Archives of Toxicology (2020) 94:1135–1149
https://doi.org/10.1007/s00204-020-02690-w

REGULATORY TOXICOLOGY

Alternatives to statistical decision trees in regulatory (eco-) toxicological bioassays

Felix M. Kluxen[1] · **Ludwig A. Hothorn**[2]

Received: 23 June 2019 / Accepted: 25 February 2020 / Published online: 19 March 2020

Abstract
The goal of (eco-) toxicological testing is to experimentally establish a dose or concentration–response and to identify a threshold with a biologically relevant and probably non-random deviation from "normal". Statistical tests aid this process. Most statistical tests have distributional assumptions that need to be satisfied for reliable performance. Therefore, most statistical analyses used in (eco-)toxicological bioassays use subsequent pre- or assumption-tests to identify the most appropriate main test, so-called statistical decision trees. There are however several deficiencies with the approach, based on study design, type of tests used and subsequent statistical testing in general. When multiple comparisons are used to identify a non-random change against negative control, we propose to use robust testing, which can be generically applied without the need of decision trees. Visualization techniques and reference ranges also offer advantages over the current pre-testing approaches. We aim to promulgate the concepts in the (eco-) toxicological community and initiate a discussion for regulatory acceptance.

Keywords Hazard identification · Hazard characterization · Assumption tests · Pre-tests · Robust statistics · Regulatory toxicology

Introduction

Within the plant protection product or chemicals regulatory frameworks in the EU (Reg. (EC) 1107/2009 and 1907/2006, respectively), data generation using (eco-) toxicological bioassays is highly regulated. It requires good laboratory practise (GLP)- and Organisation for Economic Co-operation and Development (OECD) test guideline-compliance. Those who intend to register a chemical or product propose an initial data interpretation, which is peer-reviewed by competent national or international authorities. The interpretation may also be subject to guidance or regulation documents. A decision-aiding guide in the process is assessing statistical significance of the observed effects by statistical testing. Ideally, this would allow the distinction of random and treatment-related effects (note: treatment, dose and concentration are used interchangeably in the manuscript). Practically, this is limited because according to the predominant statistical method used in experimental toxicity, null hypotheses statistical testing (NHST), certain error rates must be accepted, both for falsely attributing and for falsely not attributing statistical significance.

At least in regulatory toxicology, testing against a negative control—pair-wise or by a Dunnett (Dunnett 1955) test—is the standard statistical method used (Hamada 2018; Hothorn 2014; Jarvis et al. 2011; Na et al. 2014) and is also specifically required for hazard identification. One example is found in OECD test guideline 487 (2016b), the in vitro micronucleus test. Here, a test item is considered, among other criteria, to possess genotoxic potential when "[…] at least one of the test concentrations exhibits a statistically significant increase compared with the concurrent negative control […]". It is currently not possible to satisfy this requirement by dose–response modelling.

There is some regulatory guidance available on how the statistical analysis should be conducted (OECD 2014a, b, c) but most analyses resort to decision-tree approaches for pair-wise or Dunnett-type testing: The basic idea is to

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00204-020-02690-w) contains supplementary material, which is available to authorized users.

✉ Felix M. Kluxen
felix.kluxen@adama.com

[1] ADAMA Deutschland GmbH, Edmund-Rumpler-Strasse 6, 51149 Cologne, Germany

[2] Lauenau, Germany

data-dependently select an appropriate main test, out of several alternatives, using pre-tests.

Decision trees might vary depending on the endpoints that are investigated, i.e. continuous (e.g. organ weights), proportion (e.g. tumour incidences), count (e.g. number of micronuclei) or time-to-event data (e.g. mortality). However, most decision trees consider continuous data.

The pre-tests, also called assumption tests, commonly assess normal distribution, variance homogeneity (compare, e.g. Schmidt et al. 2016)—sometimes conditional after transformation—or outliers (OECD 2010). The main test used in decision trees for continuous data, is usually an analysis of variance (ANOVA) with subsequent post hoc comparisons, a Dunnett test or rank-sum non-parametric alternatives (Dunn 1961). Sometimes the Dunnett test is used as a post hoc test, even though the Dunnett test is actually an ANOVA conducted with a priori-set contrasts to test against control, as described in Hothorn (2016b).

