the case of continuous endpoints for which a lognormal distribution can be assumed, without much risk of outliers. The statistical approaches presented in this document should be adapted in more complex situations.

### 4.4. Simultaneous assessment of multiple endpoints

As stated previously, this opinion only provides a limited introduction to the possible application of statistical methods for comparative risk assessment on multiple endpoints. Substantial more work on this subject is needed. Therefore this section of the opinion is intended only to present material to provoke further discussion and future research.

In risk assessment studies many endpoints are measured and in current statistical methodology they are often addressed independently, even though they may be known to be correlated. In a global assessment the relevant issues become more complex because the data from all endpoints have to be considered simultaneously. Just as for single endpoints, the evaluation of multiple endpoints should be specifically adapted for either the proof of difference or the proof of equivalence.

Basically, for both proof of difference and proof of equivalence approaches there are two ways in which a global assessment can be approached using statistical methods:

1. Multiple comparisons. Here the basic statistical approaches are univariate calculations (e.g. tests, confidence intervals) for single endpoints. Additionally, there are procedures that, on the basis of the results from the univariate statistics (e.g. *p* values, confidence intervals), allow reaching global conclusions (e.g. by constructing simultaneous confidence intervals, see Section 3.2.4).

2. Multivariate analysis. This relies on the use of statistical approaches and/or models for multivariate data, including the possibilities to estimate correlations between variables and to consider subspaces of reduced dimension (see Section 3.2.4).

Complementary to either of these, it is always useful to consider a third possibility:

3. Restrict the number of endpoints *a priori* in order to ameliorate the problems of high dimension and multiplicity.

In a multiple comparison framework statistical results obtained for single endpoints are combined. The interpretation of the combined results may proceed in various degrees of formality. An informal procedure is to graph the confidence intervals representing the comparison of the GMO vs. its conventional counterpart together (as it is proposed in Section 1.3). By visual inspection of the graph it is then decided whether there are potential hazards and/or whether the GMO and its control should be termed equivalent.

In a somewhat more formal analysis it can be investigated, e.g. by simulation studies, how many significant results can be expected under the null hypothesis of GMO and conventional counterpart

being equivalent varieties (that is, allowing for the same variation as found between commercial varieties). Such considerations can account for specific information that is available, e.g. observed correlations between endpoints, and/or observed variability between commercial varieties which have a history of safe use. Such an approach was used in a recent EFSA review of the MON 863 maize 90-day rat feeding study (EFSA 2007, Appendix 5), and is also partly illustrated in the example of this opinion (Section 4). Of course, a statement that the number of significant differences is or is not higher than expected should still be accompanied by an expert evaluation of the biological and/or toxicological relevance of the observed pattern of statistically significant results.

More formal approaches to multiple hypothesis testing can be found in the statistical literature (see Shaffer, 1995; or Dudoit *et al*., 2003 for a review). The basic idea is that the Type I and Type II error rates discussed in Section 1.2 are redefined in terms of a set (family) of hypotheses. The family-wise error rate (FWER) is the probability of at least one error in the family of hypotheses.

The objective in the proof of difference is to identify which of the endpoints are different, i.e. changed with respect to the conventional counterpart. The question arises whether in the proof of difference an adjustment against multiplicity (i.e. many endpoints) is appropriate and, if so, which concept of error control is preferred. On the one hand an adjustment reduces the power, i.e. the false negative rate increases. This conservatism induces considerable loss of power in trials where there are many endpoints and/or small sample sizes. On the other hand, without multiplicity adjustment the false positive rate increases as the number of endpoints increases. Whether the control of the family-wise error rate (FWER) or of the false discovery rate (FDR) is more appropriate is a topic of recent research (e.g. Dudoit *et al*. 2003). In any case only those procedures taking the correlations between the endpoints into account, i.e. that restrict the degree of conservatism, can be recommended.

The objective in the proof of equivalence is to characterize equivalence for all endpoints or at least a subset of endpoints. In contrast to the proof of difference, there is an intersection-union test problem. Although the inference is performed on the marginal $(1-\alpha)$ confidence level for each individual endpoint, the global (or subset) decision becomes conservative with increasing number of endpoints.

More research is needed for appropriate simultaneous confidence intervals for multiple endpoints, both in the case of a proof of difference and in the case of a proof of equivalence. In particular, the effects of small sample sizes and the required balance between false positive and false negative error rates must be taken into account.

In a multivariate analysis framework all relevant concepts have to be reformulated in a multivariate context. For example, confidence intervals become confidence regions in multivariate space, and equivalence limits (points on a line) should be replaced by contours of concern in multivariate space. Although this may seem daunting, it may well be possible to apply standard statistical models based on multivariate normality leading to both confidence regions and contours-of-concern of ellipsoidal shape. In multivariate statistics there are many methods to investigate which subspaces are most relevant to describe natural variation, the most well-known of these methods being principal component analysis (PCA).

As a result of the above mentioned methodological difficulties, we recommend for current use: the independent univariate evaluation of single endpoints, a joint graphical presentation, and the reporting and discussion of the frequency of significant results in the set of investigated endpoints.

## 5. Example

### 5.1. Data, experimental design, coding and analysis protocol

This section provides an example of a statistical analysis as part of a comparative assessment regarding GMO safety. The data are a subset of a real dataset obtained from industry concerning a

field study. Here we restrict consideration to the compositional characteristics of maize grain for a GM variety, a comparator variety (the conventional counterpart) and 13 commercial varieties.

The experiment was a randomised block design conducted at four sites in one year. The design protocol called for each site to be planted with the GM variety, the conventional counterpart and four commercial varieties in three replications, but there were some deviations caused by practical considerations. The GM variety was investigated with 3 replications at each site. The conventional counterpart had 3 replications at two of the sites, and 2 replications at the other two sites. Three commercial varieties were investigated at two sites and the remaining 10 commercial varieties at one site only, mostly with 3 replications per site (in some cases only 2 or even 1). The data analysed here concern 14-18 fields per site, for a total of 67 fields.

It may be noted that this experimental design does not conform to the proposed guidelines as set out in this opinion. For example, the number of sites and the replication per site were lower than asked for in this opinion, the conventional counterpart was not included in all blocks, and with a total of 15 varieties a complete block design should have been used. However, in spite of these shortcomings of the experimental design, the data were suitable to illustrate the statistical analysis.

The maize grain was analysed for 68 endpoints, all compositional characteristics. However, for 15 analytes (13 fatty acids, furfural and sodium) all results (or, in one case, all but one) were below a given limit of reporting. As there was no variation in these results which could be used for a comparative evaluation, they were omitted from the further statistical analysis.

Seven results in the remaining set of 53 analytes were reported as less than a certain limit (non-detects): six results for 16:1 palmitoleic acid and one result for phytic acid. The problem seemed minor, and, whereas more advanced statistical methods exist to incorporate such results in modelling, here the non-detects were simply set to half the reporting limit.

Outliers were identified by visual inspection of graphs showing the log-transformed results for each of the three groups (GMO, conventional counterpart, reference). Outliers were identified for four analytes as shown in Figure 2, and also the seven non-detects set to half the limit of reporting were outlying. Purely for ease of illustration, these outliers were omitted from the further statistical analysis.

The purpose of the statistical analysis was to:

1. Calculate confidence limits for the difference between the mean of the GM and the mean of the conventional counterpart: $d_{GC} = m_G - m_C$, as outlined in Section 3.3.2, using model 1;

2. Calculate confidence limits for the difference between the mean of the GM and the mean of the reference varieties $d_{GR} = m_G - m_R$, as outlined in Section 3.3.2, using model 1;

3. Calculate equivalence limits as outlined in Section 3.3.2 using model 2;

4. Prepare graphs for tests of difference and equivalence, first making the necessary adjustments to the equivalence limits, as outlined in Section 4.1.

The logarithmically-transformed data were analysed with the following mixed model (the following information applies both to model 1 and to model 2):

$$y_{ijkl} = m + e_i + r_{ij} + t_k + g_l + \varepsilon_{ijkl}$$

Where the notation used is as follows: $i$, $j$, $k$ and $l$ are indices for environment (site), replication within site, treatment group (conventional counterpart, GMO or reference) and (commercial) reference genotype, respectively. The response $y_{ijkl}$ is the logarithmically-transformed result, using the natural logarithm (ln). The fixed factors in this model are $m$, the overall mean, and $t_k$, the average deviation from the overall mean for each of the three treatment groups ($k$ = 1: conventional counterpart, 2:

GMO, 3: reference genotypes). The random factors in the model are $e_i$, the average deviation for environment $i$, $r_{ij}$, the average deviation for replicate $j$ of environment $i$, $g_l$, the average deviation for reference genotype $l$, and $\varepsilon_{ijkl}$, the residual term for each measurement. As usual in mixed modelling, the random terms are assumed to arise independently from normal distributions with mean 0 and a certain variance component that is to be estimated ($V_e$, $V_r$,$V_g$ and $V_\varepsilon$, respectively). A common way to fit mixed models to data is the residual maximum likelihood (REML) algorithm, which is available in all major statistical packages.

No genotype-environment interaction term was included in the model. Whereas it would usually be of interest to study such interaction, in the current dataset there was insufficient replication of commercial varieties at different sites (environments).

Estimated means, $m_C$ , $m_G$ and $m_R$ , the differences of means, $d_{GC} = m_G - m_C$ and $d_{GR} = m_G - m_R$, and the standard errors of differences, *sed*(GC;*i*) and *sed*(GC;*i*) for model *i* (as defined in Section 3.3.2), are easily available for the fixed effects in mixed models from standard software.

These were used to construct, for the difference test, the two-sided 90% confidence limits $m_G \pm lsd$(GC;1;95); and, for the equivalence test, the two-sided 95% equivalence limits $m_R \pm lsd$(GR;2;97.5) and the two-sided 90% confidence limits $m_G \pm lsd$(GR;1;95) as given in Section 3.3.2. Recall from that section that for the calculation of *df,* where *df* denotes the appropriate number of degrees of freedom, the method of Kenward and Roger (1997) was recommended (Spilke *et al.* 2005). Also, that the least significant difference between the means of any two test materials, X and Y, using model *i,* calculated as the product of *t*(*df*;*i;a*) and *sed*(XY;*i*), was denoted *lsd*(XY;*i*;*a*), and this too is available from many statistical packages.

The differences of means on the logarithmic scale can be back-transformed to ratios of geometric means on the original scale. so the point estimate of the ratio is $10^d$ or $e^d$ depending on the type of logarithm used, and the approximate 100(1-α)% confidence interval is:

$$\left[10^{d-lsd},\quad 10^{d+lsd}\right] \text{ or } \left[e^{d-lsd},\quad e^{d+lsd}\right]$$

Two different implementations of this general mixed model were needed for testing differences and equivalences. The practical implementation of the mixed model for calculating confidence limits of differences (i.e. model 1) in some major software packages is as follows:

---

Genstat:

FACTOR [labels=!T(compGMO,ref)] ref_aside

CALC ref_aside = 1*(genotypegroup.in.!(1,2))+2*(genotypegroup==3)

VCOMPONENTS [fixed=ref_aside/genotypegroup; cadjust=none]\

random = site + site.rep + genotype.indref; constraint=pos

REML y

---

SAS:

proc mixed data=example CL=WALD;

class site rep genotype genotypegroup;

```
        model y =  genotypegroup /s covb outp=out  ddfm=kenwardroger;
        random site site*rep indref*genotype;
        estimate 'gmo_comp' genotypegroup -1 1 0 / CL alpha=0.1;
run;
```

The practical implementation of the mixed model for estimating equivalence limits (i.e. model 2) in some major software packages is as follows:

---

Genstat:

```
FACTOR [labels=!T(comp,GMOref)] comp_aside
CALC comp_aside = 1*(genotypegroup==1)+2*(genotypegroup.in.!(2,3))
VCOMPONENTS [fixed=comp_aside/genotypegroup; cadjust=none]\
        random = site + site.rep + genotype; constraint=pos
REML y
```

---

SAS:

```
proc mixed data=example CL=WALD;
        class site rep genotype genotypegroup;
        model y =  genotypegroup /s covb outp=out  ddfm=kenwardroger;
        random site site*rep genotype;
        estimate 'gmo_ref' genotypegroup 0 1 -1 / CL alpha=0.05;
run;
```

---

For equivalence testing (again using model 1) the relevant code is:

---

Genstat:

```
FACTOR [labels=!T(comp,GMOref)] comp_aside
CALC comp_aside = 1*(genotypegroup==1)+2*(genotypegroup.in.!(2,3))
VCOMPONENTS [fixed=comp_aside/genotypegroup; cadjust=none]
        random = site + site.rep + indref.genotype; constraint=pos
```

REML y

SAS:

```
proc mixed data=example CL=WALD;
        class site rep genotype genotypegroup;
        model y =  genotypegroup /s covb outp=out  ddfm=kenwardroger;
        random site site*rep indref*genotype;
        estimate 'gmo_ref' genotypegroup 0 1 -1 / CL alpha=0.1;
run;
```

These program fragments give only the essential central mixed model calculation. Obviously more programming is needed to read in the data, outlier control, data transformation, and post-processing the results to calculate confidence intervals, equivalence limits and plot the graphs.

Here, a remark may be helpful. Recall that model 2 is identical to model 1 except that the random factor representing the interaction between the group of commercial varieties and the indicator variable, *I*, is omitted. The basic information needed from the mixed model are the means, the standard errors of difference and the corresponding degrees of freedom. With the above two specifications of code for the mixed model (code for model 1 with *genotype.indref* and for model 2 with *genotype* among the random terms) the means and variance components are exactly the same. Only the *sed*s and the *df*s are different. Actually, the *sed*s from the two models are related by:

$$\left(sed_{GC;2}\right)^2 = \left(sed_{GC;1}\right)^2 + 2 \cdot V_g$$
$$\left(sed_{GR;2}\right)^2 = \left(sed_{GR;1}\right)^2 + V_g$$

Where $V_g$ is the variance component for the genotypes. However, it is emphasised that it remains essential to fit the two models to the same dataset because of the separate calculations of the degrees of freedom by the Kenward-Roger method for the two cases.

### 5.2. Results

A graphical overview of the results of the comparative analysis is shown in Figures 3 and 4. More detailed results are given in Figures 5-7, and in Tables 1–8.

