Food and Chemical Toxicology 125 (2019) 540–548

Contents lists available at ScienceDirect

Food and Chemical Toxicology

journal homepage: www.elsevier.com/locate/foodchemtox

Equivalence limit scaled differences for untargeted safety assessments: Comparative analyses to guard against unintended effects on the environment or human health of genetically modified maize

Hilko van der Voet[a,*], Paul W. Goedhart[a], Esteban García-Ruiz[b], Concepción Escorial[b], Jana Tulinská[c]

[a] Wageningen University & Research (WUR), Biometris, Droevendaalsesteeg 1, 6708PB, Wageningen, Netherlands
[b] Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Plant Protection Department, Ctra. La Coruña Km. 7,5, 28040, Madrid, Spain
[c] Slovak Medical University (SZU), Faculty of Medicine, Limbová 12, 83303, Bratislava, Slovakia

ARTICLE INFO

Keywords:
Risk assessment
Unintended effects
Arthropods
Rat feeding study
Equivalence test
Standardised effect size

ABSTRACT

Safety assessments guard against unintended effects for human health and the environment. When new products are compared with accepted reference products by broad arrays of measurements, statistical analyses are usually summarised by significance tests or confidence intervals per endpoint. The traditional approach is to test for statistical significance of differences. However, absence or presence of significant differences is not a statement about safety. Equivalence limits are essential for safety assessment. We propose graphs to present the results of equivalence tests over the array of endpoints. It is argued that plots of the equivalence limit scaled difference (ELSD) are preferable over plots of the standardised effect size (SES) used previously for similar assessments. The ELSD method can be used either with externally specified equivalence limits or with equivalence limits estimated from (historical) data. The method is illustrated with two examples: first, environmental safety of MON810 Bt maize was assessed using field trial count data of arthropods; second, human safety of herbicide tolerant NK603 maize was assessed using haematological, biochemical and organ weight data from a 90-day rat feeding study. All assessed endpoints were classified in EFSA equivalence categories I or II, implying full equivalence or equivalence more likely than not.

1. Introduction

When a new product is investigated in a risk or safety assessment, unintended effects are commonly guarded against by comparing the new product to one or more reference products with a history of safe use. The core task of safety evaluations is to demonstrate that any unintended effect is small enough to not be a safety concern. This is demonstrated by an equivalence testing approach, which employs a null hypothesis of non-equivalence, that is, that the difference between the new product and the reference product is larger than an equivalence limit. Rejection of the null hypothesis implies that the difference is smaller than the equivalence limit and thus can be regarded as a biological irrelevant difference or, more strongly, as a "proof of safety" (OECD, 1993; FDA, 2003; EMA, 2010; EFSA, 2010b; EFSA, 2011a).

There are three major ways in which equivalence limits can be set. The first approach is that experts specify appropriate values, which is a common procedure in drug equivalence testing (FDA, 2003; EMA, 2010). In the second approach, the new product and a set of reference products are tested in the same study, and equivalence limits are derived from the variability among the reference products (van der Voet et al., 2011; Vahl and Kang, 2016). The third approach employs historical data to estimate the variability among reference products, which is used to set equivalence limits (van der Voet et al., 2017, Steinberg et al., accepted).

Usually, comparisons are made for a large number of measured variables in an effort to probe the relevant underlying biological pathways, which might have been affected unintentionally in the new product. For example, for plant materials a compositional analysis with up to 80 variables is usually performed (e.g. Oberdoerfer et al., 2005; van der Voet et al., 2011; or studies cited in Delaney, 2015). In animal studies, it is common to follow guidance documents of the Organisation for Economic Co-operation and Development (OECD) that ask for measuring at least 50 standard haematological, clinical chemical and organ weight variables (OECD, 1998, 2009; e.g. Zeljenková et al., 2014,

* Corresponding author.
E-mail address: hilko.vandervoet@wur.nl (H. van der Voet).

https://doi.org/10.1016/j.fct.2019.02.007
Received 19 November 2018; Received in revised form 1 February 2019; Accepted 4 February 2019
Available online 05 February 2019

2016). In environmental risk assessments, it is customary to compare abundance data for 10 or more relevant species or other higher taxa (e.g. Brooks et al., 2003; de la Poza et al., 2005; Marvier et al., 2007). Such studies might be considered as untargeted assessments of possible adverse effects.

Equivalence limits have also been termed "limits of concern" in environmental risk assessment (EFSA, 2010b; van der Voet et al., accepted). However, the word "concern" might suggest that exceeding such a limit would imply a biological harm. While toxicity limits such as benchmark doses may be available in toxicological studies, it is almost always very difficult to set a reliable toxicity limit for each endpoint in untargeted studies. The specified limits should be regarded as screening thresholds, and equivalence tests then serve as a screening tool. In this paper, we therefore prefer to use the term "equivalence limit" rather than "limit of concern".

The results of a comparative analysis can be represented by a set of significance statements from statistical tests. However, it is commonly advised to represent effect sizes, i.e. the observed differences between the tested and reference products, by confidence intervals (Perry et al., 2009; EFSA, 2011b). A common ratio scale can be obtained by log-transforming continuous data before a statistical analysis is performed such that, when looking at differences at the log scale, ratios are in fact of interest. Significance test results can be derived from confidence intervals by comparison of the confidence limits to zero (for difference tests) or to the equivalence limits (for equivalence tests). The use of confidence intervals also facilitates the interpretation of the observed effects as being biologically relevant or not.

Variables have different units and orders of magnitude. For example, blood sodium levels are around 145 mMol/L, and alanine aminotransferase (ALT) levels are around 0.5 μkat/L. In order to create a graphical overview of confidence intervals for effects with different values and units, it has been suggested to divide the difference by the estimated standard deviation to obtain a dimensionless quantity termed Cohen's d or Standardised Effect Size (SES) (Nakagawa and Cuthill, 2007; EFSA, 2011b; Schmidt et al., 2016; Bell et al., 2017). Cohen (1988) suggested benchmarks for the SES for "small", "medium" or "large" effects with values 0.2, 0.5 and 0.8 respectively.

In this paper we propose to scale effect sizes in another, more interpretable way. Instead of the estimated standard deviation we propose to use the equivalence limit as scaling factor. The resulting statistic is appropriately called the equivalence limit scaled difference (ELSD). ELSDs have been introduced previously for a specific situation (van der Voet et al., 2017), and were subsequently applied in the G-TwYST project (Steinberg et al., accepted). Here we emphasise their wider applicability and compare the ELSD with the SES approach.

In the remainder of this paper, section 2 describes the proposed procedure for preparing graphical overviews of equivalence limit scaled differences for untargeted safety assessments. An environmental safety assessment example using field study arthropod count data is presented in section 3, while section 4 gives an example of food/feed safety assessment using animal feeding study data. Finally, the proposed graphical device is discussed in section 5.