Decision-tree approaches are focusing on endpoints in a design with a concurrent negative control and some treatment or dose groups (usually three in standard regulatory toxicity testing). They provide a seemingly consistent approach to the optimal analysis of real data in a well-documented GLP-style. There are however several deficiencies with the approach because the tests do not necessarily achieve what they intend to do, i.e. selecting for an optimal main test, disregard a potential treatment effect and result in main tests with different or partly absent effect sizes. There are further issues with subsequent (multiple) testing in general, all of which is reviewed below.

We therefore propose to use robust test methods, which have very relaxed distributional assumptions, and make the application of decision trees unnecessary. However, all generic application of statistical tests is problematic when conducted in isolation, i.e. without relating to biological effect size or plausibility, because the tests have very limited power in the small sample size design common in toxicological bioassays. Alternatively, graphical analysis of model assumptions can be used to argue for the standard approach, which may however be too labour intensive for a generic analysis, as conducted for bioassays with multiple endpoints. Also reference ranges can be used to identify toxic responses.

The presented methods are considered helpful for both experimental and regulatory toxicologists that evaluate the outcome of bioassays.

Problem

Decision trees are a collection of subsequent binary yes-and-no decisions based on data, assumptions and thresholds. In statistical testing, a decision is based on a statistical test result. Depending on the statistical decision tree's design, there are up to 2–5 decisions before the hypotheses of interest are tested; see Fig. 1 for a hypothetical example or refer to, for example, Kobayashi et al. (2008) or Schmidt et al. (2016).

It follows that the power of the main test and hence the relevant false negative decision rate, strongly depends on the performance of its pre-tests. As biological data are investigated, which often have heterogeneous variance, show non-normality, extreme values and are often of relatively small sample size (Hothorn 2016a; Wilcox 2012), the power of

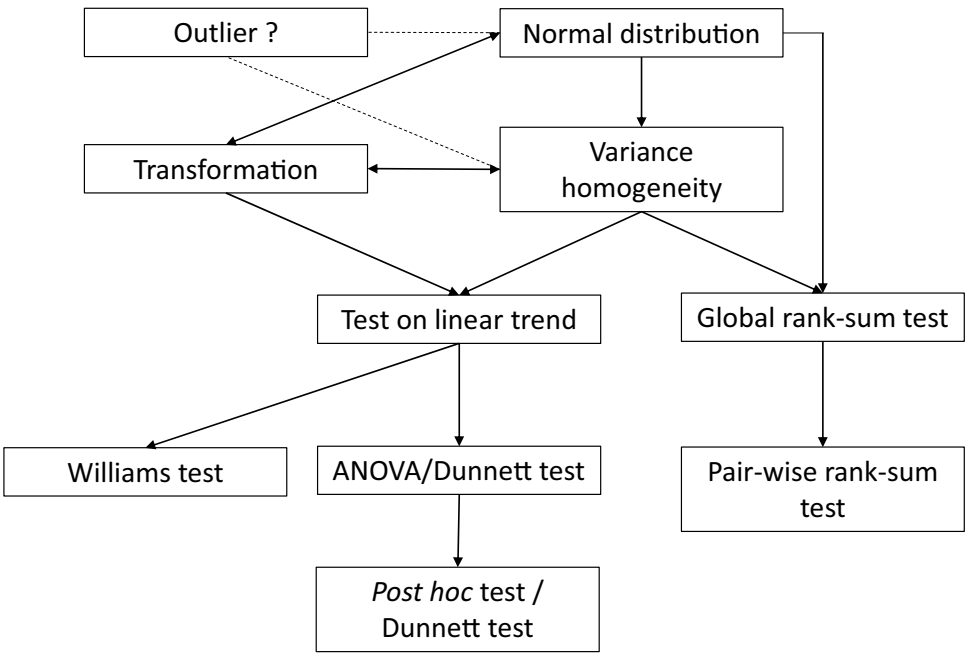

Fig. 1 Hypothetical statistical decision tree for continuous data. Shown are the various steps that might be conducted in an analysis. Sometimes a Williams-trend test is proposed instead of a Dunnett test (OECD 2014a) even if different hypotheses are tested (Jaki and Hothorn 2013; Williams 1971). Dunnett tests are also sometimes (incorrectly) used as post hoc tests. Some steps in an analysis may be repeated, e.g. transformation or outlier testing and may occur at different locations in a decision tree. Often a rank-sum test is suggested in case of assumed variance heterogeneity, which cannot be recommended (Zimmerman 1996, 2004)

the main test is usually not very high (or frankly: is inappropriately low) and the pre-tests' power, by design, even lower with respect to the main test (Kozak 2009). Hence, the pre-tests' results, and the decisions based on them, are unreliable, particularly when a combination of "deviations" occurs as often observed for biological data.