Figures 3 and 4 show the relative differences of the GMO with respect to its conventional counterpart. For example, relative large deviations are seen for Acid Detergent Fiber (+10%), Ferulic Acid (-13%), Folic Acid (+14%), Neutral Detergent Fiber (+14%), Niacin (-13%) and Total Dietary Fiber (+12%). However, due to different variabilities, large differences need not be statistically significant (e.g. the interval for Acid Detergent Fiber includes 1, so the difference is not significant), and on the other hand smaller differences may be (e.g. Glycin is significantly higher in the GMO than in the conventional counterpart, with a point estimate of only +3.5%). Note that the significance tests are based on a standard error of differences (see Table 3) which is calculated from the residual variance as *seddiff* (see Table 2) in the programme code, where $seddif = \sqrt{V_0\left(\frac{1}{12} + \frac{1}{10}\right)}$, where 12 and 10 are the number of replications in this experiment for GMO and conventional counterpart, respectively. The number of degrees of freedom estimated by the Kenward-Roger method varies between 38.7 (16:1 Palmitoleic) and 54.6 (Ash).

In total twenty-three analytes were found to have a significant difference between GMO and its conventional counterpart (which is 43% of the 53 investigated analytes). These analytes are shown in blue (or in black or red if there was also a potential equivalence problem) in Figures 3 and 4, and boxplot representations of these data are shown in Figures 5-7 to assist further interpretation. Note, however, that these boxplot representations ignore some aspects of the model, such as site and replication variation, and furthermore display data on the natural, untransformed scale.

Regarding the estimated equivalence limits, although conceptually there is a single pair of limits (calculated using the methods in Section 3.3.2), here three methods of display of these are presented in order to aid interpretation. Thus, the limits are calculated on three different scales; each scale is useful for a specific purpose.

i. The first scale is the natural, raw, untransformed scale, which allows easy recognition and facilitates initial data exploration by biologists prior to analysis:

exp ($m_R \pm lsd$(GR;2;97.5)) For instance, Niacin has equivalence limits which, when back-transformed (effectively as geometric means) onto this natural scale give values of [16.1, 27.1]. These limits are shown in the boxplots (Figures 5-7).

ii. The second scale is the ratio scale where the GMO is compared to the mean of the reference (commercial) varieties, in an equivalence test. This may be presented independently of any difference test between the GMO and the conventional counterpart, once again through back-transformed values (effectively as geometric means):

exp ($m_G - m_R \pm lsd$(GR;2;97.5))

(see Table 4). This scale provides the most direct view whether the difference between GMO and reference (commercial) varieties is significant (it is significant if the interval does not contain the value 1). This scale is therefore best for distinguishing between equivalence categories (ii) and (iii), as defined in Section 4.1. For Niacin the equivalence interval on this scale is [0.59, 0.99], so indeed the difference is significant and non-equivalence is more likely than not.

iii. Finally, the equivalence interval can be expressed on the same scale where the GMO is compared to the conventional counterpart. This is the presentation method recommended to applicants in Section 4.1 and which allows for the need to adjust equivalence limits, so that both the difference test and the equivalence test may be visually assessed by the confidence limits for the difference test alone. The adjusted equivalence limits are:

$$\exp((m_G - m_C) + \{[(m_R - m_G) \pm lsd(GR;2;97.5)]\, lsd(GC;1;95) / lsd(GR;1;95)\})$$

(see Figures 3 and 4); this scale allows a simultaneous presentation of the results for both the comparison of GMO with conventional counterpart and the comparison of GMO with the commercial reference lines. Therefore it is the easiest scale for performing a test of equivalence by the graphical equivalent of the TOST procedure advocated in this opinion (see Figures 3 and 4). This scale is best for distinguishing between equivalence categories (i) and (ii), and, similarly when considering confidence intervals completely outside the equivalence limits, for distinguishing between equivalence categories (iii) and (iv) (see Section 4.1). For the example of Niacin the equivalence interval on this scale is [0.88, 1.20] (see Table 5). This equivalence interval overlaps with the confidence interval for the comparison of the GMO with its conventional counterpart (which is [0.84, 0.90], see Table 3), therefore neither equivalence nor non-equivalence may be proven for this analyte.

In any case, the three intervals are just adjusted versions of each other and completely equivalent for statistical testing as explained more fully in Section 2.2.1. In the current example two cases were found where there was a statistically significant difference between the GMO and the references (16:0 Palmitic and Niacin). For these analytes non-equivalence is more likely than not. For further interpretation, boxplots are given in Figure 7. It can be seen that for 16:0 Palmitic, both the GMO and

the conventional counterpart are higher than the reference range, therefore on this single endpoint GMO and conventional counterpart seem to present the same potential health hazards, if any. It is outside the scope of this document to discuss the risk assessment of such cases. For Niacin the situation is different. Niacin is found 24% to be lower in the GMO than on average in the commercial varieties, and the result is also significantly lower (by 13%) than what is found for the conventional counterpart.

A problem occurs when the variance component between commercial genotypes is estimated to be zero. In the current example dataset this occurred with Ash and Phytic Acid. In these cases the calculation of standard errors of difference will be based on the assumption that there is no variation between the commercial genotypes and standard errors and degrees of freedom are derived from a model which omits the random factor for genotypes. This is not a truly believable model, so the equivalence intervals calculated are typically too narrow and must not be used (see Figure 7).

Accepting the calculated equivalence limits as null hypothesis values in a test of equivalence for the remaining 49 analytes leads to the conclusion that 44 are proven to be equivalent to the reference varieties, whereas for 5 (Lysine, Phosphorus, Potassium, Vitamin B6 and Vitamin E) the equivalence is more likely than not, but not strictly proven at the 95 % confidence level. For further interpretation, boxplots for these 5 cases are given in Figure 6.

A small simulation was performed to investigate whether the observed number of significant differences between GMO and conventional counterpart (23) is large under the null hypothesis that variation between genotypes can be described by a normal distribution with variance $V_g$ on the logarithmic scale. Here we take for $V_g$ the quantifications as obtained with the mixed model (Table 2). Under this null hypothesis and ignoring further estimation error, differences $d$ between any two varieties would have a normal distribution with variance $2V_g$. In 1000 iterations random values for $d$ were sampled from this distribution for all analytes, and a two-sided t test at the 95% confidence level was performed assuming that the *seddif* and *dfdif* from the actual experiment were appropriate characterisations of residual error. Over the 1000 iterations the average number of significant test results was 36 (approximate 95% confidence interval [30, 42]). Therefore, under a null hypothesis describing equivalence between all the varieties, the observed number of significant differences between GMO and conventional counterpart (23) is relatively small and no source of concern in itself.

Differences between GMO and conventional counterpart may not be constant over sites. This was investigated by fitting additional fixed terms coded as ref_aside.site and ref_aside.genotypegroup.site in the mixed model 1, and by performing a Wald test to obtain a $p$ value for the significance of the latter term. For 8 analytes the genotype by environment (GxE) interaction was significant ($p<0.05$), and tables with geometric means for these cases are reported in Table 6 as a help in further interpretation of the results and risk assessment. Table 7 presents all means and standard errors of means per site (on the transformed scale).

In Table 8 the outcomes are classified according to the outcome types and categories as proposed in this opinion. Apart from 2 analytes for which equivalence limits could not well be established, there are 44 analytes in category (i, Equivalence), 5 in category (ii, Equivalence more likely than not), 2 in category (iii, Non-equivalence more likely than not), and none in category (iv, Non-equivalence).

The conclusions drawn for this dataset can be summarised as follows:

1. Twenty-three analytes show statistically significant differences (at the 90% confidence level) between GMO and conventional counterpart. The differences varied between -13% and +14%. The number of significant results is not a reason of concern considering simulation results allowing for natural background variation.

2. For two analytes, 16:0 Palmitic and Niacin, a statistically significant deviation (at the 95% confidence level) from the reference lines has been found, and non-equivalence is more likely than not. Further evaluation is required.

3. For five analytes, Lysine, Phosphorus, Potassium, Vitamin B6 and Vitamin E, equivalence is more likely than not, but a strict proof of equivalence cannot be given. Further evaluation may be required.

4. For two analytes, Ash and Phytic Acid, no proper conclusion on equivalence can be formulated because of lack of observable natural variation in the commercial varieties. Further evaluation may be required.

5. For fourty-four analytes (including 20 with significant differences between GMO and conventional counterpart) equivalence is established in a formal test of equivalence (at the 95% confidence level) using the estimated equivalence limits.

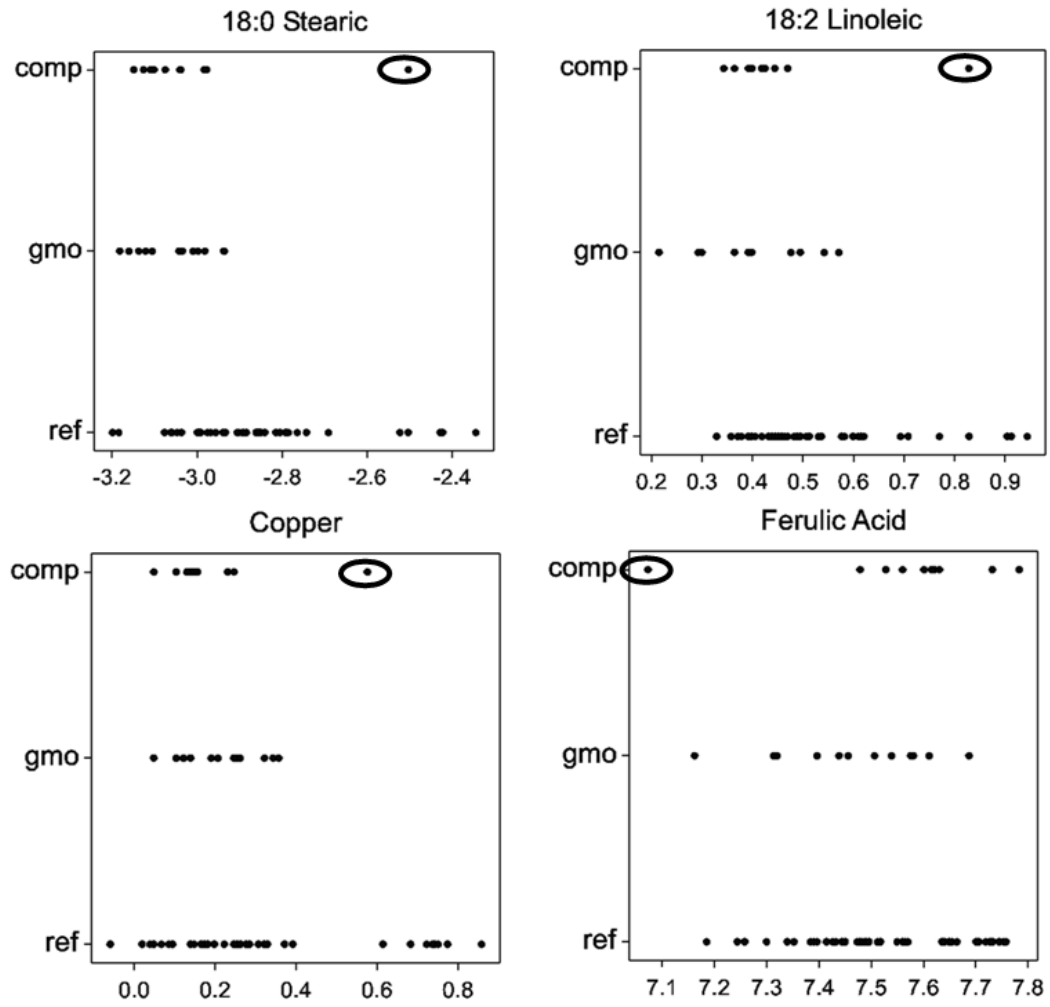


Figure 2. log10 of results for four analytes, grouped by genotypic group (comp=comparator, i.e. the conventional counterpart, gmo=GMO, ref=reference). Circles indicate visually identified outliers.

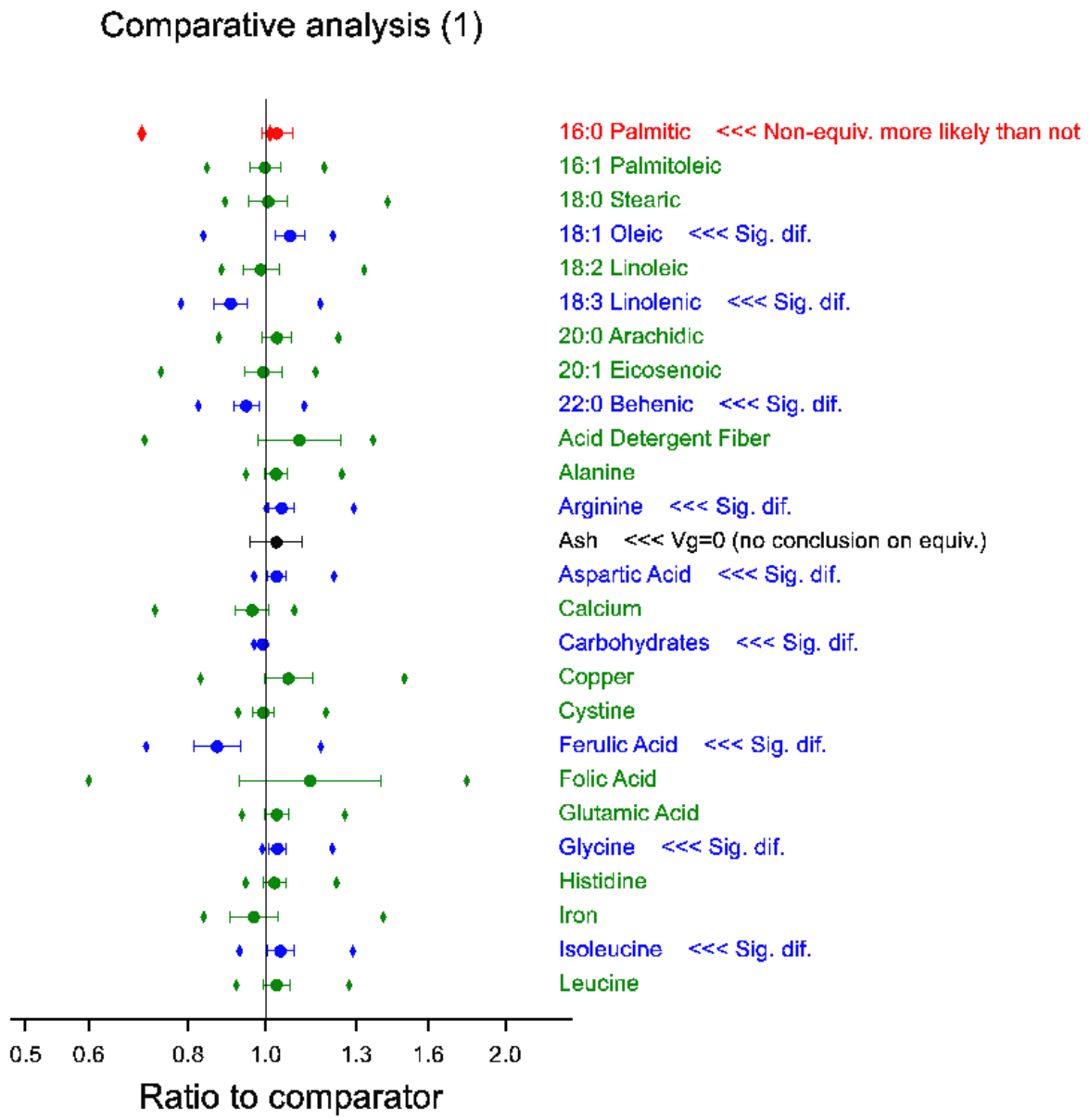


Figure 3. Part 1 of overview example comparative analysis. Circles and bars represent point estimate and 90 % confidence interval for ratio GMO to conventional counterpart. Diamonds represent adjusted equivalence limits based on reference varieties. Colours represent different types of outcome. Green: 1; Blue: 2; Black: 3-4 and cases with genotype variance (Vg) estimated zero; Red: 5-7.

