Food and Chemical Toxicology 109 (2017) 472–485

Contents lists available at ScienceDirect

Food and Chemical Toxicology

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

Equivalence testing using existing reference data: An example with genetically modified and conventional crops in animal feeding studies

Hilko van der Voet [a], [*], Paul W. Goedhart [a], Kerstin Schmidt [b]

[a] Wageningen University & Research, Biometris, Droevendaalsesteeg 1, 6708PB Wageningen, The Netherlands
[b] BioMath GmbH, Friedrich-Barnewitz-Str. 8, 18119 Rostock-Warnemünde, Germany

ARTICLE INFO

Article history:
Received 10 April 2017
Received in revised form
19 September 2017
Accepted 23 September 2017
Available online 25 September 2017

Keywords:
Food safety
Average equivalence
Distribution-wise equivalence
Linear mixed model
Generalized fiducial inference
Statistical power

ABSTRACT

An equivalence testing method is described to assess the safety of regulated products using relevant data obtained in historical studies with assumedly safe reference products. The method is illustrated using data from a series of animal feeding studies with genetically modified and reference maize varieties. Several criteria for quantifying equivalence are discussed, and study-corrected distribution-wise equivalence is selected as being appropriate for the example case study. An equivalence test is proposed based on a high probability of declaring equivalence in a simplified situation, where there is no between-group variation, where the historical and current studies have the same residual variance, and where the current study is assumed to have a sample size as set by a regulator. The method makes use of generalized fiducial inference methods to integrate uncertainties from both the historical and the current data.

1. Introduction

There are categories of innovative products that are only allowed on the market after a risk assessment has shown that the product is safe for human health and the environment. Examples of regulated products are drugs, pesticides, and, in e.g. Europe, genetically modified organisms (EFSA, 2010b, 2011a). The risk assessment generally involves a comparative trial in which the new product is compared with already established products. Assessing safety by means of such a comparative trial is fundamentally different from data analysis in other scientific fields. In most other testing situations the intention is to prove the existence of a difference or an effect, for example less incidence of a disease or a higher yield of an agricultural crop. Therefore, following the principle that the intention of testing is to prove a hypothesis to be wrong, in such studies a null hypothesis of no effect is tested against an alternative hypothesis that there is an effect. However, statistical hypothesis testing is an asymmetric procedure, and absence of a significant difference cannot be considered as a statement about the truth of the null hypothesis, or “absence of evidence is not evidence of absence” (Altman and Bland, 1995).

It is thus impossible to prove exact equality using statistical procedures. Consequently, safety assessment requires the use of alternative testing procedures, collectively known as equivalence testing, to demonstrate that an effect is small enough (Schuirmann, 1987; Walker and Nowacki, 2011). Equivalence testing is well established as the regulatory required procedure in drug testing (FDA, 2003; EMA, 2010), and has also been proposed in other fields (Garrett, 1997; Cosacov et al., 2008; van der Voet et al., 2011; EFSA, 2011a; Fessel and Snedeker, 2011; Beninger et al., 2012).

Equivalence testing raises two questions. First, on which scale should potential effects be expressed, and second, what is the threshold, also called acceptance criterion or equivalence limit, for ‘small enough’? For example, in testing the equivalence of drugs, regulatory authorities prescribe that certain pharmacokinetic parameters describing the fate of a drug in human subjects should, expressed as an average, differ less than a factor of 1.25 between the tested drug and a reference drug (FDA, 2003; EMA, 2010). This approach is called average equivalence (AE).

The optimal procedure for equivalence testing has been much debated in several fields. In drug bioequivalence testing, it has been proposed to consider the full distribution of measurements rather than an average. This leads to more refined concepts such as population bioequivalence, related to the prescribability of drugs, and

[*] Corresponding author.
E-mail addresses: hilko.vandervoet@wur.nl (H. van der Voet), paul.goedhart@wur.nl (P.W. Goedhart), kerstin.schmidt@biomath.de (K. Schmidt).

https://doi.org/10.1016/j.fct.2017.09.044
0278-6915/

individual bioequivalence, related to the switchability between drugs (Midha et al., 1999; Schall and Endrenyi, 2010). In food safety assessment, the concept of substantial equivalence was introduced (OECD, 1993; Kuiper et al., 2001), but its use was mostly in a non-statistical way and linked to a strong limitation in the number of variables that should be measured. This led to criticism of the concept (Ho and Steinbrecher, 1998; Millstone et al., 1999), although the basic concept of statistical equivalence testing based on an appropriate set of variables was never challenged.

More in general, it is clear that equivalence testing needs prior specification of the set of variables that should be measured, and of the equivalence criterion to be used. The set of variables to test should ideally be limited to those that are informative to risk assessment based on existing data and scientific information available. In this paper our focus is on the statistical procedure for equivalence testing and we will assume that the set of variables has already been established. Different classes of criteria have recently been discussed by Vahl and Kang (2016) in the context of field tests with genetically modified and reference crops. Of special interest to us are scaled average equivalence (SAE) and distribution-wise equivalence (DWE). Equivalence using SAE is assessed by comparing means over experimental units. Moreover the acceptance criterion is not a fixed value, but is scaled to some relevant measure of variation. In DWE, on the other hand, full distributions rather than means are compared.

All forms of equivalence testing need external input, in order to set a fixed equivalence limit (in AE and SAE), to obtain a scaling factor (in SAE), or to obtain a reference distribution (in DWE). In principle there are two approaches to obtain such external input. The first approach is that regulators or experts specify appropriate values. The factor of 1.25 used in drug equivalence testing (FDA, 2003; EMA, 2010) is an example of this approach. However, even experts often find it difficult to specify equivalence limits in this way. The second approach is to use additional data to generate the appropriate input.

In this paper we focus on the use of such additional data which are obtained for products that serve as a reference for the product to be tested. We distinguish between two cases. In the first case, the reference products have been measured under similar circumstances as the test product, so a direct comparison can be made. For example, in field tests, the test and reference varieties of a crop are often planted in the same experiment, and consequently previous work focussed on the comparison of a test variety with a population of reference varieties (van der Voet et al., 2011; Kang and Vahl, 2014; Vahl and Kang, 2016). In the second case, multiple studies are considered, and we cannot exclude unspecified differences between the measurements in different studies.

Here we consider animal feeding studies, in which not many feed varieties can be investigated in any single study, and there may be differences between the studies related to different experimental conditions. Consequently, the basic idea introduced in this paper is to compare the difference between a test (T) and a control (C) variety, obtained simultaneously in a current study, to the typical differences between reference (R) varieties obtained in one or more historical studies. In other words, the equivalence analysis is corrected for between-study differences, and the within-study variation between references R is used to set equivalence limits for the difference between T and C in the current study. Such an approach is in line with the traditional comparative approach in GMO risk assessment that comparison with available data on the nearest comparator, as well as with similar varieties on the market, should form the initial part of the assessment procedure (Kok and Kuiper, 2003). The variation between reference varieties is a point of departure in data-based approaches to set equivalence limits. In the simplest SAE approach of Vahl and Kang (2016) the equivalence limit is some factor times the variance between the reference varieties. Unfortunately, this prevents the ability to obtain useful equivalence limits when there is no, or very small, observed variation between the R varieties. In a DWE approach, as we will see, also the variance within the reference groups contributes to the total variation, allowing to obtain equivalence limits. Therefore DWE testing is the preferred approach when little or no variation between the R varieties is expected, as for example in biochemical and haematological measurements from animal feeding trials.

Vahl and Kang (2016) derived DWE equivalence limits based on a limit case where the variation between the reference varieties was assumed to be much larger than the residual variation. The statistical properties of this procedure in the opposite case, i.e. for a situation with little or no reference variation, are unknown. We propose an alternative strategy based on desired performance of the test in a simplified situation with no reference variation. In short, we will define an equivalence limit as the upper (1-β) limit of the upper (1-α) confidence limit for a statistic that quantifies a comparison between two reference varieties that in reality have no difference. This approach guarantees that equivalence can be declared with probability 1-β when there is no difference in reality in this simplified situation. It may be noted that power for an equivalence test is defined differently than for a traditional difference test. In the latter case it is the probability of detecting a true non-zero difference, but the power of an equivalence test is defined here as the probability of concluding equivalence when in reality there is no difference.

The use of historical data can conceptually be seen as a two-step approach: first equivalence limits are derived from the historical data, and secondly these limits are employed in the equivalence test for the current data. Indeed, the use of external fixed values (such as the factor of 1.25) can also be considered as an instance of such two-step reasoning. The two-step approach is conceptually simple, and a straight-forward use of it has been proposed as a model for GMO safety assessment (EFSA, 2010a; EFSA, 2011b; van der Voet et al., 2011). However, the price paid for this simplicity is that the uncertainty in the estimates of the equivalence limits based on the reference data (step 1) is not accounted for in step 2 (Kang and Vahl, 2014; Vahl and Kang, 2016).

To allow for all uncertainties simultaneously, it is possible to derive methods that integrate both steps and explicitly define criteria which are a function of the model parameters. Among several statistical procedures, methods of generalized fiducial inference (Weerahandi, 1993; Krishnamoorthy and Mathew, 2002; Hannig et al., 2006b, 2016; E et al., 2008; Hannig, 2009; Cisewski and Hannig, 2012) have been found useful for equivalence (McNally et al., 2003; Hannig et al., 2006a; Kang and Vahl, 2014, 2016; Vahl and Kang, 2016). We will therefore construct and use such methods.

The main aims of this paper are therefore:

1. To adapt the equivalence criteria to study-corrected testing, i.e. to base the equivalence test on the difference between a Test and a Control in the current study, rather than on the difference between a Test and a set of reference R varieties;
2. To propose a new equivalence criterion where a full distribution rather than a mean will be compared to a relevant reference distribution, based on desirable power in a simplified situation.
3. To illustrate the use of the proposed method with data from five animal feeding studies which included both conventional and genetically modified feed groups.

2. Data and methods

2.1. Data

Five animal feeding studies with GM and conventional maize have been performed as part of the EU GRACE project (http://www.grace-fp7.eu). Studies A, B, D and E were 90-day studies, and study C was a 1-year study (here we consider only the results obtained at 90 days). The data from the studies are available at the CADIMA website (https://www.cadima.info). These data have been analysed before as part of the GRACE project (Schmidt et al., 2015a,b, 2016, 2017; Zeljenková et al., 2014; 2016). Here we only consider the feed groups containing 33% maize, which was the high dose level in the GRACE studies.