Pre-tests are performed to ensure that the main test's statistical assumptions are satisfied. The data under investigation come from artificially designed experiments and the inherent problem is that the pre-tests do not consider a potential treatment factor, i.e. the main test's usual null hypothesis 'no treatment effect' is also assumed to be true for the pre-tests. If the null hypothesis 'no treatment effect' is rejected due to the result of the main test, the robustness of the pre-tests' results need to be scrutinized a posteriori. If treatment really leads to a location change, e.g. mean increase in one group as compared to control, the distribution of the response variable (e.g. body weight), which is tested in pre-tests, actually consists of (at least) two distributions, namely the control and the treatment group distribution (refer to Fig. 2 for an illustration of the issue).

Error control and multiple testing

According to Neyman-Pearson/NHST test theory, all statistical tests have some probability of giving an incorrect result. On one hand, a result could be considered to be "statistically significant" (the null hypothesis is rejected) but it is actually not truly different to what it is compared. This is called a false positive or Type-I-Error. On the other hand, a result could be considered to be not statistically different (the null hypothesis is not rejected), while it truly is different to what it is compared. This is called a false negative or Type-II-Error. Both errors should be small but cannot be completely eliminated. The reason for this is that the assumed statistical distributions used in tests are continuous and we use (scientifically unjustified) thresholds to assign results as being improbable under the test assumptions. We a priori exclude certain extreme (test) results, which would occur only seldom under the assumed statistical distribution, from what we consider normal/to be no treatment effect.

An additional issue is that there is an inherent asymmetry between false positive and false negative decision rates in NHST. Only one of them, the false positive rate/Type-I-Error, can be controlled directly, whereas the false negative rate/Type-II-Error (1—Type-II-Error is the statistical power) is mainly driven by sample and expected effect size among other factors. For example, the null-hypothesis of the common Kolmogorov–Smirnov-test on normal distribution, investigates whether a homogeneous population of a certain size follows a normal distribution, which is compared against a wide alternative of non-normal distributions. That means that a p value > 0.05 (if selected as a statistical "relevance" threshold) favours normal distribution (the null hypnoses is not rejected in this so-called lack-of-fit test) and could lead to the subsequent use of a t test. However, for $n = 10$ this error control is extremely weak (Drezner et al. 2010). Moreover, this Kolmogorov–Smirnov-test is commonly used on k-sample design data and the violation of normal distribution may be due to a location effect only in the highest dose, even if the data follow an underlying normal distribution.

An additional issue is that the Type-I or alpha error is only controlled within one test and not over the range of tests

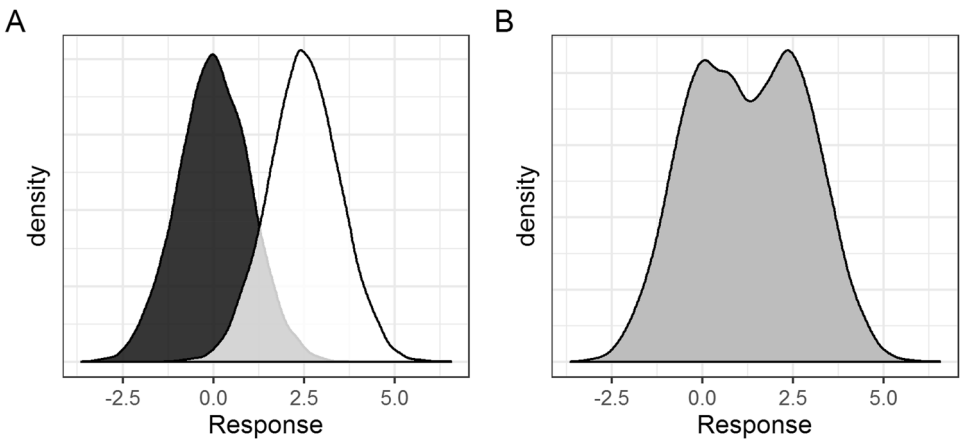

Fig. 2 The issue of not taking a treatment factor into account when using pre-tests. Shown are the density plots of a hypothetical response. **a** Shows that the response variable is actually affected by treatment (white) as compared to control (black)—the distributions are similar but shifted towards a higher response. However, a pre-test would be conducted on the joint response distribution (**b**), which does not take the treatment factor into account and is prone to bias the assessment. An alternative approach is to test each group independently on assumptions, however, this increases the total amount of statistical tests conducted, which would increase the probability for the Type-I-Error

subsequently performed in a decision tree. As one can easily calculate with the following formula, a Type-I-Error is only 5% when a single test is conducted.

$$\begin{aligned}&\text{Probability}\left[\text{at least one significant result}\right]\\&\quad=1-\text{Probability}\left[\text{no significant results}\right]\\&\quad=1-(1-0.05)^{Number\ of\ Hypotheses}\end{aligned}$$

Therefore, Type-I-Errors accumulate over the different levels of the decision tree. The use of pre-tests is therefore not recommended by several statisticians (Kozak 2009; Ramsey and Schafer 2002; Schucany and Tony Ng 2006; Zimmerman 2004).