2. Methods

We assume that an experiment has been performed in which a new product, further called the test group, is compared with a single reference. We propose graphical overviews for untargeted safety assessments according to the following stepwise scheme:

1. An appropriate statistical analysis is performed, separately for all endpoints, and these analyses are summarised by confidence intervals for differences between the test group and the reference group at an appropriate scale.
2. Equivalence limits at the same scale are obtained for each endpoint, either derived from values specified by experts, or estimated with uncertainty from (historical) data.
3. A single graph of all confidence intervals is made, grouping endpoints for clarity, in which the interpretation of the area within the equivalence limits is stressed by using a green background colour. The confidence intervals are such that they allow for difference testing and, in the case of fixed equivalence limits, also for equivalence testing.
4. In addition, the differences and confidence intervals are scaled by means of the equivalence limits to give ELSD confidence intervals which then have common scaled equivalence limits of -1 and 1. These scaled confidence intervals are also graphically depicted with a green area between -1 and 1. Again, the ELSD confidence intervals are such that they allow for difference testing and for equivalence testing. Note than when the fixed equivalence limits are asymmetric, estimates and confidence limits on the right of the no difference value should be scaled by the upper equivalence limit, and those on the left by the lower equivalence limit.

We now describe how the confidence intervals are plotted such that they can be used for both difference and equivalence testing. We assume that testing is done at the 5% significance level with obvious alterations for other significance levels. We first consider the case where fixed equivalence limits are available. In that case we advocate the use of the two one-sided tests (TOST) approach for equivalence testing (Schuirmann, 1987). This boils down to constructing a two-sided 90% confidence interval (with 5% outside the interval on each side) which is plotted alongside the equivalence limits. If this interval is within the equivalence limits, the null hypothesis of non-equivalence is rejected in favour of equivalence at the 5% significance level. The same interval can be used when there is just a single equivalence limit. It can also be used for one-sided difference testing by checking whether the relevant one-sided part of the interval encompasses the value zero, in which case the difference test is not rejected. However, the 90% interval cannot be used for two-sided difference testing at the 5% level since this requires checking whether the two-sided 95% interval (with 2.5% outside the interval on each side) contains the value zero. We therefore propose to always plot the 90% interval and, when this gives a false impression of the two-sided difference test, to extend the plotted interval to the corresponding 95% limit.

A different procedure is applied when equivalence limits are estimated, as fully described in van der Voet et al. (2017). In short, a one-sided equivalence limit is derived on a quadratic scale from a one-sided confidence region under the (alternative) hypothesis that the difference between test and reference is zero. Consequently, in the first step of the procedure the equivalence interval on the ELSD scale is constructed as a symmetric 95% confidence interval, i.e. of the form [-EL, + EL]. The total probability outside the interval is 5%, but is typically not distributed equally over the two tails. E.g. for positive point estimates, the right-hand tail of this interval will contain more than 2.5% and the left-hand tail will contain less than 2.5%. In a second step, the intervals are adapted as follows. For positive point estimates the lower confidence limit is replaced by the lower limit of the 95% two-sided confidence interval. The lower limit of the modified interval can then be used for difference testing while the upper limit can still be used for equivalence testing. Similarly, for negative point estimates the upper limit of the equivalence interval is replaced by the upper limit of the 95% two-sided confidence interval. Although the adapted intervals cover between 92.5% and 95%, they can be used for both types of testing at the 95% confidence level.

In the second example given below the proposed graphs are compared with the more traditional SES graphs where the estimated standard deviation was used for scaling. Exact SES intervals were calculated with the MBESS package in R which implements the inversion method of Kelley (2007).

3. Example 1: effect sizes for arthropod counts in a three-year field study comparing MON810 maize and its near-isogenic line

In the project Assessing and Monitoring the Impacts of Genetically Modified Plants on Agro-ecosystems (AMIGA), a range of field studies were performed to assess the environmental safety of Bt maize, which expresses the Cry1Ab protein. Non-target arthropods, classified into five functional groups, were counted in a field trial comparing the genetically modified Bt maize MON810 variety DKC6451YG (GMO) to the near-isogenic non-Bt line DKC6450 as the comparator (CMP) (van der Voet et al., accepted). Equivalence limits for counts were externally specified as factors 0.5 and 2 for taxa with overall observed means m of 10 and higher, and the logarithm of the equivalence limits was scaled by $\sqrt{10/m}$ for lower means. These adapted equivalence limits correct for the increased variability at lower abundance levels (van der Voet et al., 2016; van der Voet et al., accepted).

Here we report field trials performed in Seseña, Spain, using the same field and plots in the years 2012, 2013 and 2014. The experimental design was a randomized block design with five lines in the field serving as blocks. Two replicates of the GMO and CMP were randomized within each line. In each of the $5 \times 4 = 20$ plots, maize was sown in 10 rows of 10 m long, separated by 0.95 m. Two pitfall traps were placed in two different maize sowing rows separated by another row in the middle, so that a barrier cutting the middle row diagonally led ground-dwelling arthropods to the traps in the outer rows. Count data for ground-dwelling arthropods were obtained in 9 sampling periods per year. After their taxonomical identification to the genus/species level in the most relevant groups and to at least the family/order level in the others, the taxa were sorted into 5 functional categories: herbivores, predators, parasitoids, detritivores, and an 'Other' group for taxa with different or unknown feeding habits. 'NI' at the end of a taxon label stands for 'Not Identified at a lower taxonomical level'. For example, the genus of predatory ground beetles *Pseudoophonus* is represented by two explicitly named species and one remaining group 'PseudoophonusNI'. Counts were summed over the two traps per plot and over the nine sampling periods per year. For the multi-year analysis presented here the counts were also summed over the three years such that summed counts are available for the 20 plots.

The mean counts for the two maize varieties and the overall mean (m) were calculated. For all 85 taxa where both variety means were positive an over-dispersed Poisson log-linear model (McCullagh and Nelder, 1989) was fitted with, on the logarithmic link scale, additive effects for block and variety. This model assumes that the variance of a count is proportional to the mean and that effects are additive on the log-scale. The dispersion factor σ^2 was estimated employing the Pearson statistic, and was set to 1 if the estimate was lower than 1. From this estimate of σ^2 the coefficient of variation (CV), expressed as a percentage, was estimated as $CV = 100\sqrt{\hat{\sigma}^2/m}$, which follows from the over-dispersed Poisson model. The log-linear model directly estimates the log-ratio (or log of fold change) of the mean of the GMO and CMP variety and its associated standard error. These were used to construct 90% and 95% confidence intervals, as described in Section 2, for the log-ratio employing the Student distribution with appropriate degrees of freedom.

Back-transformed estimates of fold change and adapted 90% confidence intervals were graphically depicted in Fig. 1, together with the equivalence region between the equivalence limits. A further scaling to equivalence limit scaled differences (ELSDs) led to the representation in Fig. 2.

For taxa with a zero average for the GMO or the CMP it is not possible to estimate the log-ratio using the OP model. For these taxa, we calculated a ratio where the zero count was replaced with the lowest possible mean value based on a count of 1, e.g. 0.1 in the current case with 10 replications. If these estimates fell outside the equivalence limits they would be included in the graphical display to focus attention on these possibly relevant changes. However, this situation did not occur for the current dataset.