In studies A-E there were 3, 3, 2, 1 and 1 non-GM (reference) groups, respectively, including the isogenic (control) varieties of the GM (test) varieties. All 6 non-GM varieties in studies A and B were different, whereas in studies C, D and E some of these were used again. The variability in the reference data after 90 days in all studies was summarized in Schmidt et al. (2017). Cages were used as the experimental unit in a completely randomized design. The group sample size in studies A and B was 8 cages for all groups. In study C the group sample size was 10 cages (but only 5 for most measurements), and in studies D and E the group sample size was 5 cages.

The statistical analysis presented here was performed with results of 13 haematological variables, 15 clinical chemical variables, the final body weight (at week 13) and 11 organ weights (relative to final body weight). In addition, the growth rate r was estimated from the weights W of each individual animal by fitting the following exponential growth curve against week number t:

$$W = W_{final} - \left(W_{final} - W_{initial}\right) r^t$$

All variables were transformed to the natural logarithmic scale and then averaged to the cage level. Any outliers were identified using Grubb's test at the 1% level. For variables in which outlying observations were identified, statistical analyses were performed with outliers included and excluded to check if this made major differences. Details of the data pre-processing are documented in Supplement 1. All calculations in this work were performed with the statistical program Genstat 18 (https://www.vsni.co.uk/software/genstat/), and should be easily reproducible with other statistical software such as R or SAS.

2.2. Statistical model

The basic structure of the statistical model is similar to the model used by van der Voet et al. (2011), Kang and Vahl (2014) and Vahl and Kang (2016). Let y_{ijk} be the log-transformed response of feed i in study j for unit (cage) k. The following linear mixed model can then be used for the historical studies $j = 1 \ldots n_S$ with reference feeds $i = 1 \ldots n_R$, and the current study with test feed $i = T$ and control feed $i = C$:

$$y_{ijk} = \begin{cases} \mu_R + R_i + S_j + E_{ijk} & i = 1 \ldots n_R \quad j = 1 \ldots n_S \quad k = 1 \ldots n_{ij} \\ \mu_T + F_{Tk} & i = T \qquad\qquad\qquad\quad k = 1 \ldots n_T \\ \mu_C + F_{Ck} & i = C \qquad\qquad\qquad\quad k = 1 \ldots n_C \end{cases} \tag{1}$$

Parameters μ_R, μ_T and μ_C correspond to the expected means for the population of reference feeds R, the test feed T and the control feed C respectively. The random effect R_i denotes deviations from μ_R for the reference feeds in the historic studies and is assumed to follow a normal distribution with mean zero and variance σ_R^2. Study effects S_j, in historical studies $j = 1 \ldots n_S$, are considered to be fixed. This implies that the random effects R_i represent variation between reference feeds within studies. The residual random effects E_{ijk}, for unit $k = 1 \ldots n_{ij}$ in the historical studies, and F_{ik}, for units $k = 1 \ldots n_i$ in the current study with $i = T, C$, are assumed to follow a normal distribution with mean zero and variances σ_E^2 and σ_F^2 respectively. The residual variances in the historic and current studies are thus allowed to be different (see Discussion for alternative possibilities). For specific experimental designs this model can be easily extended with terms for e.g. blocks within studies. Such extensions are not needed for our illustrative data.

Note that in this model there is no formal link between the model for the historical studies and the model for the current study since the two models have no parameters in common. This is the main difference with the models used by van der Voet et al. (2011), Kang and Vahl (2014) and Vahl and Kang (2016), because these were developed for experiments where reference, test and control feeds are simultaneously compared in the same experiments.

2.2.1. Statistical model for historical data

The fixed study effects S_j in the historical studies are considered to be nuisance parameters, and consequently ANOVA mean squares according to Henderson method III can be employed to estimate the variance parameters σ_R^2 and σ_E^2 (Searle et al., 1992, p. 213). Define SS_S as the sums of squares due to differences between studies, $SS_{R|S}$ as the sums of squares due to differences between feeds within studies, and SS_E as the residual sums of squares.

Writing the linear model for the historical data in obvious matrix notation as $y = X_S a + X_R b + e$, the accompanying degrees of freedom are given by $df_S = r[X_S]$, $df_{R|S} = r[X_S\, X_R] - r[X_S]$ and $df_E = N - r[X_S\, X_R]$, where $r[\cdot]$ is the rank of a matrix and $N = \Sigma_i \Sigma_j n_{ij}$ is the total number of units. The sums of squares and degrees of freedom are easily obtained by fitting the linear model. The mean squares ms_E and $ms_{R|S}$, according to Henderson method III for a two-way crossed mixed model (Searle et al., 1992, p. 213), can then be used to estimate the variance components, i.e. $\widehat{\sigma}_E^2 = ms_E = SS_E/df_E$ and $\widehat{\sigma}_R^2 = (ms_{R|S} - ms_E)/n_{eff}$, in which $n_{eff} = h_7/df_{R|S}$, $h_7 = N - \sum_j (\sum_i n_{ij}^2/N_j)$ and $N_j = \sum_i n_{ij}$ (see Searle et al., 1992, p. 211). Note that n_{eff} can be interpreted as the effective unit replication; it equals the common sample size when all n_{ij} are the same. The estimate $\widehat{\sigma}_R^2$ can become negative in which case zero will be used as an estimate. The sums of squares $SS_{R|S}$ and SS_E are independent under normality, even in the unbalanced case (as shown by Searle et al., 1992, p. 73 for a one-way lay-out). Moreover SS_E follows a scaled chi-squared distribution with df_E degrees of freedom, i.e. $SS_E \sim \sigma_E^2\, \chi^2_{df_E}$. For balanced data, i.e. all reference feeds are present in every study and the replication $n_{ij} = n$ is constant, $SS_{R|S}$ also follows a scaled chi-squared distribution now with $df_{R|S}$ degrees of freedom: $SS_{R|S} \sim (n\, \sigma_R^2 + \sigma_E^2)\, \chi^2_{df_{R|S}}$. For unbalanced data with $\sigma_R^2 > 0$ there is no chi-squared distribution associated with $SS_{R|S}$. However the distribution of $SS_{R|S}$ can be approximated for moderately unbalanced data by means of $(n_{eff}\, \sigma_R^2 + \sigma_E^2)\, \chi^2_{df_{R|S}}$; note that the expectation of the two distributions are equal. The quality of the approximation for the illustrative dataset is investigated in Supplement 2.

2.2.2. Statistical model for current data

The current study is conducted to assess equivalence of the test feed T and the control feed C. Note that C represents a control type with special status, and is measured in the same study as T. For example, if T is a feed with a GM variety, C could be a feed with the near-isogenic variety. The parameter of interest is therefore the expected difference between the means $\Delta = \mu_T - \mu_C$. This is estimated by the difference between the observed means in the current study $D = y_{T.} - y_{C.}$, which is distributed as $D \sim N(\Delta,\ a^2\sigma_F^2)$, or equivalently $(D - \Delta)/(a\sigma_F) \sim N(0,\ 1)$, in which $a = \sqrt{1/n_T + 1/n_C}$. The variance σ_F^2 can be estimated by means of the ratio of the sums of squares SS_F and the corresponding degrees of freedom $df_F = n_T + n_C - 2$, and SS_F follows a scaled chi-squared distribution: $SS_F \sim \sigma_F^2\, \chi^2_{df_F}$ independently of D.

In summary the data obtained in the historical and current studies can be summarized by means of the statistics $SS_{R|S},\ SS_E,\ D$ and SS_F which are mutually independent and have distributions of known form.

2.3. Equivalence criteria

In Table 1 six equivalence criteria are presented, organised in three rows representing the type of equivalence (AE, SAE or DWE) and two columns representing whether the mean μ_T of the test T is compared to the mean of the reference feeds μ_R, or to the mean μ_C of the control feed C. All criteria are based on (expected) squared differences between means (AE and SAE) or observations (DWE). The basic idea is that the criterion value should be small enough to be able to conclude equivalence.

We now explain why we use distribution-wise equivalence in which T is compared to C for our illustrative data. First, the 'Compare T to C' criteria are preferred if the R data are mainly or completely obtained from previous studies and there may be large between-study differences. Under these circumstances the estimation of $(\mu_T - \mu_R)^2$ will be imprecise because it will include between study effects. Note that in other types of experiments, such as field trials with several plant genotypes, the 'Compare T to R' approach may be feasible and even preferable, especially when T, C and R groups are simultaneously compared in the same study or studies. This was in fact employed by Vahl and Kang (2016), and their criteria labelled 'SAE-S' and 'DWE-C' are simple rescalings of criteria in the 'Compare T to R' column of Table 1.

Second, the DWE criterion is preferred over the AE and SAE criteria for our illustrative data. A criterion for AE would compare the difference in means between the T and R (or C) groups to a fixed, externally defined, value. For example, symmetric limit values, e.g. ln (0.8) and ln (1.25), can be written as a single limit, in this case $[\ln(1.25)]^2$. However, such external fixed values are often not available. Alternatively, the SAE criterion compares the squared difference to twice the variance σ_R^2. A common choice for an SAE equivalence limit is $z^2_{0.975}$, based on the reasoning that 95% of the reference means $(\mu_R + R_i)$ will lie in the interval $\mu_R \pm z_{0.975}\sigma_R$ (van der Voet et al., 2011; Vahl and Kang, 2016), and therefore most of the differences between two reference means will fall in the interval $\pm\sqrt{2}z_{0.975}\sigma_R$. However, a SAE criterion based on σ_R^2 runs into problems when the variance component σ_R^2 is small such that estimates of this component become zero or very small. For the experimental situation considered in this paper (animal feeding studies) this occurs quite often.

The distribution-wise equivalence (DWE) criterion avoids the problems mentioned for AE and SAE. When comparing T to C it has the form, with i_1 and i_2 representing two reference feeds in the historic studies:

$$\theta = \frac{E\left(y_{Tjk} - y_{Cjk}\right)^2}{E\left(y_{i_1jk_1} - y_{i_2jk_2}\right)^2} = \frac{(\mu_T - \mu_C)^2\ +\ 2\sigma_F^2}{2\sigma_R^2 +\ 2\sigma_E^2} = \frac{\Delta^2\ \ +\ 2\sigma_F^2}{2\sigma_R^2 +\ 2\sigma_E^2} \qquad (2)$$

The DWE criterion considers the two populations of measurements on the experimental units in their entirety, not just the mean parameters μ_T and μ_C. The numerator of θ is the expectation of the squared difference between a unit with the Test feed and a unit with the Control feed in the current study. This is compared to the denominator which is the expectation of the squared difference between a unit with a reference feed and a unit with another reference feed both in the same (historical) study. Large values of θ may indicate lack of equivalence. Note that the additions $+2\sigma_F^2$ in the numerator and $+2\sigma_E^2$ in the denominator are the only differences between the DWE and the SAE criterion. The reasons to prefer DWE over SAE for data similar to our example are further discussed in the Discussion section.