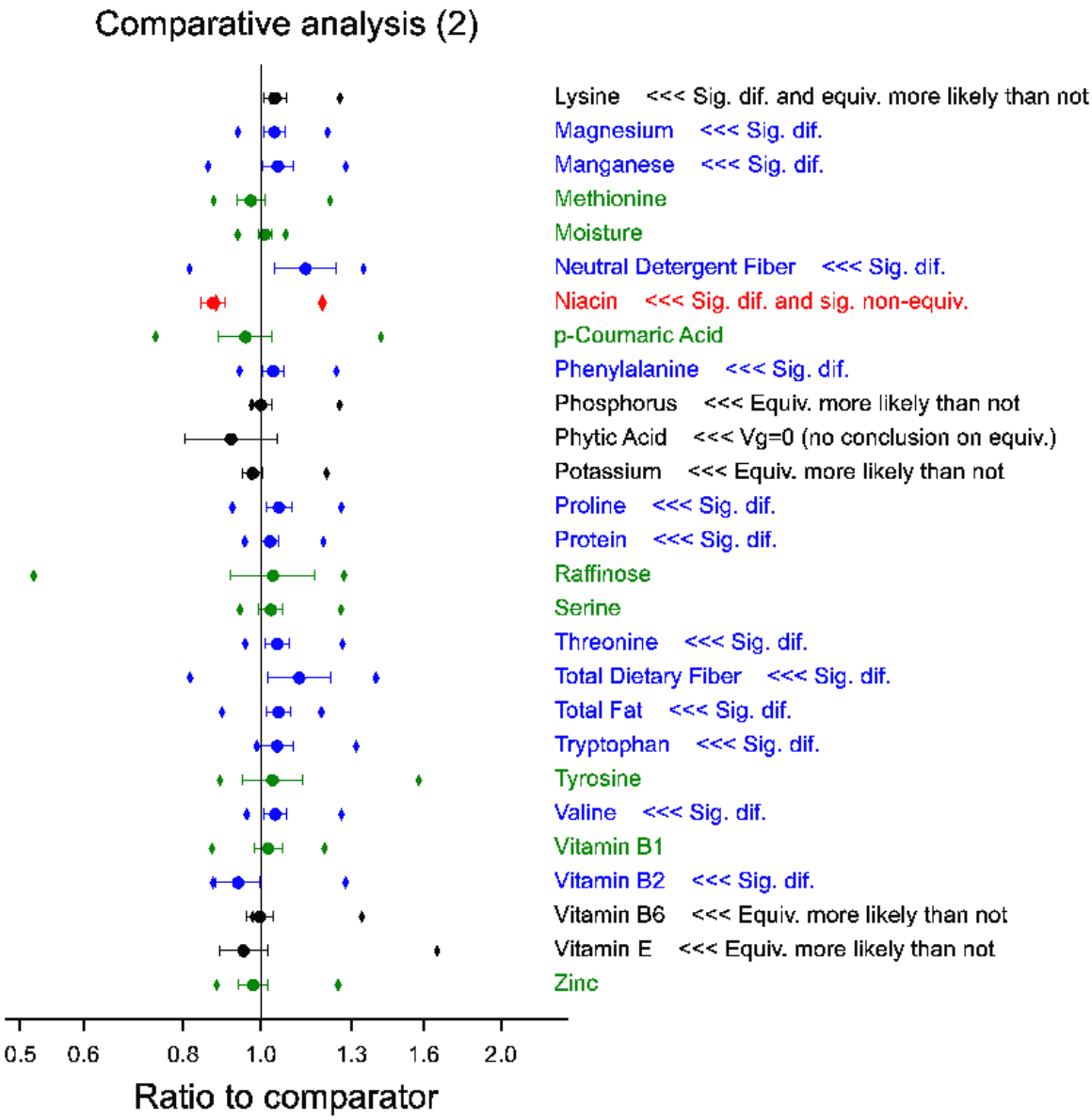


Figure 4. Part 2 of overview example comparative analysis. Circles and bars represent point estimate and 95 % confidence interval for ratio GMO to conventional counterpart. Diamonds represent adjusted equivalence limits based on reference varieties. Colours represent different types of outcome. Green: 1; Blue: 2; Black: 3-4 and cases with genotype variance (Vg) estimated zero; Red: 5-7.

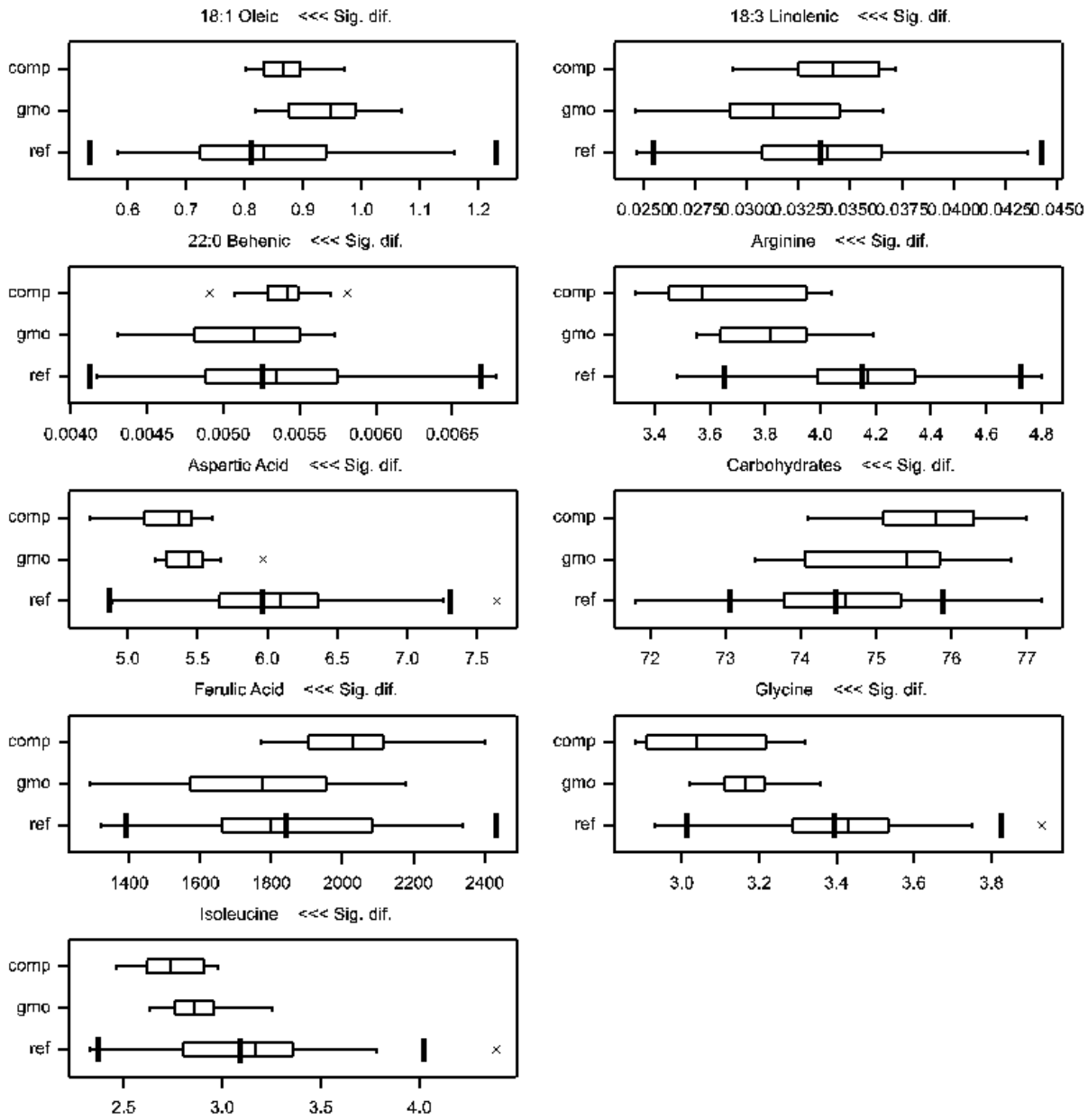


Figure 5a**.** Boxplots (part 1) for cases with significant differences to comparator (i.e. the conventional counterpart), but equivalence to commercial varieties. Note that these are presented on the raw, natural scale and that therefore no adjustment to the equivalence limits is required. The boxplots comprise schematic box-and-whisker diagrams. Each box extends from the lower to the upper quartile (p25 to p75) and the line in the middle is the median (p50). The whiskers extend to extreme data points (minimum and maximum), unless points are farther away from the quartiles than 1.5 times the box length, in which case the points are shown separately as crosses and the whiskers only cover the remaining points. comp=comparator; gmo=GMO; ref= reference lines. Additional thicker bars in the boxplot for references represent geometric mean and calculated equivalence limits.

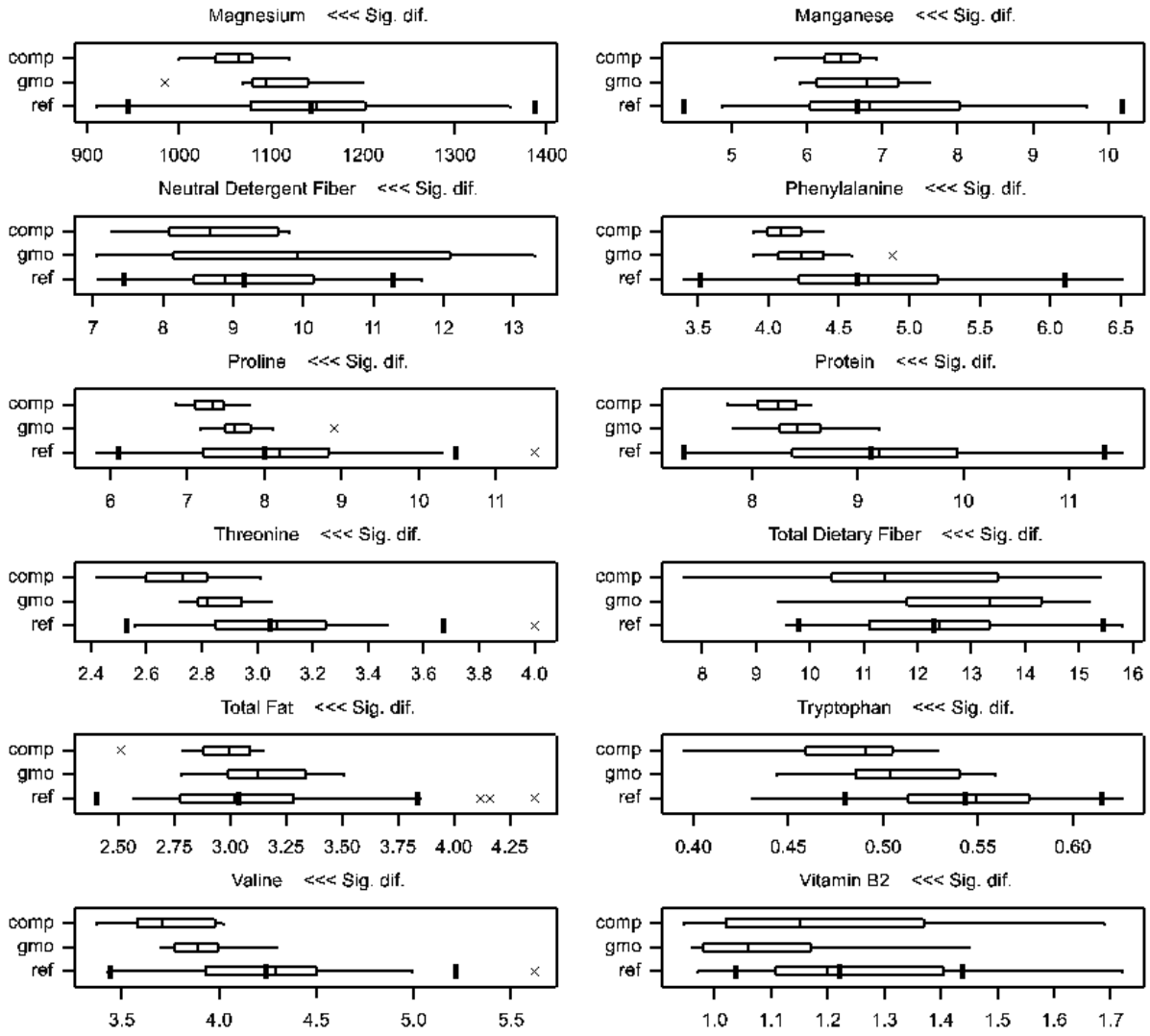


Figure 5b. Boxplots (part 2) for cases with significant differences to comparator(i.e. the conventional counterpart), but equivalence to commercial varieties. Note that these are presented on the raw, natural scale and that therefore no adjustment to the equivalence limits is required. The boxplots comprise schematic box-and-whisker diagrams. Each box extends from the lower to the upper quartile (p25 to p75) and the line in the middle is the median (p50). The whiskers extend to extreme data points (minimum and maximum), unless points are farther away from the quartiles than 1.5 times the box length, in which case the points are shown separately as crosses and the whiskers only cover the remaining points. comp=comparator; gmo=GMO; ref= reference lines. Additional thicker bars in the boxplot for references represent geometric mean and calculated equivalence limits.