In the example, the decrease or increase equalled or exceeded two-fold for 21 taxa. More than doubled counts were observed for the predatory Pompilidae (wasps), Chrysopidae (lacewings), Thomisidae and Zoropsidae (spiders), for the parasitoid Bethylidae and Aphelinidae (wasps), and for the unclassified Sphaeroceridae (flies). Less than halved counts were observed for the predatory Aeolothripidae (thrips) and Dolichopodidae (flies), for the parasitoid Ichmeumonidae (wasps), for the detritivore Chironomidae (flies) and for the unclassified Muscidae (flies). It can be observed that all these 21 taxa were relatively rare, with mean counts over ten plots of less than 1 observation per plot, both for the GMO and the CMP plots. Using the adapted equivalence limits for low abundances, the point estimates for all 85 taxa were within the equivalence region. In the terminology of EFSA (2010a), 79 taxa were classified in equivalence category I (fully equivalent, i.e. the confidence interval falls entirely between the equivalence limits), and 6 taxa in equivalence category II (equivalence more likely than not, i.e. the point estimate falls between the equivalence limits, but at least one of the confidence limits extends beyond an equivalence limit). The latter six taxa, the predatory Eurobellia (earwigs) and Chrysopidae, the detritivore Collembola (springtails), Corylophidae (beetles) and Chironomidae, and the unclassified Formicidae (ants), are indicated with red confidence limits outside the equivalence band in Figs. 1 and 2. Five taxa, indicated in blue in Figs. 1 and 2, showed significant differences, at the 95% level, between the GMO and the CMP (the herbivore Aphididae (aphids), the predatory Phytoseiidae (mites) and Opiliones (harvestmen), and the detritivore Corylophidae and Chironomidae). It can be noted that this set included more common taxa, with mean counts over ten plots almost always greater than 1 and up to 11 for the Aphididae in the GMO plots, 11 for the Opiliones in the GMO plots and 33 for the Phytoseiidae in the CMP plots.

4. Example 2: effect sizes in a 90-day rat feeding study on NK603 maize

In the project GM plants Two-Year Safety Testing (G-TwYST), three rat feeding studies were performed to assess the safety of the herbicide tolerant maize NK603 (Goedhart and van der Voet, 2017; Steinberg et al., accepted). Here we focus, for illustration of the proposed method, on the comparison of a test group of female rats fed for 90 days with a diet containing 33% NK603 maize that was treated with glyphosate (Roundup®) in the field (this was group NK33 + in G-TWYST study B) with a control group of female rats that received a diet containing 33% non-GM maize. Equivalence limits, along with confidence intervals, were estimated for 42 variables in 5 groups (body weight and growth, haematology, differential white blood counts, clinical biochemistry, organ weights). We refer to Steinberg et al. (accepted) for a full description of these variables. This approach employed data for non-GM feeding groups with 33% maize in previous studies at the same experimental facility (Schmidt et al., 2017) as described in previous work (van der Voet et al., 2017; Goedhart and van der Voet, 2017; Steinberg et al., accepted). For nine of these 42 variables fixed equivalence limits have been suggested in the literature (Hong et al., 2017).

The data were log-transformed and analysed by analysis of variance. The adapted confidence intervals, back-transformed to the ratio scale, are depicted in Fig. 3 along with the median equivalence limits (red bars) and their associated 95% intervals which are given by the blue bars. This plots shows that there were large differences in variability between the endpoints. In general, a high variability in the observed data was accompanied by a high variability in the historical data and therefore wide equivalence limits. There were two confidence intervals that did not include the value 1 for the ratio (growthRate and OvaryWeight), and therefore indicated statistically significant differences at the two-sided 95% confidence level.

For many endpoints the confidence interval for the ratio in Fig. 3 fell completely within the interval defined by the left and right median