2.4. Interval estimation using generalized fiducial inference

An estimate of θ is readily obtained by plugging in the estimates defined in the previous paragraphs for the parameters $\Delta,\ \sigma_F^2,\ \sigma_R^2$ and σ_E^2. These estimates are based on the summary statistics D and SS_F for the current study and $SS_{R|S}$ and SS_E for the historic studies. Based on these summary statistics $\boldsymbol{Y} = (D,\ SS_F,\ SS_{R|S},\ SS_E)$ confidence limits for θ can be obtained using Generalized Fiducial Inference (GFI), a theory that has been developed recently on the basis of Fisher's fiducial argument (see e.g. the review in Hannig et al., 2016). The general idea of GFI is that a Generalized Fiducial Distribution (GFD) for the unknown parameter, here θ, is constructed by inverting the data-generating equation. The data-generating equation is the model (1) extended with the calculations leading to the summary statistics $\boldsymbol{Y}$. In a more general notation it can be specified as $\boldsymbol{Y} = \boldsymbol{G}(\boldsymbol{U}, \xi)$, for data $\boldsymbol{Y}$, a deterministic function $\boldsymbol{G}(\cdot,\cdot)$, fixed but unknown parameters ξ and random components $\boldsymbol{U}$ with completely known distributions. If it is possible to invert the equation to $\xi = \boldsymbol{Q}(\boldsymbol{U}, \boldsymbol{Y})$, where $\boldsymbol{Q}(\cdot,\cdot)$ are Generalized Pivotal Quantities (GPQs), then simulating a large number (in this paper 10 000) of realizations for the random components $\boldsymbol{U}$ produces an empirical distribution representing the GFD for the unknown parameters ξ. Confidence limits can be obtained as empirical percentiles of the GFD.

As a simple example, the data generating equation for the sums of squares SS_F can be written as $SS_F = \sigma_F^2\, U_F$ where U_F is a random value from a chi-squared distribution: $U_F \sim \chi^2_{df_F}$. Inverting this

Table 1

Six equivalence criteria for comparing a Test (T) type to either a population of Reference (R) types or a single Control (C) type. The entry for DWE comparing T and C ($\boldsymbol{\theta}$) is the criterion used in this paper.

Criterion	Compare T to R	Compare T to C
Average equivalence (AE)	$(\mu_T - \mu_R)^2$	$(\mu_T - \mu_C)^2$
Scaled average equivalence (SAE)	$\frac{(\mu_T - \mu_R)^2}{E(R_{i_1} - R_{i_2})^2} = \frac{(\mu_T - \mu_R)^2}{2\sigma_R^2}$	$\frac{(\mu_T - \mu_C)^2}{E(R_{i_1} - R_{i_2})^2} = \frac{(\mu_T - \mu_C)^2}{2\sigma_R^2}$
Distribution-wise equivalence (DWE) i, i_1 and i_2 represent reference feeds in the historic study	$\frac{E(y_{Tjk} - y_{ijk})^2}{E(y_{i_1jk_1} - y_{i_2jk_2})^2} =$ $= \frac{(\mu_T - \mu_R)^2 + \sigma_R^2 + \sigma_E^2 + \sigma_F^2}{2\sigma_R^2 + 2\sigma_E^2}$	$\boldsymbol{\theta} = \frac{\boldsymbol{E}(\boldsymbol{y}_{\boldsymbol{Tjk}} - \boldsymbol{y}_{\boldsymbol{Cjk}})^2}{\boldsymbol{E}(\boldsymbol{y}_{\boldsymbol{i_1jk_1}} - \boldsymbol{y}_{\boldsymbol{i_2jk_2}})^2} =$ $= \frac{(\boldsymbol{\mu_T} - \boldsymbol{\mu_C})^2 + 2\boldsymbol{\sigma_F^2}}{2\boldsymbol{\sigma_R^2} + 2\boldsymbol{\sigma_E^2}}$

gives $GPQ(\sigma_F^2) = SS_F/U_F$ as the Generalized Pivotal Quantity for the parameter σ_F^2. Given an observed SS_F, $GPQ(\sigma_F^2)$ can be used to derive a confidence interval for the parameter σ_F^2. In this simple example this can be done exactly by employing percentiles of the $\chi^2_{df_F}$ distribution, giving the classical confidence interval for σ_F^2. In a more general approach, multiple simulations from the appropriate distribution, in this case $\chi^2_{df_F}$, can be employed to generate the Generalized Fiducial Distribution of the unknown parameter σ_F^2. Finally, 2.5th and 97.5th percentiles of the simulated GFD then give numerical estimates of the limits of a 95% two-sided confidence interval.

In the previous paragraph it was shown that the data can be summarized by means of the independent statistics $D, SS_F, SS_{R|S}$ and SS_E. These can be used in the following way to provide GPQs for the parameters $\Delta,\ \sigma_F^2,\ \sigma_R^2$ and σ_E^2, using inversion of the data generating equations given in the previous paragraphs, and defining $Z \sim N(0,1)$, $U_E \sim \chi^2_{df_E}$, $U_F \sim \chi^2_{df_F}$, $U_{R|S} \sim \chi^2_{df_{R|S}}$:

$$GPQ\left(\sigma_F^2\right) = SS_F/U_F \tag{3a}$$

$$GPQ(\Delta) = D + a\,Z\,\sqrt{GPQ\left(\sigma_F^2\right)} \tag{3b}$$

$$GPQ\left(\sigma_E^2\right) = SS_E/U_E \tag{3c}$$

$$GPQ\left(\sigma_R^2\right) = \max\left[0, \left(SS_{R|S}\middle/U_{R|S} - GPQ\left(\sigma_E^2\right)\right)\middle/n_{eff}\right] \tag{3d}$$

The maximisation in (3d) prevents negative estimates of σ_R^2, which is a commonly used strategy in estimating variance components. Note that percentiles of the simulated distribution $GPQ(\Delta)$ are just numerical approximations of the usual t-distribution based confidence limits for Δ (see e.g. Example 1 in Hannig et al., 2016). The GPQ for θ in equation (2) is then given by replacing parameters by their respective GPQ's.

$$GPQ(\theta) = \frac{[GPQ(\Delta)]^2 + 2GPQ\left(\sigma_F^2\right)}{2GPQ\left(\sigma_R^2\right) + 2GPQ\left(\sigma_E^2\right)} \tag{4}$$

So, conditional on the observed summary statistics D, SS_F, $SS_{R|S}$ and SS_E, equation (4) can be used to simulate a large number of values giving the generalized fiducial distribution of θ.

Large values of θ may indicate lack of equivalence, therefore the $100(1-\alpha)\%$ percentile of the GFD defined by (4), $\theta_{\text{upp}} = P_{100(1-\alpha)}(GPQ(\theta))$, serves as an upper confidence limit for the magnitude of Δ (whether positive or negative) translated to the θ scale.

2.5. Distribution-wise equivalence test at the $\boldsymbol{\theta}$ scale

Distribution-wise equivalence testing requires specification of a 'safe' case which forms the basis for equivalence. To set an equivalence limit at the chosen θ scale, we introduce a 'safe' case where there is no difference between T and C (i.e. $\Delta = 0$) and therefore values of θ are relatively low. For any dataset derived under this no-difference hypothesis, we define θ^0_{upp} as an upper $100(1-\alpha)\%$ confidence limit for θ related to large values of Δ (whether positive or negative). Our requirement is that for simulations from the 'safe' case, this upper confidence limit will remain below the equivalence limit (and therefore indicate equivalence) with a pre-set power $1-\beta$. In other words, the equivalence limit will be set as the $100(1-\beta)\%$ percentile of simulated upper confidence limits.

The 'safe' case should not depend on any of the current or historical data, but it should only depend on the design parameters of the historical studies. We define a simplified 'safe' case by making a number of assumptions:

a) There is no difference between the test feed T and control feed C in the current study, or $\Delta = 0$ (this corresponds to the idea of a power analysis for the equivalence test);
b) There is no variability between the reference feeds in the historical studies, or $\sigma_R^2 = 0$;
c) The residual variance in the historic and current studies are identical, or $\sigma_F^2 = \sigma_E^2$;
d) The regulator will set a minimum sample size n_0 for the T and C groups to be compared. The idea is that the current experiment should not be allowed to be too small, which would lead to unacceptable wide equivalence bands. A value for n_0 may be inspired by external guidance or by values in the historical data (summarized by the effective sample size n_{eff}). In principle, n_0 will be used to define the structure of the current study when calculating the equivalence limit, i.e. $a_0 = \sqrt{2/n_0}$ and $df_0 = 2n_0 - 2$.

Using a superscript '0' to denote that the distributions are derived under these additional assumptions, the distributions of the four summary statistics are then given by $D^0 \sim N(0, a_0^2\sigma_E^2)$, $SS_F^0 \sim \sigma_E^2\ \chi^2_{df_0}$, $SS_E^0 \sim \sigma_E^2\ \chi^2_{df_E}$, and $SS^0_{R|S} \sim \sigma_E^2\ \chi^2_{df_{R|S}}$. All four distributions have the common variance parameter σ_E^2 which will cancel out in equation (4), and without loss of generality we can set $\sigma_E^2 = 1$. The four distributions thus do not depend on any parameter; they only depend on the design of the historical studies, through the various sample sizes and degrees of freedom, and the regulatory n_0. A simulation approach is now used to find out which values of the DWE criterion θ can be expected in this simplified situation:

1) Simulate the data summary statistics D^0, SS_F^0, SS_E^0 and SS_R^0 according to the distributions given above, with $\sigma_E^2 = 1$;
2) Use equation (4) to simulate the GFD^0 of θ^0 for this simulated dataset using a large number of samples (10 000 in this paper);
3) Summarize the simulated distribution of θ^0 by means of the $100(1-\alpha)$ percentile $\theta^0_{upp} = P_{100(1-\alpha)}(GPQ^0_\theta)$ of GFD^0;
4) Repeat steps 1–3 many times (10 000 in this paper) to obtain the distribution, say G^0_α, of θ^0_{upp} under the additional assumptions;
5) Set the equivalence limit θ_0 to the $100(1-\beta)$ percentile of the distribution G^0_α: $\theta_0 = P_{100(1-\beta)}(\theta^0_{upp})$.