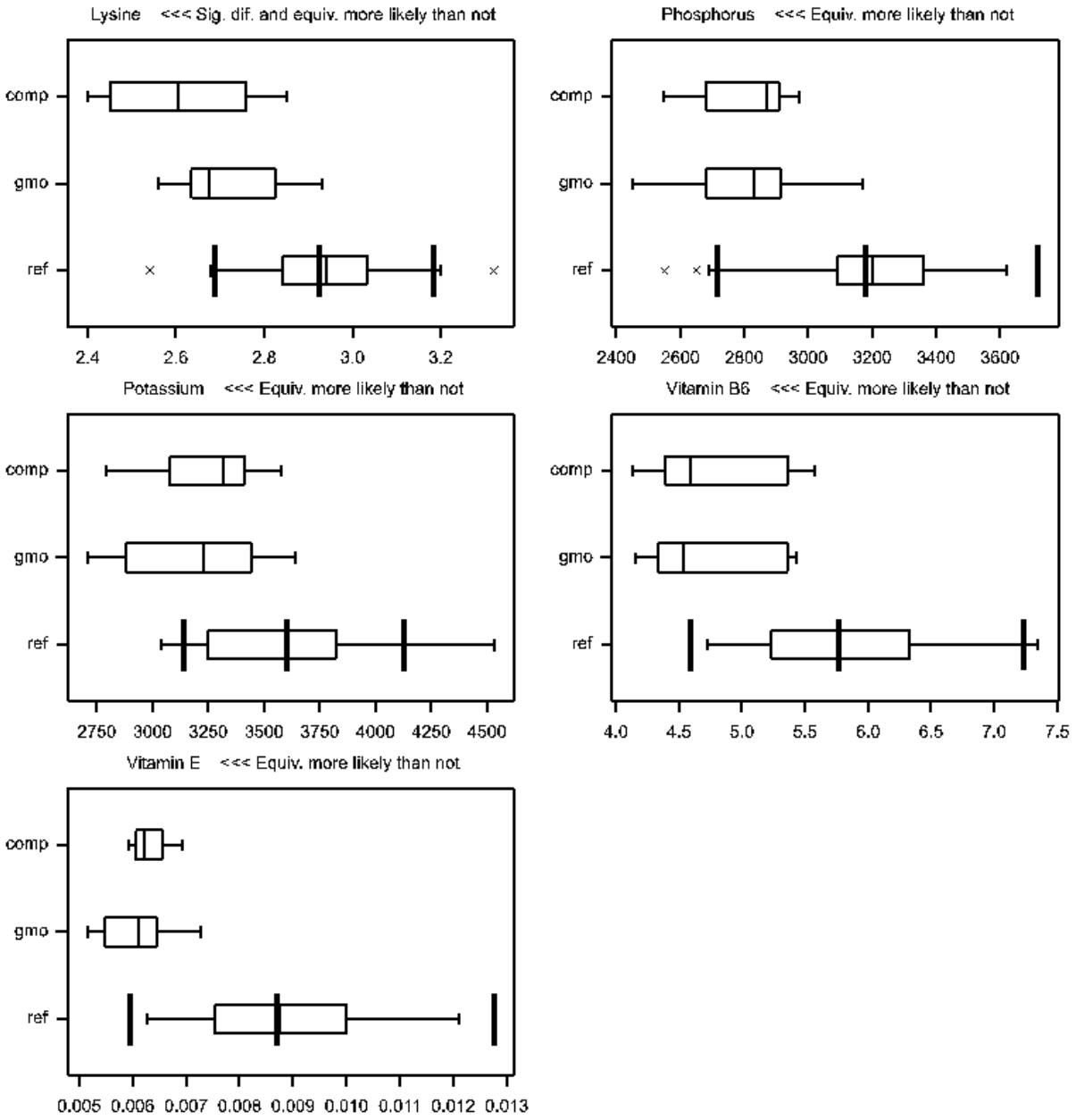


Figure 6. Boxplots for cases with equivalence more likely than not, but unproven (category (ii)). Note that these are presented on the raw, natural scale and that therefore no adjustment to the equivalence limits is required. The boxplots comprise schematic box-and-whisker diagrams. Each box extends from the lower to the upper quartile (p25 to p75) and the line in the middle is the median (p50). The whiskers extend to extreme data points (minimum and maximum), unless points are farther away from the quartiles than 1.5 times the box length, in which case the points are shown separately as crosses and the whiskers only cover the remaining points. comp=comparator; gmo=GMO; ref= reference lines. Additional thicker bars in the boxplot for references represent geometric mean and calculated equivalence limits.

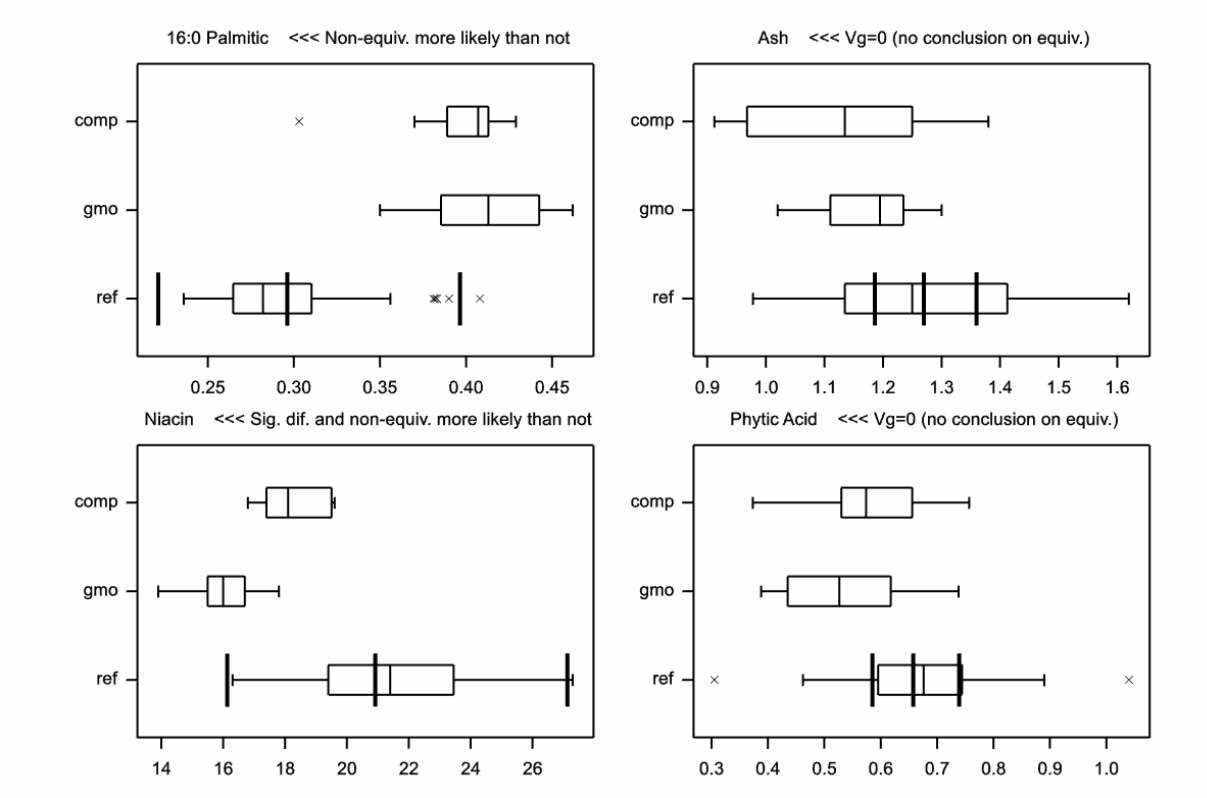


Figure 7. Boxplots for cases with non-equivalence more likely than not (category (iii)) or an impossibility to judge equivalence due to a zero estimate for variance of genotypes. Note that these are presented on the raw, natural scale and that therefore no adjustment to the equivalence limits is required. The boxplots comprise schematic box-and-whisker diagrams. Each box extends from the lower to the upper quartile (p25 to p75) and the line in the middle is the median (p50). The whiskers extend to extreme data points (minimum and maximum), unless points are farther away from the quartiles than 1.5 times the box length, in which case the points are shown separately as crosses and the whiskers only cover the remaining points. comp=comparator; gmo=GMO; ref= reference lines. Additional thicker bars in the boxplot for references represent geometric mean and calculated equivalence limits.

**Table 1:** Geometric means for comparator (*Gmcomp*=exp($m_C$)), GMO (*Gmgmo*=exp($m_G$)) and commercial varieties (*Gmref*=exp($m_R$)), where the coding notation *Gmcomp is* used to represent the back-transformed geometric mean of the comparator, etc

| Analyte | Gmcomp | Gmgmo | Gmref |
|---|---|---|---|
| 16:0 Palmitic | 0.396 | 0.409 | 0.296 |
| 16:1 Palmitoleic | 0.004 | 0.004 | 0.004 |
| 18:0 Stearic | 0.047 | 0.047 | 0.056 |
| 18:1 Oleic | 0.871 | 0.935 | 0.812 |
| 18:2 Linoleic | 1.512 | 1.492 | 1.675 |
| 18:3 Linolenic | 0.034 | 0.031 | 0.034 |
| 20:0 Arachidic | 0.013 | 0.013 | 0.013 |
| 20:1 Eicosenoic | 0.011 | 0.011 | 0.01 |
| 22:0 Behenic | 0.005 | 0.005 | 0.005 |
| Acid Detergent Fiber | 3.52 | 3.884 | 3.523 |
| Alanine | 6.172 | 6.366 | 6.999 |
| Arginine | 3.641 | 3.816 | 4.153 |
| Ash | 1.13 | 1.167 | 1.27 |
| Aspartic Acid | 5.281 | 5.453 | 5.967 |
| Calcium | 51.015 | 49.108 | 42.441 |
| Carbohydrates | 75.683 | 75.084 | 74.458 |
| Copper | 1.161 | 1.242 | 1.322 |
| Cystine | 1.699 | 1.689 | 1.819 |
| Ferulic Acid | 2007.963 | 1746.716 | 1840.686 |
| Folic Acid | 0.543 | 0.618 | 0.573 |
| Glutamic Acid | 15.536 | 16.056 | 17.57 |
| Glycine | 3.063 | 3.172 | 3.395 |
| Histidine | 2.389 | 2.452 | 2.63 |
| Iron | 17.11 | 16.539 | 18.846 |
| Isoleucine | 2.747 | 2.869 | 3.088 |
| Leucine | 10.231 | 10.562 | 11.57 |
| Lysine | 2.602 | 2.715 | 2.925 |

| Analyte | Gmcomp | Gmgmo | Gmref |
|---|---|---|---|
| Magnesium | 1060.466 | 1103.888 | 1144.278 |
| Manganese | 6.377 | 6.705 | 6.67 |
| Methionine | 1.767 | 1.718 | 1.889 |
| Moisture | 11.94 | 12.093 | 11.973 |
| Neutral Detergent Fiber | 8.629 | 9.826 | 9.166 |
| Niacin | 18.241 | 15.9 | 20.915 |
| p-Coumaric Acid | 154.488 | 147.921 | 165.492 |
| Phenylalanine | 4.102 | 4.255 | 4.631 |
| Phosphorus | 2799.113 | 2803.565 | 3177.435 |
| Phytic Acid | 0.57 | 0.523 | 0.658 |
| Potassium | 3242.469 | 3170.599 | 3603.606 |
| Proline | 7.29 | 7.69 | 7.997 |
| Protein | 8.222 | 8.449 | 9.132 |
| Raffinose | 0.113 | 0.118 | 0.09 |
| Serine | 4.119 | 4.242 | 4.628 |
| Threonine | 2.724 | 2.858 | 3.046 |
| Total Dietary Fiber | 11.448 | 12.801 | 12.301 |
| Total Fat | 2.979 | 3.138 | 3.038 |
| Tryptophan | 0.481 | 0.505 | 0.543 |
| Tyrosine | 2.674 | 2.768 | 3.152 |
| Valine | 3.739 | 3.902 | 4.239 |
| Vitamin B1 | 0.344 | 0.352 | 0.352 |
| Vitamin B2 | 1.177 | 1.103 | 1.221 |
| Vitamin B6 | 4.736 | 4.723 | 5.764 |
| Vitamin E | 0.006 | 0.006 | 0.009 |
| Zinc | 19.535 | 19.111 | 21.338 |

**Table 2:** Variance components for random terms in mixed model: genotype (Varg), site (Vars), replication within site (Varr) and residual (Var0); see description of mixed statistical model in section

| Analyte | Varg | Vars | Varr | Var0 |
|---|---|---|---|---|
| 16:0 Palmitic | 0.016 | 0.0036 | 0 | 0.0038 |
| 16:1 Palmitoleic | 0.010 | 0 | 0.0009 | 0.0038 |
| 18:0 Stearic | 0.027 | 0.0007 | 0.0008 | 0.0055 |
| 18:1 Oleic | 0.033 | 0.0003 | 0.0003 | 0.0035 |
| 18:2 Linoleic | 0.012 | 0.0024 | 0.0009 | 0.0049 |
| 18:3 Linolenic | 0.013 | 0.0031 | 0.0009 | 0.0047 |
| 20:0 Arachidic | 0.010 | 0.0015 | 0.0008 | 0.0035 |
| 20:1 Eicosenoic | 0.018 | 0.0035 | 0.0006 | 0.0057 |
| 22:0 Behenic | 0.011 | 0.0030 | 0.0005 | 0.0026 |
| Acid Detergent Fiber | 0.007 | 0.0004 | 0 | 0.0271 |
| Alanine | 0.012 | 0.0018 | 0.0011 | 0.0020 |
| Arginine | 0.003 | 0.0003 | 0.0003 | 0.0025 |
| Ash | 0 | 0.0059 | 0.0005 | 0.0110 |
| Aspartic Acid | 0.008 | 0.0005 | 0.0006 | 0.0014 |
| Calcium | 0.026 | 0.0122 | 0 | 0.0043 |
| Carbohydrates | 0.000 | 0.0002 | 0.0000 | 0.0000 |
| Copper | 0.039 | 0 | 0 | 0.0092 |
| Cystine | 0.005 | 0.0003 | 0.0010 | 0.0019 |
| Ferulic Acid | 0.013 | 0.0043 | 0.0005 | 0.0081 |
| Folic Acid | 0.013 | 0 | 0.0014 | 0.0797 |
| Glutamic Acid | 0.015 | 0.0018 | 0.0015 | 0.0023 |
| Glycine | 0.002 | 0.0002 | 0.0004 | 0.0013 |
| Histidine | 0.007 | 0 | 0.0007 | 0.0019 |
| Iron | 0.014 | 0.0122 | 0 | 0.0094 |
| Isoleucine | 0.013 | 0.0001 | 0.0015 | 0.0029 |
| Leucine | 0.019 | 0.0041 | 0.0018 | 0.0027 |
| Lysine | 0.001 | 0.0002 | 0 | 0.0023 |

| Analyte | Varg | Vars | Varr | Var0 |
|---|---|---|---|---|
| Magnesium | 0.007 | 0.0003 | 0.0000 | 0.0019 |
| Manganese | 0.034 | 0.0024 | 0.0009 | 0.0039 |
| Methionine | 0.013 | 0.0020 | 0.0018 | 0.0031 |
| Moisture | 0.001 | 0.0071 | 0 | 0.0007 |
| Neutral Detergent Fiber | 0.004 | 0.0018 | 0.0016 | 0.0153 |
| Niacin | 0.013 | 0.0021 | 0.0004 | 0.0025 |
| p-Coumaric Acid | 0.057 | 0.0060 | 0.0000 | 0.0113 |
| Phenylalanine | 0.015 | 0.0022 | 0.0013 | 0.0020 |
| Phosphorus | 0.004 | 0.0026 | 0 | 0.0020 |
| Phytic Acid | 0 | 0.0003 | 0.0095 | 0.0319 |
| Potassium | 0.003 | 0.008 | 0 | 0.0015 |
| Proline | 0.014 | 0 | 0.0020 | 0.0026 |
| Protein | 0.009 | 0.0004 | 0.0007 | 0.0013 |
| Raffinose | 0.043 | 0.0178 | 0 | 0.0284 |
| Serine | 0.009 | 0.0034 | 0.0014 | 0.0023 |
| Threonine | 0.006 | 0 | 0.0008 | 0.0024 |
| Total Dietary Fiber | 0.005 | 0.0025 | 0.0009 | 0.0158 |
| Total Fat | 0.010 | 0.0017 | 0.0004 | 0.0022 |
| Tryptophan | 0.002 | 0.0003 | 0.0003 | 0.0044 |
| Tyrosine | 0.011 | 0.0032 | 0.0006 | 0.0147 |
| Valine | 0.008 | 0 | 0.0009 | 0.0021 |
| Vitamin B1 | 0.008 | 0.000 | 0.0006 | 0.0032 |
| Vitamin B2 | 0.003 | 0.0180 | 0 | 0.0078 |
| Vitamin B6 | 0.009 | 0.0093 | 0.0003 | 0.0028 |
| Vitamin E | 0.026 | 0.0003 | 0.0001 | 0.0097 |
| Zinc | 0.014 | 0.0019 | 0.0002 | 0.0034 |