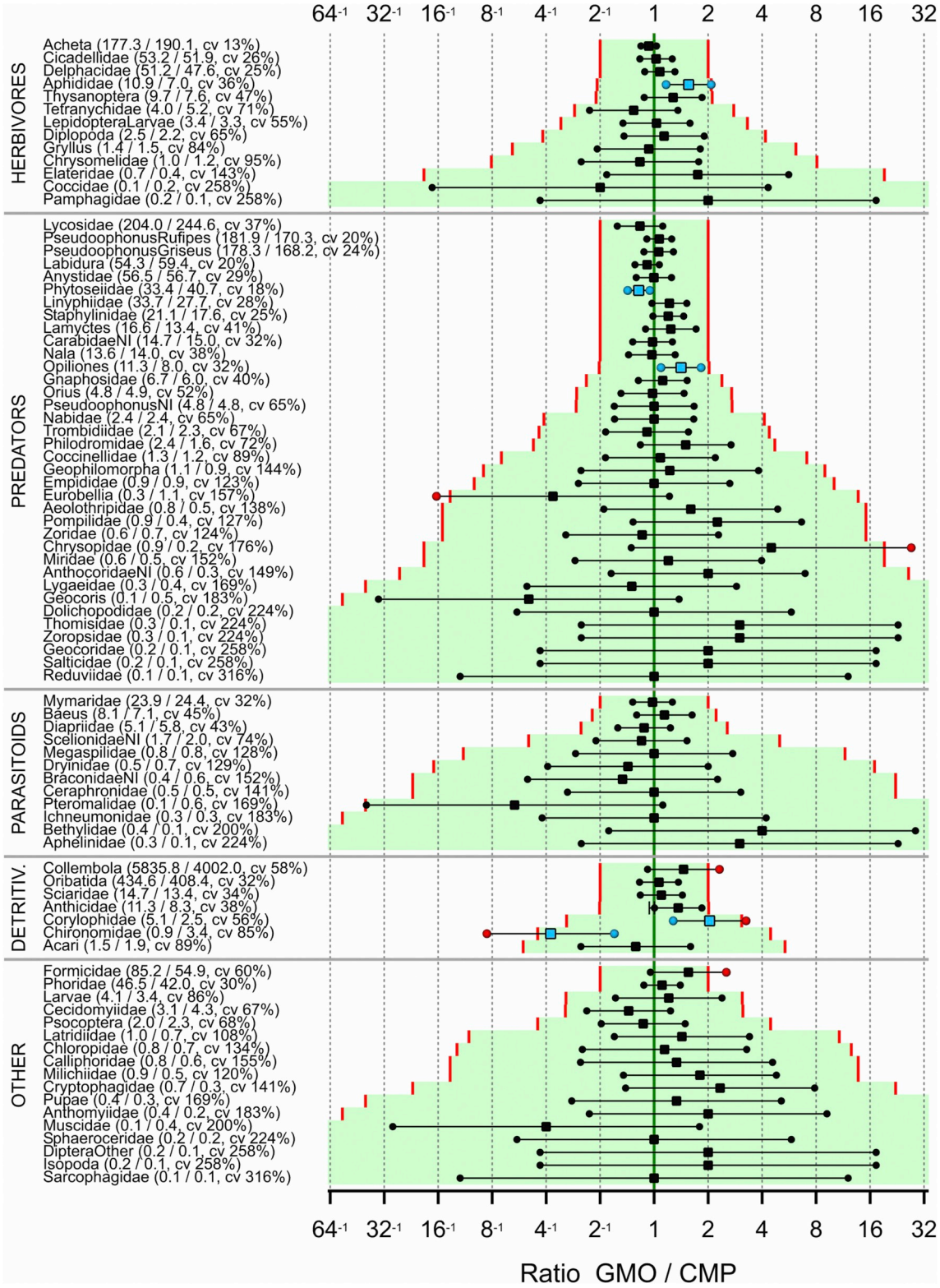

Fig. 1. Arthropods in field trials Spain, counts summed over 2012–2014. Ratios for test group (GMO) vs. comparator group (CMP) for all taxa with positive means for both varieties, with 90% confidence intervals (and an added 95% limit if needed for a correct interpretation of the difference test, in this case only for Anthicidae). Results are sorted according to decreasing abundance within functional groups. Equivalence limits (ELs, vertical red bars) are factors 0.5 and 2 for taxa with means of 10 and higher, and log(EL) is scaled by $\sqrt{10/m}$ for lower means. Means over the ten plots for GMO and CMP, and coefficient of variation (cv) are indicated in brackets in the labels. Points outside the ELs are coloured red, points inside the ELs for statistically significant differences are coloured blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

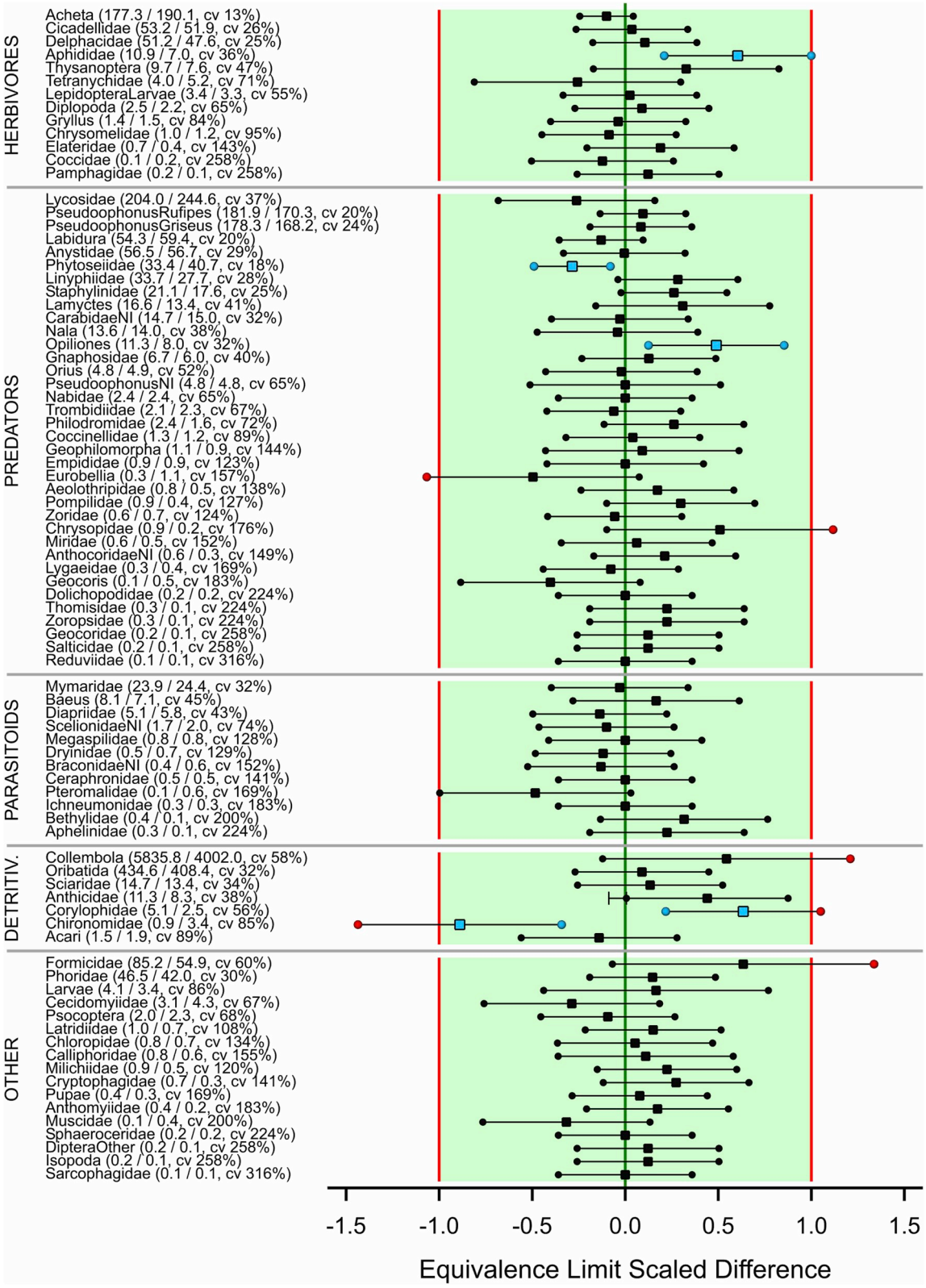

Fig. 2. Arthropods in field trials Spain, counts summed over 2012–2014. Equivalence limit scaled differences for test group (GMO) vs. comparator group (CMP) for all taxa with positive means for both varieties, shown with 90% confidence intervals (and an added 95% limit if needed for a correct interpretation of the difference test, in this case only for Anthicidae). Results are sorted according to decreasing abundance within functional groups. Equivalence limits (ELs) are factors 0.5 and 2 for taxa with means of 10 and higher, and log(EL) is scaled by $\sqrt{10/m}$ for lower means. Means over the ten plots for GMO and CMP, and the coefficient of variation (cv) are indicated in brackets in the labels. Points outside the ELs are coloured red, points inside the ELs for statistically significant differences are coloured blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