Under these assumptions we would like to reject the null hypothesis of no equivalence with a large probability, say $1-\beta$. The probability $1-\beta$ is the power of the equivalence test in the simplified situation. Note that θ_0 only depends on the design values of the historical studies and on three regulatory values, n_0, α and β.

Using the equivalence limit θ_0 as calculated above, the equivalence test can be carried out for a dataset by calculating the generalized fiducial distribution given by equation (4) for the observed summary statistics of the historical and current data. Employing the $100(1-\alpha)$ percentile θ_{upp} of this distribution, the null hypothesis of no equivalence will be rejected when $\theta_{upp} < \theta_0$.

2.6. Distribution-wise equivalence test at the equivalence limit scaled difference (ELSD) scale

The scale of the DWE criterion θ is not easily understood. It is therefore preferable to re-express results on a better recognizable scale. First we express results on the more familiar difference (Δ) scale between the test T and the control C. Note that for the illustrative data, variables were log-transformed such that differences in fact relate to ratios on the original scale. Secondly, for a full integration of uncertainties we perform an additional scaling to what we call the equivalence limit scaled difference (ELSD) scale. As the name indicates, +1 and −1 represent the equivalence limits on this scale.

First, on the Δ scale, the classical confidence interval can be obtained, either using a parametric calculation or using percentage points of $GPQ(\Delta)$. We now derive how the equivalence limit θ_0 translates to the Δ scale. Define $A = 1/\sqrt{2GPQ(\sigma_F^2)}$ and $B = [GPQ(\sigma_R^2)\ +\ GPQ(\sigma_E^2)]/GPQ(\sigma_F^2)$ then

$$GPQ(\theta) = \frac{[GPQ(\Delta)]^2\ +\ 2GPQ(\sigma_F^2)}{2GPQ(\sigma_R^2)\ +\ 2GPQ(\sigma_E^2)} = \ \frac{[A\cdot GPQ(\Delta)]^2 + 1}{B} \qquad (5)$$

For the purpose of deriving symmetric limits on the Δ scale, the equivalence limit θ_0 therefore corresponds to equivalence limits $\Delta_{0,low}, \Delta_{0,upp}$ with distributions which are found by inverting equation (5):

$$EL_{low}\ ,\ EL_{upp} = GPQ(\Delta_{0,low}\ )\ , GPQ(\Delta_{0,upp}\ ) = \pm\sqrt{(B\theta_0 - 1)}\Big/A$$

provided that $B\theta_0 - 1 > 0$. Note that A and B are transformed GFDs, and therefore the equivalence limits on the Δ scale are also GFDs. Also note that the equivalence limits $\Delta_{0,low}, \Delta_{0,upp}$ are symmetric around zero. The medians of the two GFDs can be used as point estimates, and using $100\alpha/2$ and $100(1-\alpha/2)$ percentile points, both GFDs can be represented by a confidence interval. We now have obtained three confidence intervals all based on a GFD: for Δ, for $\Delta_{0,low}$ and for $\Delta_{0,upp}$, respectively. It is not clear how to compare the interval for Δ with the intervals for the equivalence limits. Obviously, although the Δ scale has the advantage of familiarity, it is not the appropriate scale for a direct representation of the equivalence test.

Therefore, to obtain a scale where equivalence can be represented directly, we apply a further scaling using the (positive) upper equivalence limit. Define

$$GPQ(ELSD) = \frac{GPQ\ (\Delta)}{GPQ(\Delta_{0,upp})}$$

where ELSD is short for Equivalence Limit Scaled Difference. Now in this new measure all uncertainties are integrated into one distribution, and the equivalence condition is met when an appropriate confidence interval derived from $GPQ(ELSD)$ lies completely within the interval (−1,1). We now describe how to construct this confidence interval.

If $B\theta_0 - 1 \le\ 0$, the corresponding estimates of $\Delta_{0,low}, \Delta_{0,upp}$ are set both to zero, and it is clear that equivalence cannot be established. In order to have a visual indication on the ELSD scale, and also to define a distribution of $GPQ(ELSD)$ from which empirical confidence limits can be calculated, we set $GPQ(ELSD)$ in those cases to a large negative or positive value, e.g. $BIG = 2$, with sign equal to that of $GPQ(\Delta)$. The full definition of $GPQ(ELSD)$ thus becomes

$$\begin{cases} GPQ(ELSD) = \dfrac{GPQ\ (\Delta)}{GPQ(\Delta_{0,upp})} & \text{if } B\theta_0 - 1 > 0 \\ GPQ(ELSD) = -BIG & \text{if } B\theta_0 - 1 \le 0 \text{ and } GPQ(\Delta) < 0 \\ GPQ(ELSD) = +BIG & \text{if } B\theta_0 - 1 \le 0 \text{ and } GPQ(\Delta) \ge 0 \end{cases}$$

The one-sided test using $GPQ(\theta)$ can make no distinction between positive and negative differences. One-sided intervals on the θ scale therefore correspond to two-sided intervals on the ELSD scale with lower and upper confidence limits which are perforce symmetric around 0. By a simple search algorithm we therefore identify a limit $ELSD_{\text{lim}}$ such that $P[GPQ(ELSD) < -ELSD_{lim}] + P[GPQ(ELSD) > ELSD_{lim}] = \alpha$.

The ELSD scale can also be used for the difference test. Since $GPQ(\Delta)$ defines the classical confidence interval for the difference Δ, and the sign of $GPQ(ELSD)$ is always equal to that of $GPQ(\Delta)$, it follows that the classical difference test can be performed by checking whether the value zero is included in the confidence interval given by the $100\alpha/2$ and $100(1-\alpha/2)$ percentile points $ELSD_{100\alpha/2}$ and $ELSD_{100(1-\alpha/2)}$ of $GPQ(ELSD)$. There are thus two relevant intervals on the ELSD scale: the symmetric interval around zero for the equivalence test and the interval which corresponds to the classical difference test. For visualisation of both, we propose an interval with the most appropriate upper and lower limit, such that both the difference and equivalence test can be performed by this single interval. Depending on which of the difference test percentiles is closest to zero, we propose to plot intervals with limits $(ELSD_{low},\ ELSD_{upp})$:

$$\begin{cases} \left(ELSD_{100\alpha/2}\ , ELSD_{lim}\right) & \text{if } abs\left(ELSD_{100\alpha/2}\right) < abs\left(ELSD_{100(1-\alpha/2)}\right) \\ \left(-ELSD_{lim}\ ,\ ELSD_{100(1-\alpha/2)}\right) & \text{if } abs\left(ELSD_{100\alpha/2}\right) \ge abs\left(ELSD_{100(1-\alpha/2)}\right) \end{cases}$$

These intervals, while covering between $100(1-3\alpha/2)$ and $100(1-\alpha)$ % of the distribution, can then be used for equivalence and difference testing. The hypothesis of no difference is rejected in case the interval does not contain zero, while the non-equivalence hypothesis is rejected when the interval fully lies inside the interval (−1,1).

2.7. Simplified model assuming $\boldsymbol{\sigma_R^2} = 0$

In animal feeding studies there is often little evidence of variation between reference feeding groups. If the assumption $\sigma_R^2 = 0$ is made from the beginning, the statistical model (1) simplifies to

$$y_{ijk} = \begin{cases} \mu_R + S_j + E_{Rjk} & i = R & j = 1\ldots n_S & k = 1\ldots n_{Rj} \\ \mu_T + F_{Tk} & i = T & & k = 1\ldots n_T \\ \mu_C + F_{Ck} & i = C & & k = 1\ldots n_C \end{cases} \qquad (6)$$

The DWE equivalence criterion (2) reduces to

$$\theta = \ \frac{\Delta^2\ +\ 2\sigma_F^2}{2\sigma_E^2} \qquad (7)$$

The GPQ calculations described in the previous sections can be easily modified for this simplified model by omitting all terms involving σ_R^2.

2.8. Power analysis

For a power analysis 1000 datasets were simulated following the design of the A, B and C studies for the historical data, with 8 replications in studies A and B and 5 replication is study C, and a simple two-group comparison for the current study, with $n_T = n_C = 8$ replications and $df_F = 14$ degrees of freedom. It was assumed that these sample sizes were acceptable for regulators ($n_o = 8$), and residual variances σ_E^2 and σ_F^2 were arbitrarily set equal to 1. The between reference variance σ_R^2 was set to 0, 0.25 and 1 in different simulations. Moreover the effect of a larger replication in the current study was investigated using values of 16, 32, 64 and 128 for n_T and n_C. In a third set of simulations the degrees of freedom in the current study was varied between $df_F = 14$ (corresponding with two groups of 8) and 200 (corresponding with a current study where many additional feed groups would provide additional degrees of freedom to estimate the residual variation). The number of datasets used for calculation of θ_0 was 10 000 and the number of GPQ samples was also set to 10 000. The level of significance α was set at 5% and the probability for an equivalence outcome in the simplified case $(1-\beta)$ was set at 95%.

Fig. 1 shows the simulated power against the true effect size expressed as Δ/σ_E. The probability for an equivalence outcome in the simplified case, here set at $1-\beta = 0.95$, is indeed attained in the simplified situation where $\Delta = 0$ and $\sigma_R^2 = 0$. The power decreases for increasing true effect sizes. With 8 replications (Fig. 1a) feeds with a true difference of one standard deviation are still judged equivalent in 80% of the cases. True variability between reference feeds ($\sigma_R^2 = 0.25$ or 1) increases the power, and allows feeds with larger true differences to be considered equivalent. With more replications in the current study (Fig. 1b) the power curves start at higher power and are steeper. With more residual degrees of freedom in the current study (Fig. 1c) the power curves also start at higher values, but the differences are smaller in comparison to panel (b). Finally, note that the true effect size with power equal to 0.05 in panels (b) and (c) appears to be largely independent from the replication in the current study.