**Table 3:** Assessment of differences GMO vs. comparator. sed(GC;1;95) and lsd(GC;1;95) are on ln scale, ratio, and 90% confidence limits (low and upp) are back-transformed. See description of mixed statistical model in section 5.1 and see Figures 3 and 4 for graphical representation.

| Analyte | ratio | low | upp | sed(GC;1;95) | df(GC;1) | lsd(GC;1;95) |
|---|---|---|---|---|---|---|
| 16:0 Palmitic | 1.034 | 0.9887 | 1.081 | 0.02658 | 49.3 | 0.04455 |
| 16:1 Palmitoleic | 0.999 | 0.9558 | 1.045 | 0.02638 | 38.7 | 0.04445 |
| 18:0 Stearic | 1.008 | 0.9538 | 1.066 | 0.03307 | 41.7 | 0.05562 |
| 18:1 Oleic | 1.074 | 1.0294 | 1.121 | 0.02538 | 42.1 | 0.04268 |
| 18:2 Linoleic | 0.987 | 0.9366 | 1.04 | 0.03121 | 42.2 | 0.0525 |
| 18:3 Linolenic | 0.904 | 0.8606 | 0.95 | 0.02942 | 39.9 | 0.04954 |
| 20:0 Arachidic | 1.034 | 0.9906 | 1.079 | 0.02545 | 41.5 | 0.04282 |
| 20:1 Eicosenoic | 0.995 | 0.9419 | 1.051 | 0.03253 | 41.4 | 0.05473 |
| 22:0 Behenic | 0.946 | 0.9117 | 0.981 | 0.0218 | 41.4 | 0.03668 |
| Acid Detergent | 1.103 | 0.9805 | 1.241 | 0.07046 | 51.9 | 0.118 |
| Alanine | 1.031 | 0.9986 | 1.065 | 0.0192 | 41.4 | 0.03231 |
| Arginine | 1.048 | 1.0107 | 1.087 | 0.02153 | 41.9 | 0.03622 |
| Ash | 1.033 | 0.958 | 1.114 | 0.04498 | 54.6 | 0.07527 |
| Aspartic Acid | 1.033 | 1.0049 | 1.061 | 0.0161 | 41.4 | 0.02709 |
| Calcium | 0.963 | 0.9184 | 1.009 | 0.02801 | 48.5 | 0.04697 |
| Carbohydrates | 0.992 | 0.9875 | 0.997 | 0.00274 | 42 | 0.00461 |
| Copper | 1.069 | 0.9963 | 1.148 | 0.04221 | 51 | 0.07072 |
| Cystine | 0.994 | 0.9636 | 1.026 | 0.01863 | 40.1 | 0.03137 |
| Ferulic Acid | 0.87 | 0.8132 | 0.931 | 0.04008 | 41.8 | 0.06743 |
| Folic Acid | 1.138 | 0.929 | 1.395 | 0.12096 | 44.1 | 0.20323 |
| Glutamic Acid | 1.033 | 0.9984 | 1.07 | 0.02055 | 41.4 | 0.03458 |
| Glycine | 1.035 | 1.0084 | 1.063 | 0.01571 | 40.2 | 0.02644 |
| Histidine | 1.026 | 0.994 | 1.06 | 0.01903 | 40.7 | 0.03203 |
| Iron | 0.967 | 0.9015 | 1.036 | 0.04159 | 49.4 | 0.06972 |
| Isoleucine | 1.044 | 1.0045 | 1.086 | 0.0232 | 41.6 | 0.03903 |
| Leucine | 1.032 | 0.9942 | 1.072 | 0.02241 | 41.3 | 0.03771 |
| Lysine | 1.043 | 1.0084 | 1.08 | 0.02037 | 48.8 | 0.03416 |
| Magnesium | 1.041 | 1.0089 | 1.074 | 0.01863 | 42.6 | 0.03132 |
| Manganese | 1.051 | 1.0047 | 1.1 | 0.02702 | 41.1 | 0.04546 |
| Methionine | 0.972 | 0.9338 | 1.012 | 0.02398 | 41.1 | 0.04035 |
| Moisture | 1.013 | 0.9938 | 1.032 | 0.0113 | 49.5 | 0.01894 |
| NeutralDeterFiber | 1.139 | 1.0413 | 1.245 | 0.05322 | 43.2 | 0.08946 |
| Niacin | 0.872 | 0.8405 | 0.904 | 0.0216 | 41.7 | 0.03633 |
| p-Coumaric Acid | 0.957 | 0.8867 | 1.034 | 0.04565 | 42.1 | 0.07677 |
| Phenylalanine | 1.037 | 1.0043 | 1.071 | 0.01921 | 41.4 | 0.03232 |
| Phosphorus | 1.002 | 0.9698 | 1.034 | 0.01923 | 48.5 | 0.03224 |
| Phytic Acid | 0.919 | 0.8042 | 1.049 | 0.07949 | 53.2 | 0.13307 |
| Potassium | 0.978 | 0.9509 | 1.006 | 0.01665 | 47.9 | 0.02792 |
| Proline | 1.055 | 1.0166 | 1.095 | 0.02199 | 41.9 | 0.037 |
| Protein | 1.028 | 1.001 | 1.055 | 0.01566 | 41.2 | 0.02635 |
| Raffinose | 1.036 | 0.9177 | 1.17 | 0.07232 | 49.1 | 0.12125 |
| Serine | 1.03 | 0.9944 | 1.067 | 0.02087 | 41.1 | 0.03512 |
| Threonine | 1.049 | 1.0131 | 1.087 | 0.02094 | 41.5 | 0.03523 |
| Total Dietary | 1.118 | 1.0211 | 1.225 | 0.05404 | 41.2 | 0.09092 |
| Total Fat | 1.054 | 1.0181 | 1.09 | 0.02035 | 41.9 | 0.03422 |
| Tryptophan | 1.049 | 1.0002 | 1.101 | 0.02851 | 44 | 0.04791 |
| Tyrosine | 1.035 | 0.9483 | 1.13 | 0.05215 | 44.4 | 0.0876 |
| Valine | 1.043 | 1.0096 | 1.078 | 0.0196 | 41.8 | 0.03296 |
| Vitamin B1 | 1.023 | 0.9818 | 1.065 | 0.02433 | 43.3 | 0.0409 |
| Vitamin B2 | 0.938 | 0.8799 | 0.999 | 0.03784 | 49 | 0.06344 |
| Vitamin E | 0.953 | 0.8874 | 1.023 | 0.04217 | 42.5 | 0.0709 |
| Vitamin B6 | 0.997 | 0.9597 | 1.037 | 0.0229 | 41.6 | 0.03853 |
| Zinc | 0.978 | 0.9381 | 1.02 | 0.02495 | 42.5 | 0.04195 |

**Table 4:** 95% Equivalence limits (low and upp) calculated on the scale of the ratio of GMO to reference mean. The point estimate of this ratio itself is given in the column ratioThe width of the interval depends on the standard error of difference for equivalence (sed(GR;2;97.5), given on logarithmic scale), and the degrees of freedom for equivalence (df(GR;2)) calculated by the Kenward-Roger method See text in section 5.1 for further explanation.[1]

| Analyte | ratio | low | upp | sed(GR;2;97.5) | df(GR;2) | lsd(GR;2;97.5) |
|---|---|---|---|---|---|---|
| 16:0 Palmitic | 1.3811 | 1.0317 | 1.849 | 0.1317 | 10.5 | 0.2917 |
| 16:1 Palmitoleic | 0.9982 | 0.7841 | 1.271 | 0.1055 | 8.4 | 0.2414 |
| 18:0 Stearic | 0.8355 | 0.57 | 1.225 | 0.1736 | 10.9 | 0.3824 |
| 18:1 Oleic | 1.1519 | 0.7595 | 1.747 | 0.1897 | 11.2 | 0.4165 |
| 18:2 Linoleic | 0.8908 | 0.6886 | 1.152 | 0.1155 | 10 | 0.2574 |
| 18:3 Linolenic | 0.9234 | 0.7009 | 1.216 | 0.1221 | 9.1 | 0.2757 |
| 20:0 Arachidic | 0.9931 | 0.7816 | 1.262 | 0.1075 | 10 | 0.2395 |
| 20:1 Eicosenoic | 1.11 | 0.8068 | 1.527 | 0.1432 | 10 | 0.3191 |
| 22:0 Behenic | 0.9766 | 0.7672 | 1.243 | 0.1089 | 10.4 | 0.2414 |
| Acid Detergent Fiber | 1.1026 | 0.8373 | 1.452 | 0.1027 | 4.4 | 0.2752 |
| Alanine | 0.9096 | 0.7052 | 1.173 | 0.1154 | 10.8 | 0.2545 |
| Arginine | 0.9188 | 0.8077 | 1.045 | 0.0558 | 7.9 | 0.1289 |
| Ash [1] | 0.9187 | 0.858 | 0.984 | 0.0341 | 53.6 | 0.0683 |
| Aspartic Acid | 0.9138 | 0.7455 | 1.12 | 0.0921 | 10.7 | 0.2035 |
| Calcium | 1.1571 | 0.7993 | 1.675 | 0.1673 | 10.6 | 0.3699 |
| Carbohydrates | 1.0084 | 0.9895 | 1.028 | 0.0083 | 8.2 | 0.019 |
| Copper | 0.9395 | 0.5947 | 1.484 | 0.2073 | 10.8 | 0.4572 |
| Cystine | 0.9286 | 0.7797 | 1.106 | 0.0779 | 9.5 | 0.1748 |
| Ferulic Acid | 0.9489 | 0.7179 | 1.254 | 0.1234 | 9 | 0.2791 |
| Folic Acid | 1.0786 | 0.6948 | 1.674 | 0.1515 | 3.6 | 0.4397 |
| Glutamic Acid | 0.9138 | 0.6908 | 1.209 | 0.127 | 10.9 | 0.2798 |
| Glycine | 0.9343 | 0.8292 | 1.053 | 0.0525 | 8.8 | 0.1193 |
| Histidine | 0.9325 | 0.7711 | 1.128 | 0.0853 | 10 | 0.1901 |
| Iron | 0.8776 | 0.6544 | 1.177 | 0.1288 | 8.6 | 0.2934 |
| Isoleucine | 0.9292 | 0.7137 | 1.21 | 0.1192 | 10.5 | 0.2638 |
| Leucine | 0.9129 | 0.6673 | 1.249 | 0.1422 | 10.9 | 0.3134 |
| Lysine | 0.9282 | 0.8532 | 1.01 | 0.0322 | 4.7 | 0.0843 |
| Magnesium | 0.9647 | 0.796 | 1.169 | 0.0863 | 10 | 0.1922 |
| Manganese | 1.0053 | 0.6585 | 1.535 | 0.1922 | 11 | 0.4231 |
| Methionine | 0.9095 | 0.7004 | 1.181 | 0.1177 | 10.3 | 0.2612 |
| Moisture | 1.01 | 0.9357 | 1.09 | 0.0336 | 8.6 | 0.0765 |
| Neutral Detergent Fiber | 1.072 | 0.8703 | 1.32 | 0.0778 | 4.4 | 0.2084 |
| Niacin | 0.7602 | 0.5862 | 0.986 | 0.1178 | 10.8 | 0.2599 |
| p-Coumaric Acid | 0.8938 | 0.5145 | 1.553 | 0.2498 | 10.6 | 0.5523 |
| Phenylalanine | 0.9187 | 0.6965 | 1.212 | 0.1258 | 11 | 0.2769 |
| Phosphorus | 0.8823 | 0.7537 | 1.033 | 0.0694 | 8.8 | 0.1575 |
| Phytic Acid [1] | 0.7958 | 0.7082 | 0.894 | 0.0581 | 52.2 | 0.1166 |
| Potassium | 0.8798 | 0.7675 | 1.009 | 0.0595 | 8.2 | 0.1366 |
| Proline | 0.9616 | 0.7333 | 1.261 | 0.1227 | 10.7 | 0.2711 |
| Protein | 0.9252 | 0.7454 | 1.148 | 0.098 | 10.8 | 0.2161 |
| Raffinose | 1.3123 | 0.7939 | 2.169 | 0.2218 | 8.9 | 0.5025 |
| Serine | 0.9167 | 0.7343 | 1.144 | 0.0998 | 10.2 | 0.2218 |
| Threonine | 0.9383 | 0.7788 | 1.13 | 0.0834 | 9.8 | 0.1863 |
| Total Dietary Fiber | 1.0407 | 0.8284 | 1.307 | 0.087 | 4.7 | 0.2281 |
| Total Fat | 1.033 | 0.8173 | 1.306 | 0.1062 | 10.8 | 0.2342 |
| Tryptophan | 0.9292 | 0.8205 | 1.052 | 0.0487 | 5.1 | 0.1244 |
| Tyrosine | 0.8782 | 0.6674 | 1.156 | 0.117 | 7.3 | 0.2745 |
| Valine | 0.9203 | 0.7477 | 1.133 | 0.0936 | 10.3 | 0.2078 |
| Vitamin B1 | 0.9999 | 0.8059 | 1.241 | 0.0972 | 10.3 | 0.2157 |
| Vitamin B2 | 0.9036 | 0.7671 | 1.064 | 0.0611 | 4.4 | 0.1637 |
| Vitamin B6 | 0.8194 | 0.6529 | 1.028 | 0.1025 | 10.4 | 0.2272 |
| Vitamin E | 0.6908 | 0.4718 | 1.011 | 0.1709 | 9.9 | 0.3813 |
| Zinc | 0.8956 | 0.6814 | 1.177 | 0.1225 | 9.9 | 0.2734 |