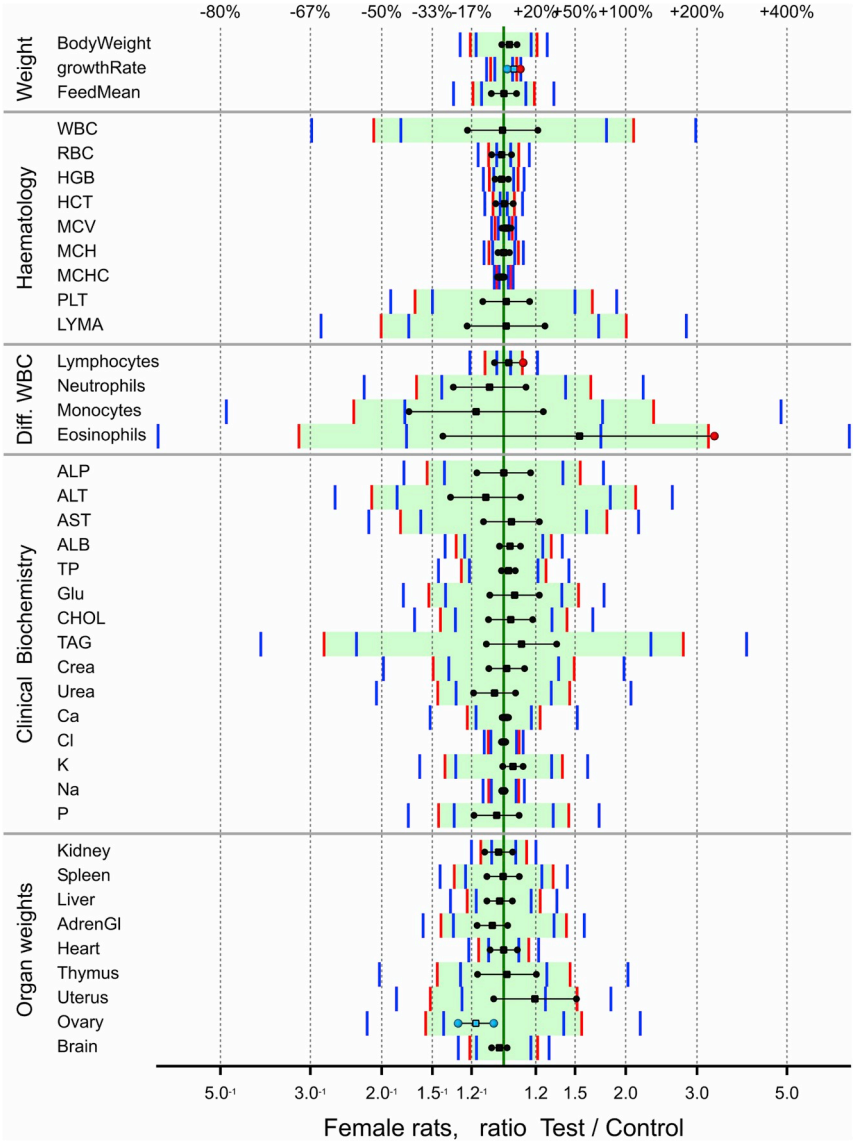

Fig. 3. Adapted confidence intervals for the ratio of NK33 + and the Control feed for female rats with added equivalence limit estimates (red bars) with 95% confidence intervals (blue bars). Points outside the ELs are coloured red, points inside the ELs for statistically significant differences are coloured blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

equivalence limits, and even within the interval defined by the 97.5% point of the left equivalence limit and the 2.5% point for the right limit. This strongly suggests that equivalence could be established for these endpoints, including BodyWeight, FeedMean, WBC, MCH, PLT, LYMA, Neurophils, Monocytes, all clinical biochemistry endpoints (ALP, ALT, AST, ALB, TP, Glu, CHOL, TAG, Crea, Urea, Ca, Cl, K, Na, P), and most organ weights (Spleen, Liver, AdrenGl, Heart, Thymus, Ovary, Brain). For other endpoints the situation was less clear. For example, RBC and MCHC had limited variability resulting in small intervals for both the ratio and the equivalence limits, and the ratio intervals for Kidney and Uterus weights had some overlap with the intervals for the equivalence limits. In any case, formal equivalence testing cannot be derived from the representation as in Fig. 3, because both the estimated ratios and the equivalence limits have uncertainties as indicated by the corresponding confidence intervals. However, the two intervals can be combined into a single interval by moving to equivalence limit scaled differences, as explained in van der Voet et al. (2017). The resulting Fig. 4a can be used both for equivalence testing and for difference testing at the 95% confidence level. In addition to the two significant differences already observed in Fig. 3 (showing an increased growthRate and a decreased Ovary weight), Fig. 4a reveals that 37 confidence intervals fell completely within the range [-1,1] indicating a proof of safety (equivalence category I). The remaining 5 confidence intervals (growthRate, HCT, Lymphocytes, Eosinophils, Uterus weight) extended beyond the range [-1, +1], but still had their central values inside, such that they are classified as equivalence category II or "equivalent more likely than not" according to the EFSA nomenclature (EFSA, 2010a). Note that the unclear equivalence status of endpoints such as RBC, MCHC, Kidney weight and Uterus weight according to Fig. 3 has been settled in Fig. 4a.

For comparison, the traditional SES plot, which employs standardisation of the estimated difference with the residual standard deviation, is given in Fig. 4b. With respect to a benchmark for SES, EFSA (2011b) states that "if experience from previous toxicity tests shows that an effect size of, say, one SD or less is of little toxicological relevance then this can be used to determine sample size in new situations". This led others (Zeljenková et al., 2014, 2016; Schmidt et al., 2016, 2017; Tulinská et al., 2018) to use benchmarks of $-1/+1$. Many intervals in Fig. 4b did not fall completely within these benchmarks.