3. Results

As an illustration of the proposed methodology for equivalence testing, the observed differences between Test (GM at 33%) and Control fed male or female rats in study D or E of the Grace project were analysed against the background of the data for the Reference and Control (all non-GM) male rats in studies A, B and C. The method has been applied both with and without the identified outliers, and with different values of the regulatory sample size, i.e. $n_0 = 5$ or 8, corresponding to the typical sample sizes in studies D/E and A/B, respectively. We assumed regulatory error rates α and β both equal to 5%. All results are presented graphically in Supplement 3. In this section we just show one example to illustrate the general points.

The summary statistics for the male rats of study D, after excluding 12 outliers (see Supplement 1), are given in Table 2 where, for a better interpretability, the sums of squares are recalculated as standard deviations ($s = \sqrt{SS/df}$). For the historical data, the variation between Reference groups s_R (estimated from 4 degrees of freedom) is always smaller than the variation within groups s_E, and for 20 out of 41 variables it is estimated as zero or near zero. For BodyWeight and growthR the effective sample size is larger because these variables are available for 10, rather than 5, cages in study C. Relative organ weights were only obtained from historical studies A and B, therefore the effective replication is smaller. Furthermore slight variations in degrees of freedom for error df_E and effective replication n_{eff} are due to exclusion of outliers. In the last two columns of Table 2 we compare estimates of the standard deviations related to the denominator of the DWE criterion θ, i.e. $\sqrt{s_R^2 + s_E^2}$ for model (1) and s_E for model (6). It can be noted that $\sqrt{s_R^2 + s_E^2}$ is only slightly larger than s_E, with a maximum change of less than 5% (for TAG).

In a first example, we have assumed that the sample sizes in the historical studies (characterised by n_{eff} between 7.7 and 10.5) roughly represent regulatory needs, and purely for illustration the regulatory minimum sample size was set to $n_0 = 8$ (cages). It must be noted that in study D the sample sizes were smaller, $n_T = n_C = 5$, so that a power of 0.95 will not be reached in the simplified situation.

In Fig. 2 the observed differences are expressed as ratios of Test vs. Control, i.e. differences at a logarithmic scale. The black line segments in Fig. 2 are just the GFD versions of the ordinary 95% confidence intervals for the differences at the log scale, and their intersection with the vertical line at ratio = 1 indicates non-

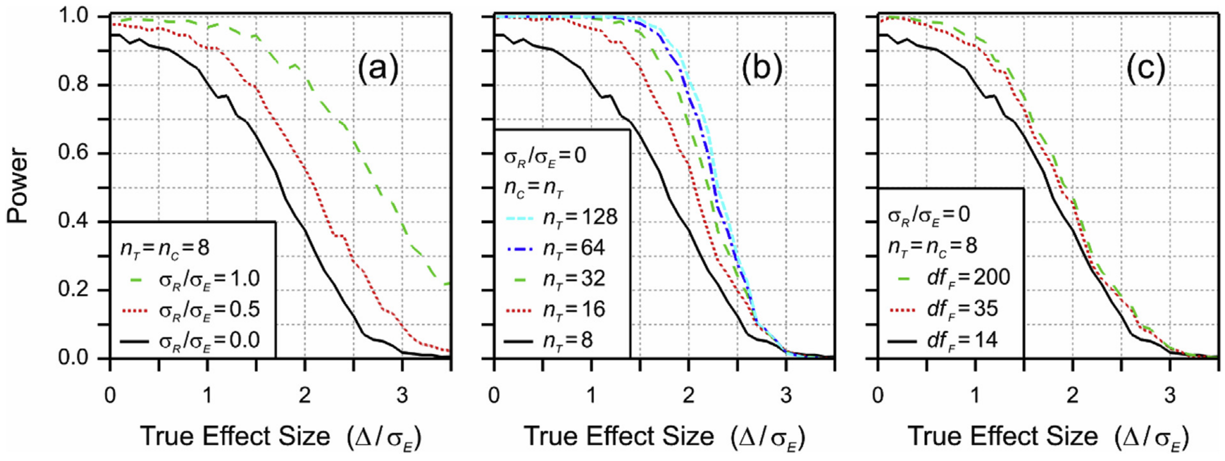

Fig. 1. Power of equivalence test in simulations (1000 runs) vs. a background in historical studies with similar design as that of most variables in the A, B and C studies (8 replicates in A and B, 5 replicates in C). Power is shown as a function of the true effect size between T and C (expressed in SD). (a) results for three values of the true variation between reference foods ($\boldsymbol{\sigma_R/\sigma_E} = 0, 0.5, 1$), $\boldsymbol{n_T} = \boldsymbol{n_C} = 8$ replications; (b) results for larger sample sizes in the current study, $\boldsymbol{n_T} = \boldsymbol{n_C} = 8, 16, 32, 64, 128$ replications, for $\boldsymbol{\sigma_R/\sigma_E} = 0$; (c) results when using additional groups to have more degrees of freedom in the current study, $\boldsymbol{df_F} = 14, 35, 200$, for $\boldsymbol{n_T} = \boldsymbol{n_C} = 8$ replications and $\boldsymbol{\sigma_R/\sigma_E} = 0$.

Table 2
Summary statistics Study D against background from studies A-C. Male rats, outliers excluded. $\boldsymbol{D}$ = difference Test (GM, 33%) and Control in current study, $\boldsymbol{s_F}$ is current-study within-group standard deviation with $\boldsymbol{df_F}$ degrees of freedom, $\boldsymbol{s_E}$ and $\boldsymbol{s_R}$ are historical-study within- and between-group standard deviations with $\boldsymbol{df_E}$ and $\boldsymbol{df_R}$ degrees of freedom, $\boldsymbol{n_{eff}}$ is effective sample size in the historical studies. All statistics are calculated under model (1), except $\boldsymbol{s_E}$ in the last column, which is calculated under model (6).

study	D (GM vs. Control)			A, B, C (non-GM)						
variable	$\boldsymbol{D}$	$\boldsymbol{s_F}$	$\boldsymbol{df_F}$	$\boldsymbol{s_E}$	$\boldsymbol{df_E}$	$\boldsymbol{s_R}$	$\boldsymbol{df_R}$	$\boldsymbol{n_{eff}}$	$\sqrt{\boldsymbol{s_R^2}+\boldsymbol{s_E^2}}$	$\boldsymbol{s_E}$ (M6)
BodyWeight	0.012	0.052	8	0.056	61	0.000	4	10.5	0.056	0.055
growthR	0.000	0.019	8	0.019	60	0.000	4	10.3	0.019	0.019
Kidney	−0.041	0.054	8	0.054	42	0.000	4	8.0	0.054	0.053
Spleen	0.070	0.080	8	0.086	42	0.000	4	8.0	0.086	0.086
Liver	0.028	0.049	8	0.051	40	0.006	4	7.7	0.051	0.051
AdrenGl	0.026	0.085	8	0.120	42	0.000	4	8.0	0.120	0.119
Lung	−0.041	0.064	8	0.077	42	0.035	4	8.0	0.085	0.083
Heart	0.007	0.035	8	0.051	42	0.013	4	8.0	0.052	0.052
Thymus	−0.083	0.087	8	0.184	42	0.004	4	8.0	0.184	0.184
Pancreas	0.063	0.133	8	0.145	41	0.056	4	7.8	0.156	0.153
Testis	−0.057	0.074	8	0.076	42	0.000	4	8.0	0.076	0.075
Epididymis	−0.041	0.104	8	0.081	42	0.000	4	8.0	0.081	0.079
Brain	−0.016	0.055	8	0.050	42	0.016	4	8.0	0.052	0.052
WBC	0.244	0.154	8	0.220	51	0.089	4	9.2	0.237	0.232
RBC	−0.023	0.034	8	0.053	51	0.000	4	9.2	0.053	0.052
HGB	0.011	0.028	8	0.043	50	0.006	4	9.1	0.044	0.043
HCT	−0.004	0.023	8	0.051	51	0.000	4	9.2	0.051	0.050
MCV	0.019	0.020	8	0.019	51	0.000	4	9.2	0.019	0.018
MCH	0.033	0.032	8	0.027	50	0.000	4	9.1	0.027	0.026
MCHC	0.015	0.017	8	0.016	50	0.003	4	9.1	0.016	0.016
PLT	−0.071	0.111	8	0.270	51	0.000	4	9.2	0.270	0.262
LYMcount	0.228	0.153	8	0.217	51	0.076	4	9.2	0.230	0.225
Lymphocyte	0.058	0.046	8	0.045	51	0.003	4	9.2	0.045	0.045
Neutrophil	−0.254	0.153	8	0.166	51	0.000	4	9.2	0.166	0.165
Monocyte	0.065	0.279	8	0.294	51	0.126	4	9.2	0.320	0.312
Eosinophil	0.478	0.380	8	0.457	51	0.164	4	9.2	0.486	0.476
ALP	−0.206	0.162	8	0.176	51	0.000	4	9.2	0.176	0.172
ALT	0.045	0.119	8	0.120	49	0.000	4	8.9	0.120	0.119
AST	0.171	0.159	8	0.156	50	0.000	4	9.1	0.156	0.155
Alb	−0.023	0.051	8	0.055	50	0.000	4	9.1	0.055	0.054
Glu	−0.038	0.090	8	0.146	51	0.038	4	9.2	0.150	0.149
Krea	0.000	0.126	8	0.132	51	0.029	4	9.2	0.135	0.134
TP	−0.020	0.041	8	0.040	51	0.000	4	9.2	0.040	0.040
Urea	−0.106	0.062	8	0.100	51	0.057	4	9.2	0.116	0.111
CHOL	0.043	0.120	8	0.110	51	0.054	4	9.2	0.123	0.119
Ca	0.017	0.016	8	0.064	51	0.012	4	9.2	0.065	0.065
Cl	−0.016	0.010	8	0.051	51	0.002	4	9.2	0.051	0.051
K	0.066	0.071	8	0.112	51	0.000	4	9.2	0.112	0.110
Na	−0.003	0.013	8	0.056	51	0.000	4	9.2	0.056	0.056
P	0.074	0.072	8	0.101	50	0.029	4	9.1	0.105	0.103
TAG	0.065	0.233	8	0.294	51	0.178	4	9.2	0.343	0.328

significance of a traditional two-sided difference test. Neutrophils and Urea are seen to be significantly smaller in the Test group than in the Control group, and WBC and LYMcount are significantly greater. The new elements in Fig. 2 are the estimated equivalence limits EL_{low} and EL_{upp} together with their 95% confidence bounds. The variables have been sorted within their category in order of increasing median equivalence limit. Except for MCH all interval estimates of the ratio Test/Control (the black lines) are between the median equivalence limits (the red limits), including those for Neutrophils and Urea. In several cases, however, the 95% confidence interval for the ratio and the 95% confidence interval for the equivalence limit do overlap. The Δ scale in Fig. 2 cannot be used to perform an equivalence test. For that purpose we consider the ELSD scale in Fig. 3 where all uncertainties have been incorporated in the *ELSD* statistic. As explained in the Methods section, the confidence limit closest to 0, which can be used for the two-sided difference test, excludes $100\alpha/2$ percent of the distribution, whereas the other confidence limit, to be used in the one-sided equivalence test, excludes between $100\alpha/2$ and 100α percent of the distribution. From Fig. 3 it follows that 36 out of the 41 intervals (78%) are between the standardised equivalence limits -1 and $+1$. For 5 of the 41 variables the experiment with 5 cages per groups is insufficient to produce ELSD intervals that are fully in the equivalence range. Note, however, that all point estimates are still in the interval ±1, so in the terminology of EFSA (2010a, 2011a,b) equivalence is ‘more likely than not’.