Confidence intervals not trustworthy, because the estimate of the variance between commercial genotypes was 0 and *sed* is based on lower strata (note also the high df)

**Table 5:** Assessment of equivalences on the scale GMO vs. comparator. seds and lsds are on ln scale, adjusted equivalence limits are back-transformed. See description of mixed statistical model in section 5.1 and see Figures 3 and 4 for graphical representation.

| Analyte | sed(GR;1;95) | df(GR;1) | lsd(GR;1;95) | lsd(GC;1;95)/ lsd(GR;1;95) | Eq.limit low adjusted | Eq.limit upp adjusted |
|---|---|---|---|---|---|---|
| 16:0 Palmitic | 0.04038 | 17.9 | 0.07004 | 0.6361 | 0.6993 | 1.013 |
| 16:1 Palmitoleic | 0.03628 | 16.5 | 0.06322 | 0.7032 | 0.8444 | 1.186 |
| 18:0 Stearic | 0.05214 | 17.1 | 0.09067 | 0.6135 | 0.8904 | 1.424 |
| 18:1 Oleic | 0.0541 | 14.5 | 0.09505 | 0.449 | 0.8362 | 1.215 |
| 18:2 Linoleic | 0.03815 | 21.9 | 0.06552 | 0.8012 | 0.8811 | 1.331 |
| 18:3 Linolenic | 0.03924 | 18.9 | 0.06787 | 0.7299 | 0.7838 | 1.172 |
| 20:0 Arachidic | 0.03437 | 19.9 | 0.05928 | 0.7222 | 0.8741 | 1.235 |
| 20:1 Eicosenoic | 0.04525 | 19.1 | 0.07822 | 0.6997 | 0.7398 | 1.156 |
| 22:0 Behenic | 0.0333 | 17.6 | 0.05781 | 0.6344 | 0.8238 | 1.119 |
| Acid Detergent | 0.05871 | 49.4 | 0.09842 | 1.199 | 0.7055 | 1.365 |
| Alanine | 0.03396 | 15.8 | 0.05934 | 0.5444 | 0.9455 | 1.247 |
| Arginine | 0.02181 | 30.9 | 0.03698 | 0.9793 | 1.0035 | 1.292 |
| Ash | 0.03408 | 53.6 | 0.05704 | 1.3196 | NA | NA |
| Aspartic Acid | 0.02737 | 16.2 | 0.04775 | 0.5673 | 0.9682 | 1.22 |
| Calcium | 0.04933 | 15.6 | 0.08626 | 0.5445 | 0.7269 | 1.088 |
| Carbohydrates | 0.003 | 25.8 | 0.00512 | 0.9005 | 0.968 | 1.002 |
| Copper | 0.06325 | 18.1 | 0.10965 | 0.6449 | 0.8289 | 1.495 |
| Cystine | 0.02497 | 19.4 | 0.04314 | 0.7272 | 0.9241 | 1.192 |
| Ferulic Acid | 0.0437 | 25.3 | 0.07461 | 0.9037 | 0.7088 | 1.174 |
| Folic Acid | 0.09769 | 50.1 | 0.16371 | 1.2414 | 0.6004 | 1.789 |
| Glutamic Acid | 0.03721 | 15.6 | 0.06506 | 0.5314 | 0.9344 | 1.258 |
| Glycine | 0.01823 | 23.8 | 0.0312 | 0.8476 | 0.9914 | 1.214 |
| Histidine | 0.02682 | 18.9 | 0.04639 | 0.6905 | 0.9446 | 1.228 |
| Iron | 0.04621 | 25.8 | 0.07884 | 0.8843 | 0.837 | 1.406 |
| Isoleucine | 0.03621 | 17.2 | 0.06294 | 0.6202 | 0.9282 | 1.288 |
| Leucine | 0.04149 | 15.4 | 0.07261 | 0.5194 | 0.9198 | 1.274 |
| Lysine | 0.0173 | 48.5 | 0.02901 | 1.1774 | 1.0314 | 1.258 |
| Magnesium | 0.02689 | 18.2 | 0.04661 | 0.672 | 0.9372 | 1.213 |
| Manganese | 0.05516 | 14.5 | 0.09692 | 0.4691 | 0.86 | 1.279 |
| Methionine | 0.03613 | 17.7 | 0.06271 | 0.6434 | 0.8736 | 1.223 |
| Moisture | 0.01227 | 27.4 | 0.02089 | 0.9068 | 0.9365 | 1.076 |
| Neutral Detergent | 0.04428 | 47 | 0.0743 | 1.2042 | 0.8149 | 1.346 |
| Niacin | 0.03534 | 16.9 | 0.0615 | 0.5908 | 0.879 | 1.195 |
| p-Coumaric Acid | 0.0749 | 16.7 | 0.13043 | 0.5886 | 0.739 | 1.416 |
| Phenylalanine | 0.03652 | 15.2 | 0.06397 | 0.5052 | 0.9413 | 1.245 |
| Phosphorus | 0.0234 | 21.7 | 0.0402 | 0.8019 | 0.976 | 1.256 |
| Phytic Acid | 0.05812 | 52.2 | 0.09733 | 1.3671 | NA | NA |
| Potassium | 0.02013 | 20.8 | 0.03465 | 0.8057 | 0.9711 | 1.21 |
| Proline | 0.03664 | 16.5 | 0.06385 | 0.5794 | 0.9222 | 1.263 |
| Protein | 0.02864 | 15.4 | 0.05013 | 0.5256 | 0.9556 | 1.199 |
| Raffinose | 0.07993 | 27.2 | 0.1361 | 0.8909 | 0.5197 | 1.272 |
| Serine | 0.03083 | 17.9 | 0.05348 | 0.6566 | 0.9427 | 1.261 |
| Threonine | 0.02716 | 20.8 | 0.04675 | 0.7536 | 0.9568 | 1.267 |
| Total Dietary | 0.04607 | 45.2 | 0.07736 | 1.1753 | 0.8161 | 1.395 |
| Total Fat | 0.03215 | 17.4 | 0.05585 | 0.6128 | 0.8947 | 1.192 |
| Tryptophan | 0.0247 | 41 | 0.04157 | 1.1525 | 0.9894 | 1.318 |
| Tyrosine | 0.04963 | 35 | 0.08385 | 1.0448 | 0.8899 | 1.579 |
| Valine | 0.02894 | 18.2 | 0.05015 | 0.6574 | 0.9613 | 1.263 |
| Vitamin B1 | 0.03163 | 21.4 | 0.05438 | 0.7521 | 0.8697 | 1.203 |
| Vitamin B2 | 0.0323 | 45.6 | 0.05423 | 1.1698 | 0.8717 | 1.278 |
| Vitamin B6 | 0.03223 | 19.4 | 0.05567 | 0.6921 | 0.9782 | 1.34 |
| Vitamin E | 0.0554 | 20.7 | 0.09539 | 0.7433 | 0.9446 | 1.665 |
| Zinc | 0.03763 | 17.1 | 0.06544 | 0.641 | 0.8811 | 1.251 |

**Table 6:** Analytes with a significant ($p<0.05$) GxE interaction (p value given), and geometric means per site (rows) and genotypes (first column: comparator, second column: GMO). See description of mixed statistical model in section 5.1.

| | | |
|---|---|---|
| 20:0 Arachidic | p=0.049 | |
| 1 | 0.01224 | 0.01172 |
| 2 | 0.01193 | 0.01389 |
| 3 | 0.01334 | 0.01335 |
| 4 | 0.01242 | 0.01283 |
| Ash | p=0.026 | |
| 1 | 1.112 | 1.079 |
| 2 | 1.215 | 1.199 |
| 3 | 1.279 | 1.162 |
| 4 | 0.977 | 1.232 |
| Carbohydrates | p=0.043 | |
| 1 | 74.63 | 73.70 |
| 2 | 75.96 | 74.73 |
| 3 | 75.35 | 75.63 |
| 4 | 76.67 | 76.30 |
| Ferulic Acid | p=0.036 | |
| 1 | 1849 | 1686 |
| 2 | 2059 | 1430 |
| 3 | 2038 | 1931 |
| 4 | 2242 | 1999 |
| Folic Acid | p=0.011 | |
| 1 | 0.6622 | 0.4817 |
| 2 | 0.5206 | 0.7715 |
| 3 | 0.3684 | 0.7524 |
| 4 | 0.5931 | 0.5226 |
| Isoleucine | p=0.037 | |
| 1 | 2.820 | 2.829 |
| 2 | 2.641 | 3.059 |
| 3 | 2.559 | 2.762 |
| 4 | 2.886 | 2.836 |
| Neutral Detergent Fiber | p=0.003 | |
| 1 | 8.547 | 12.288 |
| 2 | 8.953 | 7.760 |
| 3 | 8.834 | 10.753 |
| 4 | 8.485 | 9.090 |
| Total Dietary Fiber | p=0.021 | |
| 1 | 10.21 | 14.45 |
| 2 | 10.69 | 10.38 |
| 3 | 12.55 | 14.23 |
| 4 | 12.67 | 12.59 |

**Table 7:** Geometric means and geometric standard errors of means (sem) per site (in columns labelled 1-4) for comparator (comp) and GMO. GM=exp(mean) and GSEM=exp(sem) where mean and sem refer to quantities calculated at the logarithmic scale. Approximate 95% confidence interval for each GM is [GM x GSEM-2, GM x GSEM2]. See description of mixed statistical model in section 5.1.