With respect to the SES, there is a direct link between the t-value of

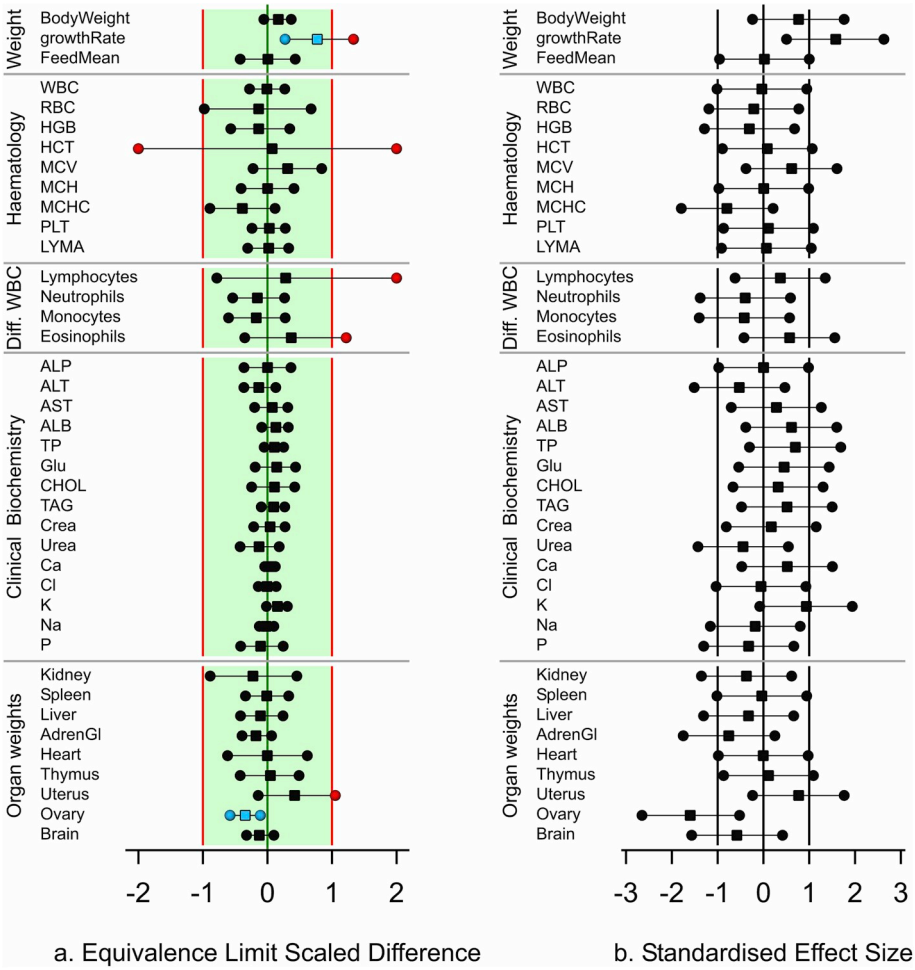

Fig. 4. Test versus Control feeds for female rats. **a**. Equivalence limit scaled differences (ELSDs), with adapted confidence intervals. Points outside the ELs $(-1, +1)$ are coloured red, points for statistically significant differences are coloured blue inside the ELs. **b**. For comparison: Standardised effect sizes, with 95% confidence intervals. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

the difference test and whether the SES fully lies within a certain interval. For example, for a two sample *t*-test with 8 replications, whenever the absolute t-value is larger than 0.04 the SES interval will not fully fall with the [-1, +1] range. This is why in Fig. 4b only intervals for which the estimate is very close to zero fall within the range [-1, +1]. So for any design and any SES benchmark there is direct link between the absolute t-value and accepting an SES value.

Finally, for nine endpoints Hong et al. (2017) provided provisional fixed equivalence limits. Seven of the limits were one-sided, specifying only an increase (ALP, Crea, Urea, CHOL, Kidney weight, Liver weight) or decrease (BodyWeight). Fig. 5 displays 90% confidence intervals for these nine endpoints. For the scaling of the plot, the given one-sided equivalence limit was also applied to the other side. In this example, all nine endpoints had intervals amply within the equivalence limits.

5. Discussion

In this paper we have proposed a graphical device for the representation of equivalence and difference tests in comparative untargeted assessments of safety. The crucial element is the specification of a relevant scale for the judgment of differences. The choice of equivalence limits as scaling factors allows the region between -1 and $+1$ to be identified as a region of equivalence or a region of safety, which has been emphasised in the graphs by applying a green background colour. It should be noted that the region outside these limits is not labelled in any way. Often the analyst is agnostic about the biological harm or toxicity that would result from values just outside the equivalence limits, and consequently no background colour is provided for the region outside the equivalence limits. Of course, if real toxicity limits were available, these toxicity limits and e.g. a red background colour *outside* such limits could be added to the graphs.

Equivalence limits, either specified by experts or estimated from concurrent or historical data, are needed for safety assessments, at least when real toxicity limits are lacking (Vahl and Kang, 2016). Without equivalence limits or other relevant limits there is no possibility to judge the relevance of observed differences between the test and reference groups. In this paper we have shown three different methods to obtain values for equivalence limits. The simplest case is represented in Fig. 5, where the equivalence limits were set to targeted effect sizes from the literature (Hong et al., 2017). In the maize example from the AMIGA study (Figs. 1 and 2), and also in a study with genetically modified potato (van der Voet et al., accepted), reliance was made on external expert opinion stating that a factor of 2 (e.g. +100% or −50%) on the observed abundances would constitute a level of potential concern for counts at a high level (implemented as 10 or higher). For lower counts, with increased coefficients of variation, we applied an automatic scaling of the equivalence limits based on the Poisson distribution. Computations for cases with externally specified equivalence limits are typically easy: the point estimates and confidence intervals are just re-scaled using the externally specified values.

The alternative and statistically more challenging approach is to derive equivalence limits from data. This was pioneered by van der

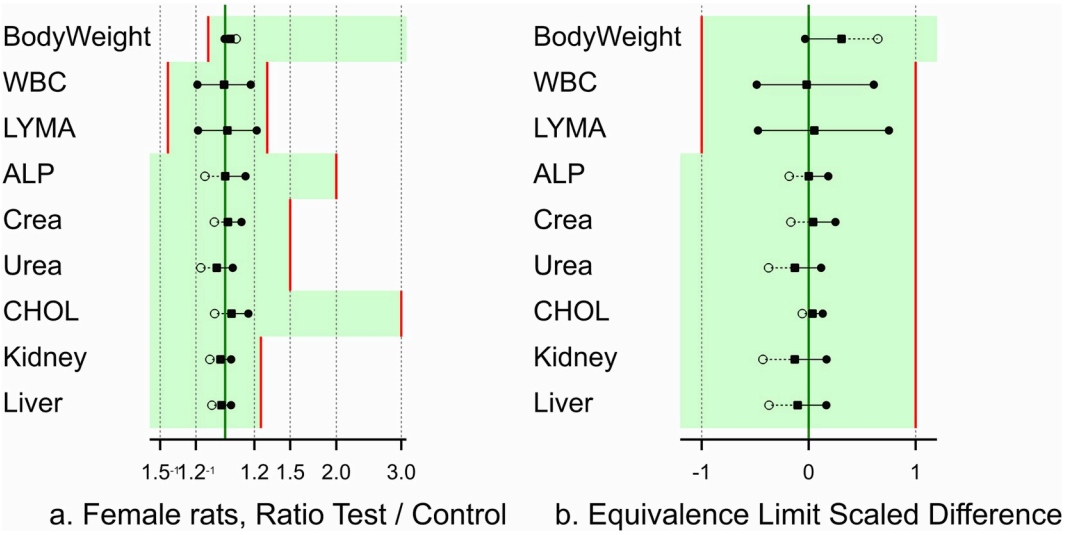

Fig. 5. Test versus Control feeds for female rats for selected variables with equivalence limits defined by targeted effect sizes of Hong et al. (2017). **a**. 90% confidence intervals for the ratio of the mean of GMO and Control feeds. **b**. Equivalence testing, using 90% confidence intervals on the ELSD scale. Open symbols have been used for confidence limits when there is no corresponding equivalence limit.

Voet et al. (2011), and improved by Kang and Vahl (2014), for cases with a set of non-GM genotypes in the same experiment as a GM genotype. Fiducial inference methods were applied to account for the uncertainties in the estimated limits for such experiments. Figs. 3 and 4, for example 2 on the G-TwYST animal feeding study, illustrate the use of historical non-GM data for the same purpose again employing fiducial inference (van der Voet et al., 2017). ELSD graphs have been applied to present all equivalence test results of three rat feeding studies in the G-TwYST project (Steinberg et al., accepted, and references therein). Finally, it can be added that a similar ELSD graph is possible when the equivalence limits are derived from reference data in the same study rather than historical data. This approach would then update the methodology proposed by EFSA for plant compositional data (EFSA, 2010a; EFSA, 2011a; van der Voet et al., 2011).