Just for illustration of the method we have also assumed that the sample size of study D was in agreement with regulatory needs, e.g. the regulatory minimum sample size was set to $n_0 = 5$ (cages). Fig. 4 shows that in this case all confidence intervals are in the interval ±1. Note that we are separately testing equivalence for 41 variables without any attempt to correct for the multiplicity, therefore around 5% of the intervals extending outside the ±1 limits (around 2 out of 41) can be expected without statistically indicating a lack of equivalence.

Fig. 4 also shows the effects of including or excluding the identified outliers. In this example all 12 outliers occurred in the historical data (none in the current data), and therefore the result of excluding outliers is a more narrow equivalence bandwidth and consequently a somewhat wider ELSD interval. The conclusions of the equivalence tests are however not changed.

Finally, Fig. 5 compares the ELSD intervals derived under the standard model (1), which includes a term R_i for between-reference group variation, and the simplified model (6), where $\sigma_R^2 =$

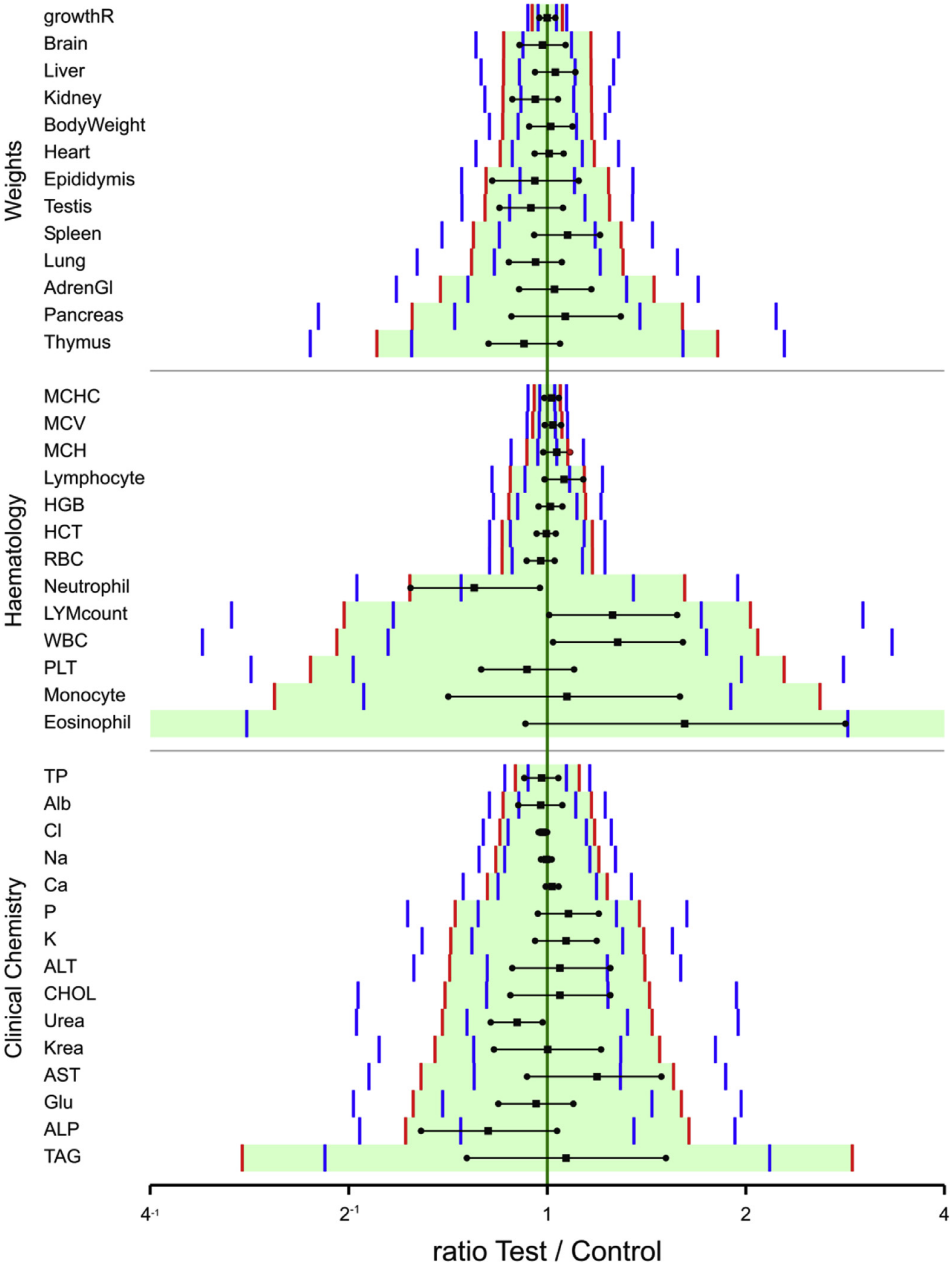

Fig. 2. Confidence intervals Test vs. Control in study D, Males, compared to the non-GM background in studies ABC (12 outliers excluded), requiring $\boldsymbol{n}_0 = 8$ as the minimum number of cages per group. Shown are the observed ratios corresponding to differences $\boldsymbol{D}$ on the log scale, with 95% confidence limits (black symbols and line segments), the median equivalence limits EL (red bars delimiting a green equivalence band width), and the 2.5% and 97.5% confidence limits for EL (blue bars). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

0 is assumed throughout. The differences found in the ELSD intervals were very small.

4. Discussion

A new method for safety assessment of innovative products has been proposed and was applied to existing data. The advantage of the proposed method is that it employs available data from previous studies to characterise the variation of observations on different reference groups that are assumed to be safe. The equivalence limits are derived in such a way that an experiment with two groups with no difference would give a confidence interval for the group difference that would lie in between the equivalence limits with a predetermined probability. In this respect the proposed approach falls under the concept of tolerance intervals which have similar properties (Kang and Vahl, 2014; Hong et al., 2014).

Statistical testing is an asymmetric procedure, both for difference and for equivalence testing. Note that in our procedure we attempt to demonstrate equivalence by rejecting a hypothesis of

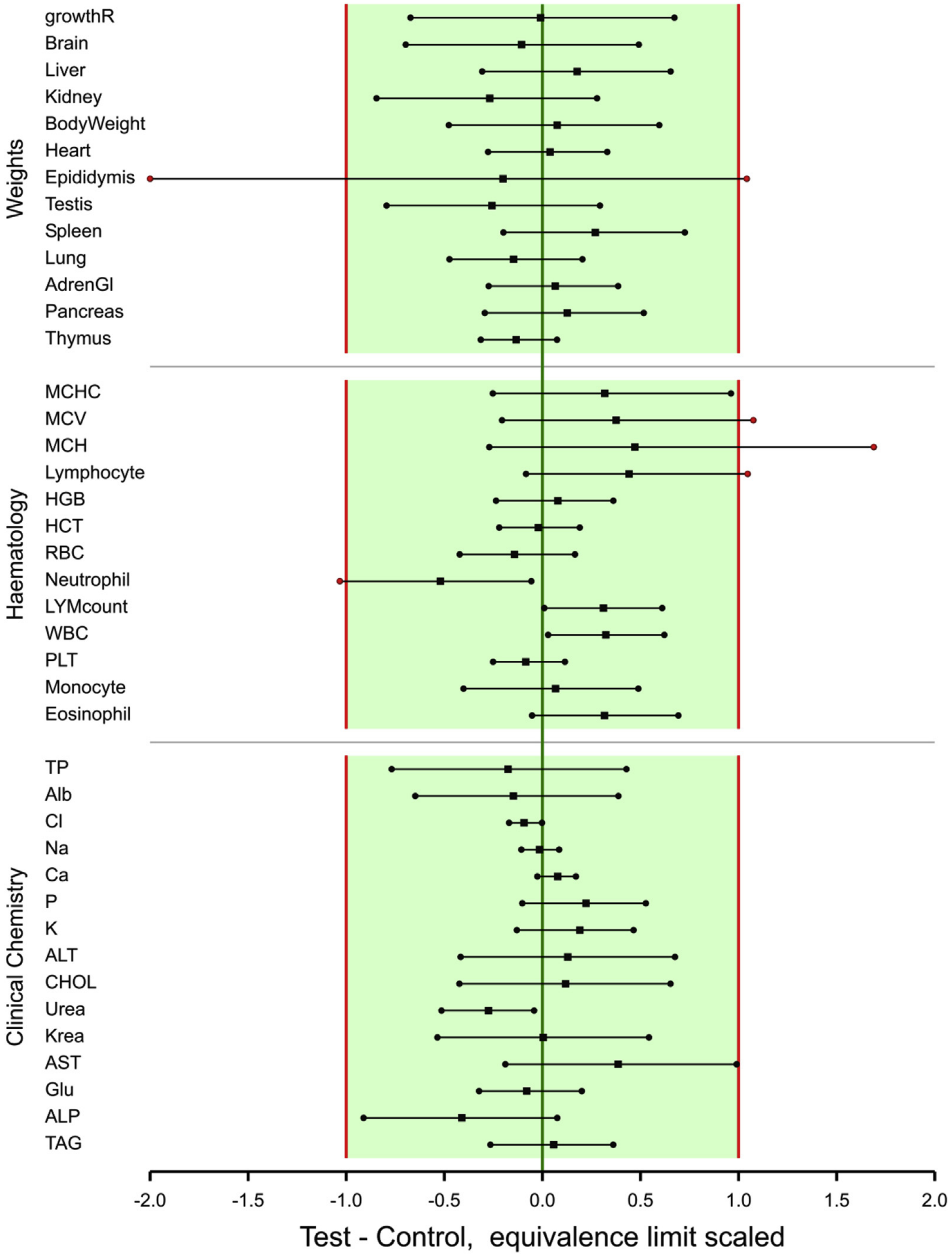

Fig. 3. Equivalence testing Test vs. Control in study D, Males, compared to the non-GM background in studies ABC (12 outliers excluded), requiring $\mathbf{n}_0 = 8$ as the minimum number of cages per group. Median ELSD with 92.5–95% confidence interval (see text).

non-equivalence. Consequently a failure to do so is not a proof of non-equivalence. Further, non-equivalence of a Test group versus a Control groups does not imply a verdict about safety, but only about an observed difference which is larger than has been seen previously for the reference groups. Safety assessment is always more than just a statistical equivalence test, and needs the interpretation of the results by experts.