| Analyte | Line | geometric mean (GM) | | | | geometric sem (GSEM) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 16:0 Palmitic | comp | 0.3523 | 0.407 | 0.4185 | 0.41 | 1.0737 | 1.0783 | 1.0783 | 1.0737 |
| 16:0 Palmitic | GMO | 0.3733 | 0.4488 | 0.4074 | 0.4099 | 1.0737 | 1.0737 | 1.0737 | 1.0737 |
| 16:1 Palmitoleic | comp | 0.0038 | 0.0036 | 0.0037 | 0.0036 | 1.0777 | 1.0828 | 1.0829 | 1.0777 |
| 16:1 Palmitoleic | GMO | 0.0035 | 0.0038 | 0.0037 | 0.0038 | 1.0777 | 1.0777 | 1.0777 | 1.0777 |
| 18:0 Stearic | comp | 0.0446 | 0.045 | 0.0499 | 0.0477 | 1.0972 | 1.0972 | 1.0972 | 1.0911 |
| 18:0 Stearic | GMO | 0.0434 | 0.0512 | 0.0469 | 0.0475 | 1.0911 | 1.0911 | 1.0911 | 1.0911 |
| 18:1 Oleic | comp | 0.866 | 0.8498 | 0.894 | 0.8689 | 1.0709 | 1.0755 | 1.0756 | 1.0709 |
| 18:1 Oleic | GMO | 0.8823 | 1.0155 | 0.9131 | 0.9348 | 1.0709 | 1.0709 | 1.0709 | 1.0709 |
| 18:2 Linoleic | comp | 1.469 | 1.5799 | 1.5032 | 1.483 | 1.0929 | 1.0929 | 1.0929 | 1.0871 |
| 18:2 Linoleic | GMO | 1.4386 | 1.7092 | 1.4059 | 1.4349 | 1.0871 | 1.0871 | 1.0871 | 1.0871 |
| 18:3 Linolenic | comp | 0.033 | 0.0349 | 0.0353 | 0.0336 | 1.0809 | 1.0863 | 1.0864 | 1.0809 |
| 18:3 Linolenic | GMO | 0.0279 | 0.0358 | 0.0313 | 0.0297 | 1.0809 | 1.0809 | 1.0809 | 1.0809 |
| 20:0 Arachidic | comp | 0.0122 | 0.0119 | 0.0133 | 0.0124 | 1.0688 | 1.0733 | 1.0734 | 1.0688 |
| 20:0 Arachidic | GMO | 0.0117 | 0.0139 | 0.0133 | 0.0128 | 1.0688 | 1.0688 | 1.0688 | 1.0688 |
| 20:1 Eicosenoic | comp | 0.0103 | 0.0108 | 0.0116 | 0.011 | 1.0915 | 1.0976 | 1.0976 | 1.0915 |
| 20:1 Eicosenoic | GMO | 0.0094 | 0.012 | 0.0112 | 0.0111 | 1.0915 | 1.0915 | 1.0915 | 1.0915 |
| 22:0 Behenic | comp | 0.0052 | 0.0055 | 0.0055 | 0.0054 | 1.0604 | 1.0644 | 1.0644 | 1.0604 |
| 22:0 Behenic | GMO | 0.0046 | 0.0054 | 0.0053 | 0.0052 | 1.0604 | 1.0604 | 1.0604 | 1.0604 |
| Acid Detergent | comp | 3.3755 | 3.9577 | 3.5598 | 3.371 | 1.2127 | 1.227 | 1.227 | 1.2127 |
| Acid Detergent | GMO | 4.4561 | 3.5188 | 4.053 | 3.5809 | 1.2127 | 1.2127 | 1.2127 | 1.2127 |
| Alanine | comp | 6.1783 | 6.3636 | 5.747 | 6.3188 | 1.0557 | 1.0592 | 1.0593 | 1.0557 |
| Alanine | GMO | 6.2663 | 6.9212 | 6.0936 | 6.216 | 1.0557 | 1.0557 | 1.0557 | 1.0557 |
| Arginine | comp | 3.4874 | 3.6121 | 3.4367 | 3.9932 | 1.0554 | 1.059 | 1.0591 | 1.0554 |
| Arginine | GMO | 3.757 | 3.85 | 3.7828 | 3.8738 | 1.0554 | 1.0554 | 1.0554 | 1.0554 |
| Ash | comp | 1.1122 | 1.2149 | 1.2794 | 0.9768 | 1.114 | 1.1216 | 1.1217 | 1.114 |
| Ash | GMO | 1.0791 | 1.1986 | 1.1624 | 1.2324 | 1.114 | 1.114 | 1.114 | 1.114 |
| Aspartic Acid | comp | 5.1443 | 5.3191 | 5.0386 | 5.5297 | 1.0443 | 1.0471 | 1.0472 | 1.0443 |
| Aspartic Acid | GMO | 5.2465 | 5.7101 | 5.402 | 5.463 | 1.0443 | 1.0443 | 1.0443 | 1.0443 |
| Calcium | comp | 48.6135 | 57.1976 | 43.1981 | 55.2638 | 1.075 | 1.0797 | 1.0797 | 1.075 |
| Calcium | GMO | 48.0419 | 58.9964 | 42.2952 | 48.5132 | 1.075 | 1.075 | 1.075 | 1.075 |
| Carbohydrates | comp | 74.632 | 75.9555 | 75.3542 | 76.6661 | 1.0066 | 1.007 | 1.007 | 1.0066 |
| Carbohydrates | GMO | 73.6996 | 74.7317 | 75.6331 | 76.2989 | 1.0066 | 1.0066 | 1.0066 | 1.0066 |
| Copper | comp | 1.1924 | 1.2133 | 1.0941 | 1.1347 | 1.1161 | 1.1236 | 1.1236 | 1.1236 |
| Copper | GMO | 1.1855 | 1.2149 | 1.198 | 1.3788 | 1.1161 | 1.1161 | 1.1161 | 1.1161 |
| Cystine | comp | 1.732 | 1.6747 | 1.6404 | 1.7428 | 1.0544 | 1.0578 | 1.0579 | 1.0544 |
| Cystine | GMO | 1.703 | 1.6947 | 1.5984 | 1.7661 | 1.0544 | 1.0544 | 1.0544 | 1.0544 |
| Ferulic Acid | comp | 1848.967 | 2058.695 | 2037.814 | 2242.183 | 1.0884 | 1.1102 | 1.0943 | 1.0884 |
| Ferulic Acid | GMO | 1686.142 | 1429.616 | 1931.492 | 1999.323 | 1.0884 | 1.0884 | 1.0884 | 1.0884 |
| Folic Acid | comp | 0.6622 | 0.5206 | 0.3684 | 0.5931 | 1.3577 | 1.3837 | 1.3837 | 1.3577 |
| Folic Acid | GMO | 0.4817 | 0.7715 | 0.7524 | 0.5226 | 1.3577 | 1.3577 | 1.3577 | 1.3577 |
| Glutamic Acid | comp | 15.688 | 16.0293 | 14.4335 | 15.8283 | 1.0612 | 1.0651 | 1.0652 | 1.0612 |
| Glutamic Acid | GMO | 15.932 | 17.3535 | 15.2838 | 15.7264 | 1.0612 | 1.0612 | 1.0612 | 1.0612 |
| Glycine | comp | 2.9947 | 3.0362 | 2.8948 | 3.283 | 1.0418 | 1.0445 | 1.0445 | 1.0418 |
| Glycine | GMO | 3.0761 | 3.2185 | 3.1497 | 3.2455 | 1.0418 | 1.0418 | 1.0418 | 1.0418 |
| Histidine | comp | 2.3813 | 2.3527 | 2.2401 | 2.5318 | 1.053 | 1.0563 | 1.0564 | 1.053 |
| Histidine | GMO | 2.4295 | 2.5324 | 2.3923 | 2.4563 | 1.053 | 1.053 | 1.053 | 1.053 |
| Iron | comp | 20.5231 | 16.5481 | 14.4395 | 16.775 | 1.1189 | 1.1265 | 1.1265 | 1.1189 |
| Iron | GMO | 17.3006 | 18.0875 | 14.5993 | 16.3788 | 1.1189 | 1.1189 | 1.1189 | 1.1189 |
| Isoleucine | comp | 2.8195 | 2.6406 | 2.5589 | 2.8856 | 1.0618 | 1.0656 | 1.0657 | 1.0618 |
| Isoleucine | GMO | 2.829 | 3.0591 | 2.7625 | 2.8356 | 1.0618 | 1.0618 | 1.0618 | 1.0618 |
| Leucine | comp | 10.5587 | 10.6517 | 9.3433 | 10.2534 | 1.0666 | 1.0707 | 1.0708 | 1.0666 |
| Leucine | GMO | 10.6997 | 11.7228 | 9.841 | 10.0837 | 1.0666 | 1.0666 | 1.0666 | 1.0666 |
| Lysine | comp | 2.541 | 2.5574 | 2.4428 | 2.8164 | 1.053 | 1.0564 | 1.0564 | 1.053 |

| Analyte | Line | geometric mean (GM) | | | | geometric sem (GSEM) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| Lysine | GMO | 2.6524 | 2.7018 | 2.7221 | 2.787 | 1.053 | 1.053 | 1.053 | 1.053 |
| Magnesium | comp | 1029.583 | 1080.284 | 1050.303 | 1079.969 | 1.053 | 1.0563 | 1.0563 | 1.053 |
| Magnesium | GMO | 1059.888 | 1159.391 | 1086.656 | 1112.034 | 1.053 | 1.053 | 1.053 | 1.053 |
| Manganese | comp | 6.5591 | 6.4184 | 5.7665 | 6.734 | 1.0753 | 1.0803 | 1.0803 | 1.0753 |
| Manganese | GMO | 6.9418 | 7.1551 | 5.9496 | 6.838 | 1.0753 | 1.0753 | 1.0753 | 1.0753 |
| Methionine | comp | 1.8314 | 1.5648 | 1.7427 | 1.9091 | 1.071 | 1.0756 | 1.0757 | 1.071 |
| Methionine | GMO | 1.712 | 1.6016 | 1.7096 | 1.8589 | 1.071 | 1.071 | 1.071 | 1.071 |
| Moisture | comp | 13.3322 | 11.3996 | 12.2491 | 10.9997 | 1.0299 | 1.0317 | 1.0317 | 1.0299 |
| Moisture | GMO | 13.6651 | 11.7975 | 12.0327 | 11.0256 | 1.0299 | 1.0299 | 1.0299 | 1.0299 |
| Neutral Detergent | comp | 8.5471 | 8.9531 | 8.8343 | 8.4852 | 1.1293 | 1.1381 | 1.1382 | 1.1293 |
| Neutral Detergent | GMO | 12.2884 | 7.76 | 10.7532 | 9.0896 | 1.1293 | 1.1293 | 1.1293 | 1.1293 |
| Niacin | comp | 17.2959 | 18.3014 | 18.3277 | 19.2263 | 1.0629 | 1.067 | 1.067 | 1.0629 |
| Niacin | GMO | 15.1891 | 15.2667 | 16.0604 | 17.1608 | 1.0629 | 1.0629 | 1.0629 | 1.0629 |
| p-Coumaric Acid | comp | 155.6574 | 131.9272 | 136.3596 | 191.097 | 1.1177 | 1.1254 | 1.1254 | 1.1177 |
| p-Coumaric Acid | GMO | 145.3234 | 126.3005 | 154.8215 | 168.4816 | 1.1177 | 1.1177 | 1.1177 | 1.1177 |
| Phenylalanine | comp | 4.1835 | 4.2106 | 3.8323 | 4.1327 | 1.0559 | 1.0594 | 1.0595 | 1.0559 |
| Phenylalanine | GMO | 4.2861 | 4.6191 | 4.0376 | 4.1005 | 1.0559 | 1.0559 | 1.0559 | 1.0559 |
| Phosphorus | comp | 2645.937 | 2751.999 | 2870 | 2926.572 | 1.0537 | 1.0571 | 1.0571 | 1.0537 |
| Phosphorus | GMO | 2571.828 | 2808.924 | 2859.64 | 2990.537 | 1.0537 | 1.0537 | 1.0537 | 1.0537 |
| Phytic Acid | comp | 0.5367 | 0.6714 | 0.6613 | 0.5324 | 1.2416 | 1.2586 | 1.3075 | 1.2416 |
| Phytic Acid | GMO | 0.5626 | 0.5697 | 0.4593 | 0.51 | 1.2416 | 1.2416 | 1.2416 | 1.2416 |
| Potassium | comp | 2986.567 | 3161.202 | 3384.908 | 3451.695 | 1.0463 | 1.0492 | 1.0492 | 1.0463 |
| Potassium | GMO | 2815.598 | 3030.508 | 3372.244 | 3512.048 | 1.0463 | 1.0463 | 1.0463 | 1.0463 |
| Proline | comp | 7.3072 | 7.3341 | 6.871 | 7.5242 | 1.0654 | 1.0693 | 1.0694 | 1.0654 |
| Proline | GMO | 7.6084 | 8.1899 | 7.5316 | 7.4517 | 1.0654 | 1.0654 | 1.0654 | 1.0654 |
| Protein | comp | 8.1201 | 8.4489 | 7.9523 | 8.386 | 1.0453 | 1.0481 | 1.0482 | 1.0453 |
| Protein | GMO | 8.5555 | 8.8292 | 8.1298 | 8.2988 | 1.0453 | 1.0453 | 1.0453 | 1.0453 |
| Raffinose | comp | 0.104 | 0.126 | 0.0862 | 0.1388 | 1.2135 | 1.2278 | 1.2278 | 1.2135 |
| Raffinose | GMO | 0.0995 | 0.1351 | 0.1107 | 0.1282 | 1.2135 | 1.2135 | 1.2135 | 1.2135 |
| Serine | comp | 4.1717 | 4.3768 | 3.6869 | 4.189 | 1.0619 | 1.0658 | 1.0659 | 1.0619 |
| Serine | GMO | 4.1366 | 4.6562 | 4.0165 | 4.1865 | 1.0619 | 1.0619 | 1.0619 | 1.0619 |
| Threonine | comp | 2.5739 | 2.6989 | 2.7115 | 2.9123 | 1.0581 | 1.0619 | 1.062 | 1.0581 |
| Threonine | GMO | 2.8464 | 2.9543 | 2.7599 | 2.8756 | 1.0581 | 1.0581 | 1.0581 | 1.0581 |
| Total Dietary Fiber | comp | 10.215 | 10.686 | 12.5464 | 12.6739 | 1.1343 | 1.1434 | 1.1435 | 1.1343 |
| Total Dietary Fiber | GMO | 14.4456 | 10.3752 | 14.2321 | 12.5904 | 1.1343 | 1.1343 | 1.1343 | 1.1343 |
| Total Fat | comp | 2.7593 | 3.0895 | 3.1242 | 2.9917 | 1.0533 | 1.0567 | 1.0568 | 1.0533 |
| Total Fat | GMO | 2.9795 | 3.4425 | 3.0225 | 3.1293 | 1.0533 | 1.0533 | 1.0533 | 1.0533 |
| Tryptophan | comp | 0.493 | 0.4741 | 0.4598 | 0.4902 | 1.086 | 1.0916 | 1.0916 | 1.086 |
| Tryptophan | GMO | 0.5104 | 0.5055 | 0.4823 | 0.5222 | 1.086 | 1.086 | 1.086 | 1.086 |
| Tyrosine | comp | 2.3683 | 2.9282 | 2.8336 | 2.7466 | 1.1438 | 1.1534 | 1.1535 | 1.1438 |
| Tyrosine | GMO | 2.9766 | 2.9338 | 2.489 | 2.7004 | 1.1438 | 1.1438 | 1.1438 | 1.1438 |
| Valine | comp | 3.7585 | 3.6025 | 3.5535 | 3.9497 | 1.0525 | 1.0558 | 1.0559 | 1.0525 |
| Valine | GMO | 3.8319 | 4.0519 | 3.8355 | 3.8908 | 1.0525 | 1.0525 | 1.0525 | 1.0525 |
| Vitamin B1 | comp | 0.3566 | 0.3517 | 0.3217 | 0.3433 | 1.0684 | 1.0729 | 1.073 | 1.0684 |
| Vitamin B1 | GMO | 0.3533 | 0.3531 | 0.3631 | 0.3397 | 1.0684 | 1.0684 | 1.0684 | 1.0684 |
| Vitamin B2 | comp | 1.1726 | 1.0592 | 0.9628 | 1.5177 | 1.1066 | 1.1135 | 1.1135 | 1.1066 |
| Vitamin B2 | GMO | 1.0212 | 1.0681 | 1.0347 | 1.3123 | 1.1066 | 1.1066 | 1.1066 | 1.1066 |
| Vitamin B6 | comp | 4.2384 | 4.5327 | 4.802 | 5.4492 | 1.0683 | 1.0727 | 1.0727 | 1.0683 |
| Vitamin B6 | GMO | 4.2988 | 4.454 | 4.8347 | 5.3766 | 1.0683 | 1.0683 | 1.0683 | 1.0683 |
| Vitamin E | comp | 0.0066 | 0.006 | 0.0063 | 0.0063 | 1.1203 | 1.1281 | 1.1281 | 1.1203 |
| Vitamin E | GMO | 0.0064 | 0.0057 | 0.0056 | 0.0064 | 1.1203 | 1.1203 | 1.1203 | 1.1203 |
| Zinc | comp | 19.9744 | 20.7256 | 17.9435 | 19.3028 | 1.0716 | 1.0762 | 1.0762 | 1.0716 |

**Table 8:** Analytes classified by outcome category and type (see Section 4.1).

| Category I | Category II | Category III | Category IV | Not categorized |
|---|---|---|---|---|
| **Type 1** | **Type 3** | **Type 5** | **Type 7** | **Vg=0** |
| 16:1 Palmitoleic | Phosphorus | 16:0 Palmitic | - | Ash |
| 18:0 Stearic | Potassium | | | Phytic Acid |
| 18:2 Linoleic | Vitamin B6 | **Type 6** | | |
| 20:0 Arachidic | Vitamin E | Niacin | | |
| 20:1 Eicosenoic | | | | |
| Acid Detergent Fiber | **Type 4** | | | |
| Alanine | Lysine | | | |
| Calcium | | | | |
| Copper | | | | |
| Cystine | | | | |
| Folic Acid | | | | |
| Glutamic Acid | | | | |
| Histidine | | | | |
| Iron | | | | |
| Leucine | | | | |
| Methionine | | | | |
| Moisture | | | | |
| p-Coumaric Acid | | | | |
| Raffinose | | | | |
| Serine | | | | |
| Tyrosine | | | | |
| Vitamin B1 | | | | |
| Zinc | | | | |
| **Type 2** | | | | |
| 18:1 Oleic | | | | |
| 18:3 Linolenic | | | | |
| 22:0 Behenic | | | | |
| Arginine | | | | |
| Aspartic Acid | | | | |
| Carbohydrates | | | | |
| Ferulic Acid | | | | |
| Glycine | | | | |
| Isoleucine | | | | |
| Magnesium | | | | |
| Manganese | | | | |
| Neutral Detergent Fiber | | | | |
| Phenylalanine | | | | |
| Proline | | | | |
| Protein | | | | |
| Threonine | | | | |
| Total Dietary Fiber | | | | |
| Total Fat | | | | |
| Tryptophan | | | | |
| Valine | | | | |
| Vitamin B2 | | | | |

## 6. Conclusions and Recommendations

The GMO Panel concludes that whereas general guidance may certainly be given now, it is not possible to provide rules for experimental design and analysis that are optimal in every situation. The scientific state of the art is not unanimous on approaches for risk assessment and equivalence testing, and particular issues are highlighted in this opinion that may be clarified by further research. Nevertheless, general rules can be proposed now that may need to be further modified by experience gained and development of scientific knowledge, as for all guidance.