An additional aim of this paper was to compare the ELSD graph with the SES graph. The SES benchmarks values of 0.2, 0.5 and 0.8, which were suggested by Cohen (1988), were criticised by e.g. Nakagawa and Cuthill (2007) who noted that the biological interpretation of (standardised) effect sizes is context dependent. Such benchmarks are however still commonly used. A recent example is Bell et al. (2017) who employed a value of 0.5. Others have used a benchmark value of 1 (Zeljenková et al., 2014, 2016; Schmidt et al., 2016, 2017; Tulinská et al., 2018) following the possibly mis-interpreted statement in EFSA (2011b), which mentioned an effect size of one SD as a limit of "little toxicological relevance" only in a conditional sentence ("If experience from previous toxicity tests shows that …"). In our view fixed benchmark values for the SES are strongly artificial, i.e. their interpretation assumes that the residual standard deviation would provide a relevant limit for all variables. There is no reason why this would be true in general. Moreover, for any experimental design, there is a direct link between the t-value of the difference test and whether the SES interval falls completely within certain limits. This seems to makes the SES interval superfluous. In addition SES can only be used for continuous data and not for count data analysed by means of a loglinear model such as in example 1. The proposed ELSD method on the other hand has all the positive aspects of the ordinary SES method, such as presenting test results as confidence intervals in a single graph, and reduces the arbitrariness of the scale by forcing safety assessors to specify the scale using equivalence limits in one of a number of possible ways.

6. Conclusions

Equivalence tests are a primary tool for human health and environmental safety assessments. In this paper we have proposed graphical overviews to show the results of equivalence tests. Scaling the differences between a test and a control or reference group with equivalence limits leads to graphs with a correct interpretation regarding equivalence. Equivalence limits can be externally specified values, can be derived from externally specified values allowing for increased variability following a statistical model, or can be estimated from already available (historical) data. Examples for all three types have been shown in this paper.

Finally, the conclusions of the equivalence tests in the examples shown can be summarised in terms of the categories used by EFSA. We stress that our results do not represent formal risk assessments, which address many more aspects, but are just examples meant to illustrate the proposed methodology. We also stress that it is not our intention to suggest that animal feeding studies would have added value for GMO safety assessment over compositional analysis studies (for further discussion of this matter, see Steinberg et al., accepted). Under the chosen assumptions in the example regarding the environmental safety of MON810 maize, full equivalence (equivalence category I) was shown for 79 arthropod taxa counts and equivalence was more likely than non-equivalence (equivalence category II) for 6 arthropod taxa counts. Under the chosen assumptions in the example regarding a rat feeding study to assess the human health safety of NK603 maize, full equivalence (equivalence category I) was shown for 35 endpoints and equivalence was more likely than non-equivalence (equivalence category II) for 5 endpoints. In both examples there were no cases of non-equivalence (equivalence categories III or IV).

Acknowledgements

This paper is a follow-up on work performed in EU projects Assessing and Monitoring the Impacts of Genetically Modified Plants on Agro-ecosystems (AMIGA) and Genetically modified plants Two Year Safety Testing (G-TwYST), funded by the European Commission in the Framework programme 7 under grant agreements 289706 and 632165, respectively. For AMIGA this is publication No. 41. The entomological

count data in example 1 were provided by MC Chueca, M Gonzalez-Nuñez, S Pascual, I Sanchez-Ramos, I Loureiro, I Santin and G Cobos from the National Institute for Agricultural and Food Research and Technology (INIA) in Madrid, Spain, with help on the classification of taxa into functional groups by Joop van Loon (Wageningen University, The Netherlands). The animal study data in Example 2 were provided by Dagmar Zeljenková and team from the Slovak Medical University (SZU) in Bratislava, Slovakia. We thank AMIGA and G-TwYST colleagues for their support of this work.

Transparency document

Transparency document related to this article can be found online at https://doi.org/10.1016/j.fct.2019.02.007.

References

Bell, M.L., Fiero, M.H., Dhillon, H.M., Bray, V.J., Vardy, J.L., 2017. Statistical controversies in cancer research: using standardized effect size graphs to enhance interpretability of cancer-related clinical trials with patient-reported outcomes. Ann. Oncol. 28, 1730–1733. https://doi.org/10.1093/annonc/mdx064.

Brooks, D.R., Bohan, D.A., Champion, G.T., et al., 2003. Invertebrate responses to the management of genetically modified herbicide–tolerant and conventional spring crops. I. Soil-surface-active invertebrates. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 358, 1847–1862.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, second ed. Lawrence Earlbaum Associates, Mahwah, United States.

Delaney, B., 2015. Safety assessment of foods from genetically modified crops in countries with developing economies. Food Chem. Toxicol. 86, 132–143.

de la Poza, M., Pons, X., Farinós, G.P., et al., 2005. Impact of farm-scale Bt maize on abundance of predatory arthropods in Spain. Crop Protect. 24, 677–684.

EFSA, 2010a. Statistical considerations for the safety evaluation of GMOs. EFSA J. 8, 1250. https://doi.org/10.2903/j.efsa.2010.1250.

EFSA, 2010b. Guidance on the environmental risk assessment of genetically modified plants. EFSA J. 8, 1879. https://doi.org/10.2903/j.efsa.2010.1879.

EFSA, 2011a. Scientific Opinion on Guidance for risk assessment of food and feed from genetically modified plants. EFSA J. 9, 2150. https://doi.org/10.2903/j.efsa.2011.2150.

EFSA, 2011b. Guidance on conducting repeated-dose 90-day oral toxicity study in rodents on whole food/feed. EFSA J. 9, 2438. https://doi.org/10.2903/j.efsa.2011.2438.

EMA, 2010. Guideline on the investigation of bioequivalence. Doc. Ref.: CPMP/EWP/QWP/1401/98 Rev. 1/Corr. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/01/WC500070039.pdf.

FDA, 2003. Guidance for Industry: Bioavailability and Bioequivalence Studies for Orally Administered Drug Products — General Considerations. http://www.fda.gov/ohrms/dockets/ac/03/briefing/3995B1_07_GFI-BioAvail-BioEquiv.pdf.

Goedhart, P.W., van der Voet, H., 2017. G-TwYST Study B: A 90-day Toxicity Study in Rats Fed GM Maize NK603. Statistical Report. Report 31.10.17, Biometris, Wageningen, The Netherlands. .

Hong, B., Du, Y., Mukerji, P., Roper, J.M., Appenzeller, L.M., 2017. Safety assessment of food and feed from GM crops in Europe: Evaluating EFSA's alternative Framework for the rat 90-day feeding study. J. Agric. Food Chem. 65, 5545–5560.

Kang, Q., Vahl, C.I., 2014. Statistical analysis in the safety evaluation of genetically-modified crops: equivalence tests. Crop Sci. 54, 2183–2200.