In equivalence testing for drugs, expert-based fixed values are commonly used for equivalence limits (FDA, 2003; EMA, 2010). When experts are able to set such limits, this is a reasonable procedure. For cases where experts have difficulties to translate their expertise to numerical values even when historical datasets on reference groups are available, the proposed procedure may be helpful as an alternative.

In the context of animal feeding studies the use of standardised effect size (SES) has been suggested by EFSA (2011b). SES is the observed difference between group means divided by one standard deviation between experimental units. The reasoning of EFSA was: '*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*'. In other words, EFSA sketches an example where one SD is considered an appropriate equivalence limit. In the absence of other externally provided equivalence limits SES has been used in previous analyses of the

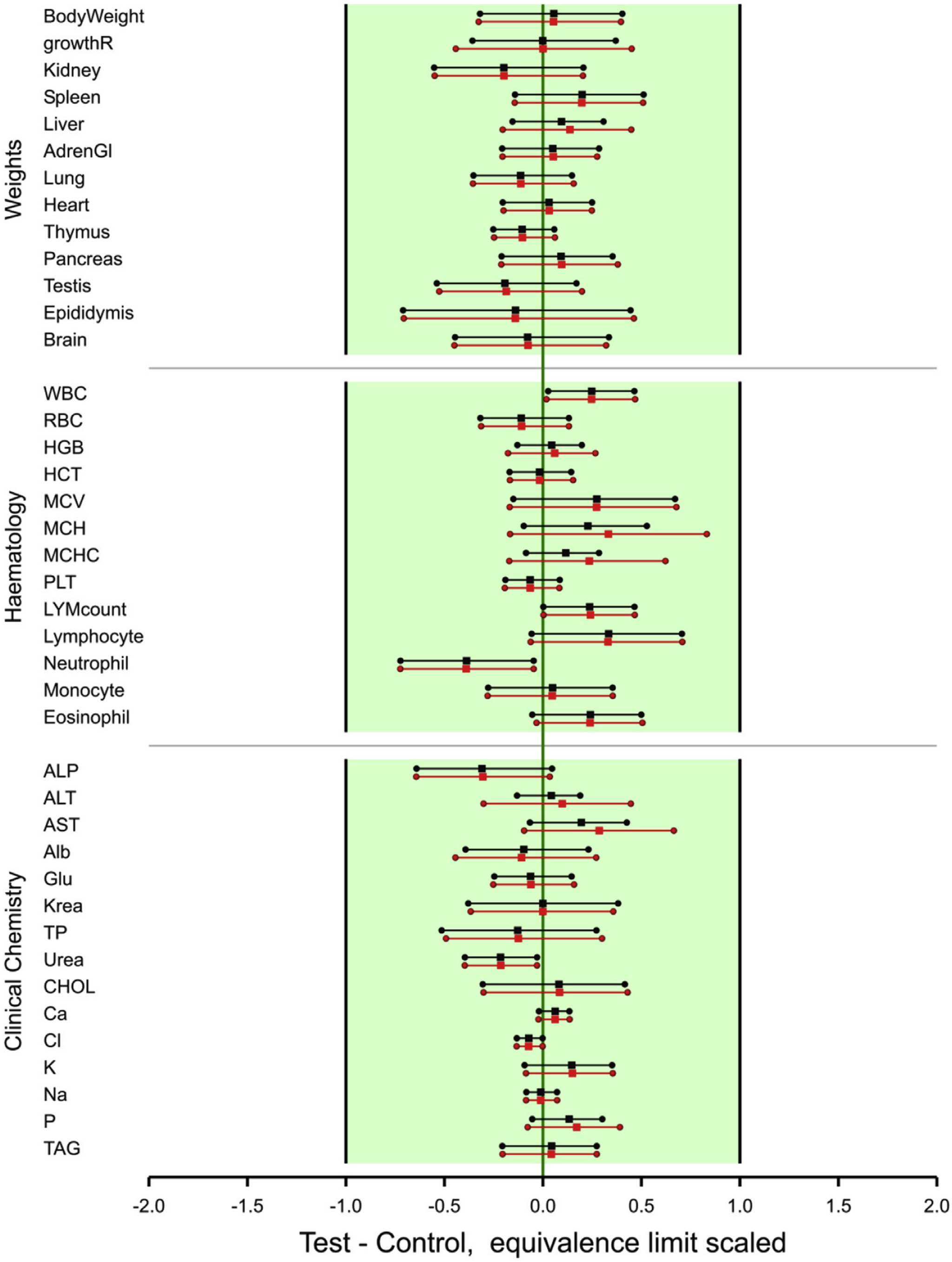

Fig. 4. Equivalence testing Test vs. Control in study D, Males, compared to the non-GM background in studies ABC, requiring $\mathbf{n}_0 = 5$ as the minimum number of cages per group. Median ELSD with 92.5–95% confidence interval (see text). Results shown before (upper, black) and after (lower, red) excluding 12 outliers from the historical data (two outliers for Liver and ALT, one outlier for growthR, Pancreas, HGB, MCH, MCHC, AST, Alb and P). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

data from the GRACE study to provide a first step to show that equivalence testing is to be preferred over difference testing in the safety assessment of GM plants (Schmidt et al., 2015a, b, 2016; 2017; Zeljenková et al., 2014; 2016). In this paper we have proposed an alternative to the purely hypothetical assumption of EFSA that one SD would be a reasonable equivalence limit, by using historical data to specify equivalence limits.

We have assumed potentially different residual variances for the historical and the current study (σ_E^2 and σ_F^2, respectively). If the variances would in fact be the same, there would be a benefit of pooling the variance estimates for the historical and current data. However, in regulatory equivalence testing, there is a potential danger in using the variances from the current experiment (typically under control of the applicant) in a role where a larger variance effectively may widen the equivalence region [EL_{low}, EL_{upp}] of the type as shown in Fig. 2. This could lead to a situation where lack of precision in the current experiment would lead to easier acceptance of equivalence, which is undesirable. For this reason we chose the model such that the historical data (reviewed and accepted by the regulators) set a standard for variances and

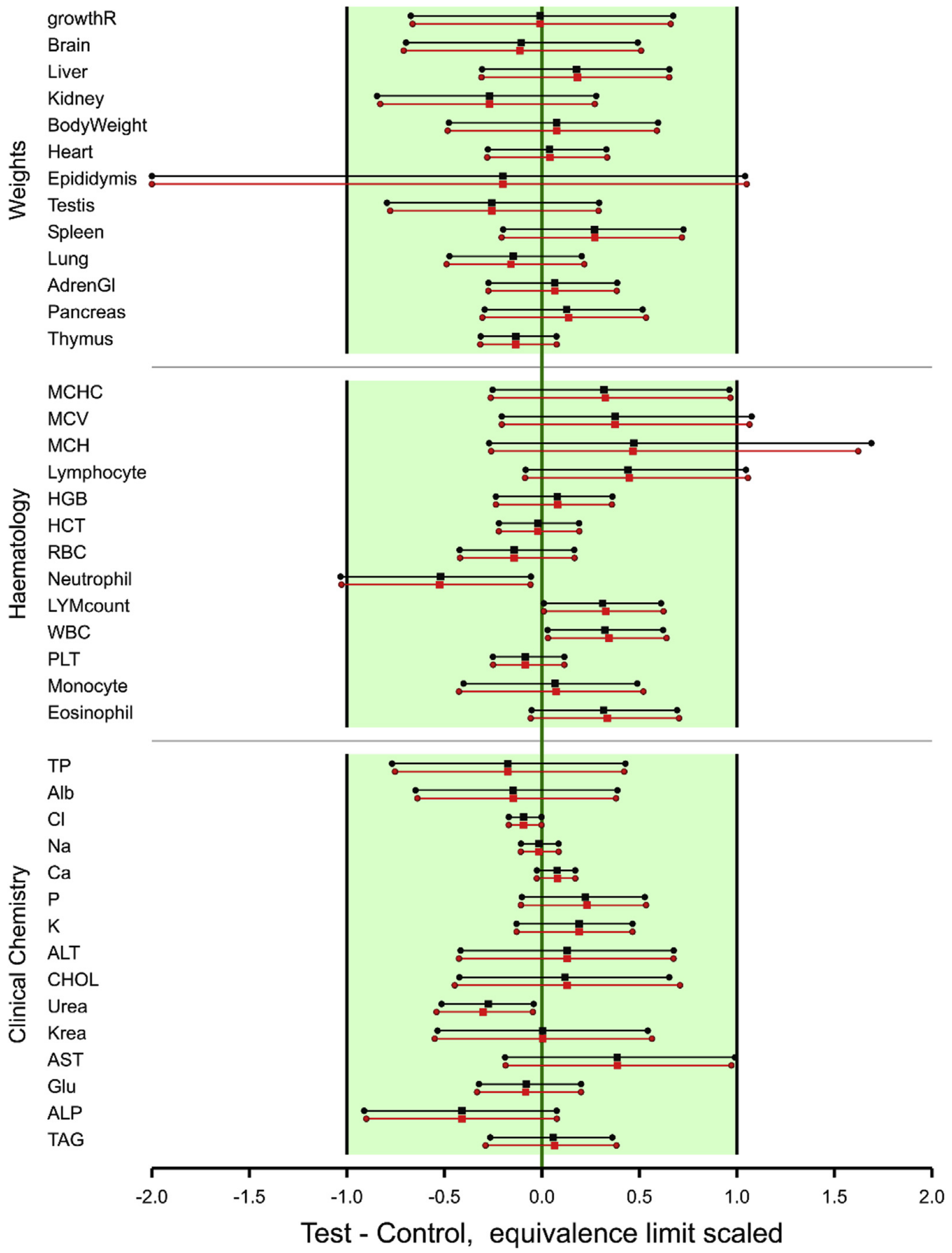

Fig. 5. Equivalence testing Test vs. Control in study D, Males, compared to the non-GM background in studies ABC (12 outliers excluded), requiring $\mathbf{n}_0 = 8$ as the minimum number of cages per group. Results for the standard model (1) (upper, black) and simplified model (6) with $\boldsymbol{\sigma}_R^2 = 0$ (lower, red). Median ELSD with 92.5–95% confidence interval (see text). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

consequently equivalence limits. This still leaves the possibility of assuming a single residual variance in model (1) or (6) which uses two GPQ estimates in equation (4): a pooled estimate in the numerator and an estimate based on only the historical data in the denominator. However, in that case, a researcher who conducts a current experiment with improved precision would not gain the full benefit of this improvement. The proposed approach can be adapted to the needs of regulators in specific cases regarding the allowed role of data from the current study for estimating precision and also for estimating between-reference variation if more references are available.