In this section we give as clearly as possible the recommendations resulting from the investigations done by Working Group members and discussions in the GMO Panel.

The recommendations are translated into definite text proposed for incorporation in the draft EC Guidelines which are currently under development.

### 6.1. Recommendations

1. Compare the GMO with its conventional counterpart, and with the mean of the commercial reference varieties with a history of safe use, by calculating appropriate differences on an appropriate scale for all relevant endpoints. Unless inappropriate, logratios (differences on log scale) should be employed for quantitative measurements.

2. Natural variation can arise from both environmental sources (i.e. between different sites and different years) and genotypic sources (variation between representative commercial varieties). In a proper experimental design the natural variation, free of environmental effects, may be quantified from experimental data which include multiple sites and multiple commercial varieties. It is recommended to fit linear mixed models for the logarithmically-transformed data including random effects for commercial genotypes.

3. Calculate 90 % confidence limits for these logratios based on a quantification of the experimental variation in the combined data of all sites. Other statistical approaches may be possible.

4. Estimate equivalence limits from the data from the same field trials as those used to test the GM and its conventional counterpart, reflecting the range of natural variation expected for those commercial reference varieties.

5. Use the confidence limits calculated in 3 and equivalence limits calculated in 4 to perform a difference test and an equivalence test.

6. Prepare a graph showing the estimated means and confidence intervals for the logratios for all relevant endpoints. Label the axis by the amount of change on the natural scale, using percent change (e.g. -20 % and +25 %) for relatively small changes, or factors (e.g. ½ and 2) for larger changes.

7. Adjust the estimated equivalence limits so that a single set of confidence limits for the difference test may validly represent both the difference and equivalence test.

8. Indicate on the graph the zero-difference point (0) and the upper and lower adjusted equivalence limits.

9. Very rarely, the above approach cannot be used to provide good estimates of natural variation to base equivalence limits on, and it may then be quantified from other sources (e.g. appropriate databases), if and only if the applicant can supply strong justification why it is reasonable to assume the representativeness of this information. The intended point of reference for judging

equivalence may be either the set of commercial varieties or the conventional counterpart. In particular:

10. When the natural variation is very small or zero, and the calculated equivalence limits are considered by experts to have little practical relevance, external data may be used to establish new equivalence limits.

11. For the interpretation, each confidence interval should be compared to 0 (proof of difference) and to the equivalence limits (proof of equivalence) consulting the scheme given in Section 4.1 of this opinion. The seven possible types of outcome should be interpreted as follows:

    I. Types 1 and 2: the GMO is equivalent to its reference;

    II. Types 3 and 4: equivalence more likely than not, further evaluation may be required;

    III. Types 5 and 6: non-equivalence more likely than not, further evaluation required;

    IV. Type 7: non-equivalence, further evaluation required.

12. Frequencies of significant results of the proof of difference tests over the complete set of considered endpoints should be reported and discussed.

13. The necessity of further assessment should be based on considering the patterns of observed logratios, and further assessments should focus on biological/toxicological relevance, taking safety limits into account when available.

14. When, in the combined data analysis across sites, biologically or toxicologically relevant unexplained differences between the GM and its conventional counterpart are demonstrated, then further analysis is required to assess to what extent such differences vary across sites.

15. Experimental designs of field trials must ensure that sufficient replication, different environmental conditions and commercial varieties are included to allow adequate quantification of natural variation. Specific minimum requirements are outlined in Section 2.

### 6.2. Issues for further consideration

The GMO Panel recognises that its recommendations leave open several issues. Partly these may be amenable to further guidance following further investigations. For these open issues applicants should find the best possible solutions, and they are encouraged to seek statistical advice to propose approaches for specific cases. Among the open issues are the following:

1. Models for data which cannot readily be transformed to normality: e.g. continuous non-normal data, or counts, or quantal, or ordinal data.

2. Power analysis for mixed model situations. Research is needed to characterize the coverage probability of the estimated confidence intervals for small sample sizes, such as three plots, two years, and four sites, because the available models are asymptotic. Moreover, research is needed for an optimal design, i.e. optimal numbers of plots and sites for a most powerful decision on equivalence.

3. The adaptation of the statistical design and analysis to more complicated designs (e.g. repeated measures).

4. Multiplicity of endpoints. Current recommendations are for single endpoints. When performing many simultaneous tests spurious significant results may be expected both in proof of difference and proof of equivalence. Further work is needed on how to handle this.

5. Multivariate analysis may give an alternative approach to the multiplicity issue, but more research is needed.

## REFERENCES

Altman D. and Bland J.M. 1995. Absence of evidence is not evidence of absence. *British Medical Journal*, 311:485.

Altman D. and Bland J.M. 2004. Confidence intervals illuminate absence of evidence. *British Medical Journal*, 328:1016-1017.

Anon. 2007. EPPO Guideline for the efficacy Evaluation of plant protection products: Design and Analysis of Efficacy Evaluation Trials, PP 1/152(3). EPPO/OEPP, Paris.

Basford K.E. and Cooper M. 1998. Genotype x environment interactions and some considerations of their implications for wheat breeding in Australia. *Australian Journal of Agricultural Research*, 49: 153-174.

Benjamini Y and Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57: 289-300.

Berger R.L. 1982. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24: 295-300.

Berger R.L. and Hsu J.C. 1996. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Sciences*, 11: 283-319.

Bofinger E. and Bofinger M. 1995. Equivalence with respect to a control: stepwise tests. *Journal of the Royal Statistical Society B*, 57: 721-733.

Brown L.D., Casella G. and Hwang J.T.G. 1995. Optimal confidence sets, bioequivalence, and the limaçon of Pascal. *Journal of the American Statistical Association*, 90: 880-889.

Chen Y.-H. and Zhou X.-H. 2006. Interval estimates for the ratio and the difference of two lognormal means. *Statistics in Medicine*, 25: 4099-4113.

Clark S.J., Rothery P. and Perry J.N. 2005. Farm Scale Evaluations of spring-sown genetically modified herbicide-tolerant crops: a statistical assessment. Proc. R. Soc. series B, 273, 237 – 243. doi:10.1098/rspb.2005.3282.

Clark S.J., Rothery P., Perry J.N. and Heard M.S. 2007. Analysis of within-field variation and assessment of sampling schemes for arable weeds using data from the Farm Scale Evaluations of genetically modified herbicide-tolerant crops. *Weed Research*, 47: 157–163.

Codex Alimentarius. 2003. Codex principles and guidelines on foods derived from biotechnology. Codex Alimentarius Commission, Joint FAO/WHO Food Standards Programme, FAO, Rome.

Dilba B., Bretz F., Guiard V., Hothorn L.A. 2004. Simultaneous confidence intervals for ratios with application to the comparison of several treatments with a control. *Meth. Inf Medicine* 43: 465-469.

Dudoit S., Shaffer J.P. and Boldrick J.C. 2003. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18: 71-103.

EC. 2003. Guidance document for the risk assessment of genetically modified plants and derived food and feed. Joint Working Group on Novel Foods and GMOs (SCP, SCF and SCAN), European Commission, Health & Consumer Protection Directorate-General.

Enot D.P. and Draper J. 2007. Statistical measures for validating plant genotype similarity assessments following multivariate analysis of metabolome data. *Metabolomics*, 3: 349-355.

EFSA. 2006. Guidance document of the Scientific Panel on genetically modified organisms for the risk assessment of genetically modified plants and derived food and feed. European Food Safety Authority, Parma, Italy.

EFSA. 2007. EFSA review of statistical analyses conducted for the assessment of the MON 863 90-day rat feeding study. European Food Safety Authority, Parma, Italy.

EFSA GMO Panel Working Group on Animal Feeding Trials. 2008. Safety and nutritional assessment of GM plants and derived food and feed: The role of animal feeding trials. *Food and Chemical Toxicology*, 46: S2–S70.

EMEA. 2001. Note for guidance on the investigation of bioavailability and bioequivalence. Document CPMP/EWP/QWP/1401/98, European Agency for the Evaluation of Medicinal Products, Committee for Proprietary Medicinal Products, London, UK.

FDA. 1997. Div. of Bio-equivalence, Office of Generic Drugs. Statistical procedures for bio-equivalence studies using standard two-treatment crossover design. In: Guidance for Industry, Rockville, MD, USA.

FDA. 2001. Guidance for Industry - Statistical approaches to establishing bioequivalence. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research.

Fieller E.C. 1954. Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B*, 16: 175-185.

Gardner M.J. and Altman D.G. 1986. Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal*, 292: 746-750.

Hammond B., Lemen J., Dudek R., Ward D., Jiang C., Nemeth M. and Burns J. 2006. Results of a 90-day safety assurance study with rats fed grain from corn rootworm-protected corn. *Food and Chemical Toxicology*, 44: 147-160.

Herman R.A., Storer N.P., Phillips A.M., Prochaska L.M. and Windels P. 2007. Compositional assessment of event DAS-59122-7 maize using substantial equivalence. *Regulatory Toxicology and Pharmacology*, 47: 37-47.

Hill R.A. and Sendashonga C. 2003. General principles for risk assessment of living modified organisms: Lessons from chemical risk assessment. *Environ. Biosafety Res.* 2: 81-88.

Hoenig J.M. and Heisley D.M. 2001. The abuse of power: The pervasive fallacy of power calculations for data analysis. *Am. Stat.,* 55: 19–24

Hothorn L.A. and Oberdoerfer R. 2006. Statistical analysis used in the nutritional assessment of novel food using the proof of safety. *Regulatory Toxicology and Pharmacology*, 44: 125-135.

Hothorn T. and Munzel U. 2002. Non-parametric confidence interval for the ratio. Report University of Erlangen, Department Medical Statistics 2002; available via: *http://www.stat.uni-muenchen.de/~hothorn/bib/TH_TR.html*.

Kenward M.G. and Roger J.H. 1997. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53: 983-997.

Kieser M. and Hauschke D. 2005. Assessment of clinical relevance by considering point estimates and associated confidence intervals. *Pharmaceutical Statistics*, 4: 101-107.

Kuiper H.A., Kleter G.A., Noteborn H.P.J.M. 2002. Substantial equivalence - an appropriate paradigm for the safety assessment of genetically modified foods? *Toxicology*, 181: 427-431.

Lindley D. 1998. Decision analysis and bioequivalence trials. *Statistical Science*, 13: 136-141.

Munk A. and Pflüger R. 1999. 1-α equivalent confidence rules for convex alternatives are α/2-level tests – with applications to the multivariate assessment of bioequivalence. *Journal of the American Statistical Association*, 94: 1311-1319.

McNaughton J.L., Roberts M., Rice D., Smith B., Hinds M., Schmidt J., Locke M, Bryant A., Rood T., Layton R., Lamb I. and Delaney B. 2007. Feeding performance in broiler chickens fed diets containing DAS-59122-7 maize grain compared to diets containing non-transgenic maize grain. *Animal Feed Science and Technology*, 132: 227-239.

Newman M.C. 2008. "What exactly are you inferring?" A loser look at hypothesis testing. *Environmental Toxicology and Chemistry*, 27: 1013-1019.

Nordic Council. 2000. Safety assessment of novel food plants. Chemical analytical approaches to the establishment of substantial equivalence. TemaNord 1998:591. Nordic Council of Ministers, Copenhagen, DK.

Oberdoerfer R.B., Shillito R.D., de Beuckeleer M and Mitten D.H. 2005. Rice (*Oryza sativa* L.) containing the *bar* gene is compositionally equivalent to the nontransgenic counterpart. *Journal of Agricultural and Food Chemistry*, 53: 1457-1465.

Pawitan Y., Michiels S., Koscielny S., Gusnanto A. and Ploner A. 2005. False discovery rate, sensitivity and sample size for microaraay studies. *Bioinformatics*, 21: 3017-3024.

Perry J.N., Rothery P., Clark S.J., Heard M.S. and Hawes C. 2003. Design, analysis and power of the Farm-Scale Evaluations of Genetically-Modified Herbicide-Tolerant crops. *Journal of Applied Ecology*: 40, 17-31.

Romano J.P. 2005. Optimal testing of equivalence hypotheses. *The Annals of Statistics*, 33: 1036-1047.

Quan H., Bolognese J. and Yuan W.Y. 2001. Assessment of equivalence on multiple endpoints. *Statistics in Medicine*, 20: 3159-3173.

Schuirmann D.J. 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15: 657-680.

Shaffer J.P. 1995. Multiple hypothesis testing. *Annual Reviews in Psychology*, 46: 561-584.

Spilke J., Piepho H.P. and Hu X. 2005. A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. *Journal of Agricultural, Biological and Environmental Statistics*, 10: 374-389.

Storey J.D. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society B*, 64: 479-498.

Storey J.D. and Tibshirani R. 2003. Statistical significance for genomewide studies. *PNAS*, 100: 9440-9445.

Tamhane A.C. and Logan B.R. 2004. Finding the maximum safe dose level for heteroscedastic data. *Journal of Biopharmaceutical Statistics*, 14: 843-856.

Tempelman R.J. 2004. Experimental design and statistical methods for classical and bioequivalence hypothesis testing with an application to dairy nutrition studies. *Journal of Animal Science*, 82: E162-E172.

Walters S.J. 2008. Consultants' forum: should post hoc sample size calculations be done? *Pharmaceutical Statistics*, published online DOI: 10.1002/pst.334.

Wang W., Hwang J.T.G. and Dasgupta A. 1999. Statistical tests for multivariate bioequivalence. *Biometrika*, 86: 395-402.

Wellek S. 2002. *Testing Statistical Hypotheses of Equivalence.* CRC Press, ISBN: 1584881607, 284 pp.

WHO. 1995. *Application of the principle of substantial equivalence to the safety evaluation of foods or food components from plants derived by modern biotechnology.* Report of a WHO workshop, WHO, Geneva.