Kelley, K., 2007. Confidence intervals for standardized effect sizes: theory, application, and implementation. J. Stat. Software 20, 8. https://doi.org/10.18637/jss.v020.i08.

Marvier, M., McCreedy, C., Regetz, J., Kareiva, P., 2007. A meta-analysis of effects of Bt cotton and maize on nontarget invertebrates. Science 316, 1475–1477.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman and Hall, U.K.

Nakagawa, S., Cuthill, I.C., 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol. Rev. Camb. Philos. Soc. 82 (4), 591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x.

Oberdoerfer, R.B., Shillito, R.D., de Beuckeleer, M., Mitten, D.H., 2005. Rice (Oryza sativa L.) containing the bar gene is compositionally equivalent to the nontransgenic counterpart. J. Agric. Food Chem. 53, 1457–1465.

OECD, 1993. Safety Evaluation of Foods Derived by Modern Biotechnology: Concepts and Principles. Organisation for Economic Co-operation and Development (OECD), Paris. https://www.oecd.org/science/biotrack/41036698.pdf.

OECD, 1998. Test No. 408: Repeated Dose 90-day Oral Toxicity Study in Rodents. OECD Publishing, Paris. https://doi.org/10.1787/9789264070707-en.

OECD, 2009. Test No. 452: Chronic Toxicity Studies. OECD Publishing, Paris. https://doi.org/10.1787/9789264071209-en.

Perry, J.N., Ter Braak, C.J.F., Dixon, P.M., et al., 2009. Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms. Environ. Biosaf. Res. 8, 65–78.

Schmidt, K., Schmidtke, J., Kohl, C., Wilhelm, R., Schiemann, J., van der Voet, H., Steinberg, P., 2016. Enhancing the interpretation of statistical P values in toxicology studies: implementation of linear mixed models (LMMs) and standardized effect sizes (SESs). Arch. Toxicol. 90, 731–751.

Schmidt, K., Schmidtke, J., Schmidt, P., Kohl, C., Wilhelm, R., Schiemann, J., van der Voet, H., Steinberg, P., 2017. Variability of control data and relevance of observed group differences in five oral toxicity studies with genetically modified maize Mon810 in rats. Arch. Toxicol. 91 (4), 1977–2006. https://dx.doi.org/10.1007/s00204-016-1857-x.

Schuirmann, D.J., 1987. A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. J. Pharmacokinet. Biopharm. 15, 657–680.

Steinberg P., van der Voet H., Goedhart P.W., Kleter G., Kok E.J., Pla M., Nadal A., Zeljenková D., Aláčová R., Babincová J., Rollerová E., Jaďuďová S., Kebis A., Szabova E., Tulinská J., Líšková A., Takácsová M., Lehotská Mikušová M., Krivošíková Z., Spök A., Racovita M., de Vriend H., Alison R., Alison C., Baumgärtner W., Becker K., Lempp C., Schmicke M., Schrenk D., Pöting A., Schiemann J. and Wilhelm R. (accepted). Lack of adverse effects in subchronic and chronic toxicity/carcinogenicity studies on the glyphosate-resistant genetically modified maize NK603 in Wistar Han RCC Rats. Arch. Toxicol.

Tulinská, J., Adel-Patient, K., Bernard, H., Líšková, A., Kuricová, M., Ilavska, S., Horváthová, M., Kebis, A., Rollerova, E., Babincová, J., Alacova, R., Wal, J.-M., Schmidt, K., Schmidtke, J., Schmidt, P., Kohl, C., Wilhelm, R., Schiemann, J., Steinberg, P., 2018. Humoral and cellular immune response in Wistar Han RCC rats fed two genetically modified maize Mon810 varieties for 90 days (EU 7th Framework Programme project GRACE). Arch. Toxicol. 92, 2385–2399. https://doi.org/10.1007/s00204-018-2230-z.

Vahl, C.I., Kang, Q., 2016. Equivalence criteria for the safety evaluation of a genetically modified crop: a statistical perspective. J. Agric. Sci. 154, 383–406.

van der Voet, H., Goedhart, P.W., Kruisselbrink, J.W., 2016. Environmental risk assessment of genetically modified organisms: protocols for statistical aspects of non-target effects field studies, Deliverable 9.4. In: AMIGA project, Report Wageningen University and Research centre, . http://edepot.wur.nl/455501.

van der Voet, H., Goedhart, P.W., Lazebnik, J., van Loon, J.J.A., Kessel, G., Mullins, E., Arpaia, S., accepted. Equivalence analysis to support environmental safety assessment: using nontarget organism count data from field trials with cisgenically modified potato. *Ecol. Evol.* https://doi.org/10.1002/ece3.4964.

van der Voet, H., Goedhart, P.W., Schmidt, K., 2017. Equivalence testing using existing reference data: an example with genetically modified and conventional crops in animal feeding studies. Food Chem. Toxicol. 109, 472–485. https://doi.org/10.1016/j.fct.2017.09.044.

van der Voet, H., Perry, J.N., Amzal, B., Paoletti, C., 2011. A statistical assessment of differences and equivalences between genetically modified and reference plant varieties. BMC Biotechnol. 11, 15.

Zeljenková, D., Ambrušová, K., Bartušová, M., Kebis, A., Kovrižnych, J., Krivošíková, Z., Kuricová, M., Líšková, A., Rollerová, E., Spustová, V., Szabová, E., Tulinská, J., Wimmerová, S., Levkut, M., Révajová, V., Ševčíková, Z., Schmidt, K., Schmidtke, J., La Paz, J.L., Corujo, M., Pla, M., Kleter, G.A., Kok, E.J., Sharbati, J., Hanisch, C., Einspanier, R., Adel-Patient, K., Wal, J.-M., Spök, A., Pöting, A., Kohl, C., Wilhelm, R., Schiemann, J., Steinberg, P., 2014. 90-day oral toxicity studies on two genetically modified maize Mon810 varieties in Wistar Han RCC rats (EU 7th Framework Programme project GRACE). Arch. Toxicol. 88, 2289–2314.

Zeljenková, D., Aláčová, R., Ondrejková, J., Ambrušová, K., Bartušová, M., Kebis, A., Kovrižnych, J., Rollerová, E., Szabová, E., Wimmerová, S., Černák, M., Krivošíková, Z., Kuricová, M., Líšková, A., Spustová, V., Tulinská, J., Levkut, M., Révajová, V., Ševčíková, Z., Schmidt, K., Schmidtke, J., Schmidt, P., La Paz, J.L., Corujo, M., Pla, M., Kleter, G.A., Kok, E.J., Sharbati, J., Bohmer, M., Bohmer, N., Einspanier, R., Adel-Patient, K., Spök, A., Pöting, A., Kohl, C., Wilhelm, R., Schiemann, J., Steinberg, P., 2016. One-year oral toxicity study on a genetically modified maize Mon810 variety in Wistar Han RCC rats (EU 7th framework programme project GRACE). Arch. Toxicol. 90, 2531–2562. https://doi.org/10.1007/s00204-016-1798-4.