We have argued that DWE is preferable over SAE when the between feed variance component (σ_R^2) is zero or small. Application of SAE is clearly impossible when σ_R^2 is zero, but in theory it could still be applied for any positive estimate. However, in our example the point estimates s_R were mostly below 0.1, and a difference of less than one standard deviation corresponds to a ratio lower than $exp(0.1) - 1 = 10\%$ at the original scale (only three estimates were higher: 0.178 for TAG, 0.164 for Eosinophils and 0.126 for Monocytes). Commonly, but not always, equivalence testing considers wider limits than 10%, and in such cases using estimates of variation with lower values seems contrary to the intentions.

Perhaps more importantly, in our example the between feed variance component (σ_R^2) could only be estimated with four degrees of freedom. As a consequence the estimates s_R in Table 2 are very imprecise. For example, 95% confidence intervals for σ_R using equation (3d) are 0.000–0.037 for BodyWeight (point estimate 0.000), and 0.067–0.563 for TAG (point estimate 0.178). In using SAE for equivalence testing, employing $0.5\, GPQ^2(\Delta)/GPQ^2(\sigma_R^2)$ in analogy with the proposed approach for DWE, we would use the 95th percentile of the distribution of this ratio. This only makes sense when at least 95% of the GPQ values for σ_R^2 are positive, but this was only the case for the five variables TAG (99.6%), Urea (99.3%), CHOL (98.3%), Monocyte (96.0%) and Lung (95.6%). Therefore, application of SAE is not an option for most variables in our case study. However, for four of the five remaining variables, the highly uncertain σ_R^2 estimate still caused the upper 95% limit of the SAE criterion to attain much higher values than the SAE criterion $z_{0.975}^2 = 3.84$. The 95% upper limits were 3.15 (TAG), 11.07 (Urea), 14.60 (CHOL), 46.91 (Monocyte) and 55.85 (Lung), respectively, so that only for TAG equivalence would be shown by SAE. Note that the uncertainty of σ_R^2 plays a smaller role in the DWE criterion, because σ_E^2 is added to the denominator which, for the current data set, is typically a larger value with less uncertainty.

When estimates of σ_R^2 are small, another option is to omit σ_R^2 from the DWE model altogether. This is model (6), and for the case study in this paper the results were very similar to the results of model (1). Experience with more data sets would be needed to justify a general preference for making the extra assumption $\sigma_R^2 = 0$.

Estimating historical variation between feeds with only four degrees of freedom is not ideal, and is due to the lack of a long history of comparable data for non-GM foods in the test facility. The proposed method using model (1) will be most useful in an infrastructure, for example a routine testing laboratory, where a longer series of historical studies, each involving at least two non-GM feeding groups, are available for the characterisation of reference variation. A requirement of at least two reference groups per feeding study would be new, but is helpful for establishing the appropriate background data for equivalence testing using model (1).

It should be stressed that the analysis in this paper is just an illustration of methodology, and not a real safety assessment. As a first point, a full safety assessment would require case-by-case interaction with regulators and risk assessors regarding the choice of variables that is needed to cover the spectrum of possible unintended effects. Such interaction could suggest a sufficient set of variables based on potential and plausible pathways to harm. A second point would be a choice between external, expert-set equivalence limits and equivalence procedures based on historical data. In the latter case, a third point to be decided by regulators would be the specification of experimental effort, for example in the form of a fixed minimum sample size n_0, e.g. conforming OECD guidance, or by specifying that a current experiment would need to have at least the same sample size as has been used in historical experiments.

Acknowledgements

This work was carried out as part of the project Genetically modified plants Two Year Safety Testing (G-TwYST), financially supported by the Seventh Framework Programme (FP7) of the European Community for Research, Technological Development and Demonstration Activities, grant agreement No. 632165. Animal feeding studies were carried out as part of the project GMO Risk Assessment and Communication of Evidence (GRACE, FP7 grant agreement No. 311957). We thank our colleagues in the GRACE and G-TwYST projects for their support of this work.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.fct.2017.09.044.

Transparency document

Transparency document related to this article can be found online at https://doi.org/10.1016/j.fct.2017.09.044.

References

Altman, D.G., Bland, J.M., 1995. Statistics notes: absence of evidence is not evidence of absence. Br. Med. J. 311, 485.

Beninger, P.G., Boldina, I., Katsanevakis, S., 2012. Strengthening statistical usage in marine ecology. J. Exp. Mar. Biol. Ecol. 426–427, 97–108.

Cisewski, J., Hannig, J., 2012. Generalized fiducial inference for normal linear mixed models. Ann. Stat. 40, 2102–2127.

Cosacov, A., Nattero, J., Cocucci, A.A., 2008. Variation of pollinator assemblages and pollen limitation in a locally specialized system: the oil-producing nierembergia linariifolia (solanaceae). Ann. Bot. 102, 723–734.

E, L., Hannig, J., Iyer, H.K., 2008. Fiducial intervals for variance components in an unbalanced two-component normal mixed linear model. J. Am. Stat. Assoc. 103, 854–865.

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 (11), 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 (5), 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.

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.

Fessel, G., Snedeker, J.G., 2011. Equivalent stiffness after glycosaminoglycan depletion in tendon — an ultra-structural finite element model and corresponding experiments. J. Theor. Biol. 268, 77–83.

Garrett, K.A., 1997. Use of statistical tests of equivalence (bioequivalence tests) in plant pathology. Phytopathology 87, 372–374.

Hannig, J., 2009. On generalized fiducial inference. Stat. Sin. 19, 491–544.

Hannig, J., E L, Abdel-Karim, A., Iyer, H., 2006a. Simultaneous fiducial generalized confidence intervals for ratios of means of lognormal distributions. Austrian J. Stat. 35, 261–269.

Hannig, J., Iyer, H., Patterson, P., 2006b. Fiducial generalized confidence intervals. J. Am. Stat. Assoc. 101, 254–269.

Hannig, J., Iyer, H., Lai, R.C.S., Lee, T.C.M., 2016. Generalized fiducial inference: a review and new results. J. Am. Stat. Assoc. 111, 1346–1361.

Ho, M.W., Steinbrecher, R., 1998. Fatal flaws in food safety assessment: critique of the joint FAO/WHO biotechnology and food safety report. Environ. Nutr. Interact. 2, 51–84.

Hong, B., Fisher, T.L., Sult, T.S., Maxwell, C.A., Mickelson, J.A., Kishino, H., Locke, M.E.H., 2014. Model-based tolerance intervals derived from cumulative historical composition data: application for substantial equivalence assessment of a genetically modified crop. J. Agric. Food Chem. 62, 9916–9926.

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

Kang, Q., Vahl, C.I., 2016. Statistical procedures for testing hypotheses of equivalence in the safety evaluation of a genetically modified crop. J. Agric. Sci. 154, 1392–1412.

Krishnamoorthy, K., Mathew, T., 2002. Assessing occupational exposure via the one-way random effects model with balanced data. J. Agric. Biol. Environ. Stat. 7, 440–451.

Kok, E.J., Kuiper, H.A., 2003. Comparative safety assessment for biotech crops. Trends Biotechnol. 21, 439–444.

Kuiper, H.A., Kleter, G.A., Noteborn, H.P.J.M., Kok, E.J., 2001. Assessment of the food safety issues related to genetically modified foods. Plant J. 27, 503–528.

McNally, R.J., Iyer, H., Mathew, T., 2003. Tests for individual and population bioequivalence based on generalized p-values. Stat. Med. 22, 31–53.

Midha, K.K., Rawson, M.J., Hubbard, J.W., 1999. Prescribability and switchability of highly variable drugs and drug products. J. Control. Release 62, 33–40.

Millstone, E., Brunner, E., Mayer, S., 1999. Beyond 'substantial equivalence'. Nature 401, 525–526.

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.

Schall, R., Endrenyi, L., 2010. Bioequivalence: tried and tested. Cardiovasc. J. Afr. 21, 7–9.

Schmidt, K., Schmidtke, J., Schmidt, P., 2015a. Statistical Analysis Report on a Chronic Toxicity (1-Year) Study of Rats With Mon810 Maize. Retrieved from. www.cadima.de.

Schmidt, K., Schmidtke, J., Schmidt, P., 2015b. Statistical Analysis Report on a Repeated Dose 90-Day Oral Toxicity/Longitudinal Study in Rodents With Mon810 Maize. Retrieved from. www.cadima.de.

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. Archives 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. Pharmacokin Biopharm. 15, 657–680.

Searle, S.R., Casella, G., McCulloch, C.E., 1992. Variance Components. Wiley, New York.

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., 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.

Walker, E., Nowacki, A.S., 2011. Understanding equivalence and noninferiority testing. J. General Intern. Med. 26 (2), 192–196. https://doi.org/10.1007/s11606-010-1513-8.

Weerahandi, S., 1993. Generalized confidence intervals. J. Am. Stat. Assoc. 88, 899–905.

Zeljenková, D., Ambrušová, K., Bartušová, M., Kebis, A., Kovriźnych, 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., Kovriźnych, 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.