This alpha error accumulation is of course even greater when one conducts an assumption test, e.g. for normality, on every treatment group, to prevent an assumption test null hypothesis rejection due to a potential (mean location shift) treatment effect, as highlighted above.

The overall alpha error increases for 2–5 subsequent but independent tests to 9.75–22.6%, assuming a nominal alpha of 5% for each test and that that alpha is controlled within each test—which is not always the case.

Type of tests used

The type of tests used relates to three issues within the decision tree test approach.

One, there are several different tests available to test the same assumption. For normality, for example, the Kolmogorov–Smirnov (K–S) test, Shapiro–Wilks (Wilk and Shapiro 1965) or Anderson–Darling test (Anderson and Darling 1952). The tests have different robustness against, e.g. extreme values. Hence, the variant of the test included in the decision tree affects the further progression of the subsequent decisions and potential main test.

Two, decision trees typically contain two types of tests, finite and asymptotic, at the same decision level, i.e. "main test". Asymptotic tests, such as Wilcoxon–Mann–Whitney (WMW) rank sum-tests (Mann and Whitney 1947), are appropriate as an approximation typically only for a large sample size ($n \to \infty$), and are therefore very different to finite tests such as *t* tests (Fisher 1925; Student 1908). An early decision will therefore not lead to similar alternatives for the main test but is a decision between finite and asymptotic, without transparency on the consequence for small sample sizes designs.

Three, decision trees also contain parametric and non-parametric tests as direct alternatives. For example in OECD 2010 "…if Levene test (a k-sample test on variance heterogeneity) is non-significant, use parametric Dunnett test, if not use the non-parametric Dunn-test (to stay in many-to-one comparisons)". However, most non-parametric tests are inappropriate for heterogeneous variances (Hothorn 2016b), they are useful for data that do not satisfy the normality assumption. The alternatives further result in different effect sizes, see below.

Effect size

The decision about which tests to use is also a decision about which effect size is calculated. While most weight is usually attributed to the *p* value, there is however a call to move to effect sizes (Cumming 2014) as they might be more appropriate to determine biological relevance and p-values have several undesired properties (Amrhein et al. 2017; Greenland et al. 2016; Wasserstein and Lazar 2016), also see discussion. The effect size of the *t* test is the difference of arithmetic means and in terms of location shift, which is easy to interpret. However, non-parametric test such as the Wilcoxon–Mann–Whitney-test gives the hard-to-interpret Hodges–Lehman estimator (when estimating confidence intervals), which is an estimate of the median. Even the *p* values are based on the different effect sizes used in the statistical tests. However, different effect sizes have a different robustness against the presence of extreme values or distributional assumptions (Hothorn 1989). This means that an effect size and confidence intervals estimated for one organ might not be comparable to an effect size of another organ, when different tests are used. Hence, only a common effect size within one bioassay allows a direct comparison.

Outliers

Some decision trees include an outlier detection step by for example Dixon's test (Dean and Dixon 1951). This introduces however at least two issues: (1) at which step should such a test be conducted (before the assumption tests or after) and (2) is the "outlier" actually an "extreme value".

Outliers come from a different sample process than the rest of the data by definition (Hawkins 1980). Thus, by a mistake such as a typo or occur due to methodological differences in the experimental conduct. An extreme value can occur in contrast due to the continuous statistical distribution assumption, as already described above. Further, also the underlying biology may give raise to extreme values, e.g. due to a particularly sensitive statistical unit; a test species might consist of responders and non-responders (compare individualized pharmacotherapy due to single-nucleotide polymorphisms or metabolomics). An extreme response value might thus be an indicator for a pathological response in safety assessment. However, an appropriate treatment of these extreme values is needed, e.g. by transformation of the response variable, otherwise

the false negative rate may increase when assuming normal distribution by the *t* test.

Therefore, single values should only be excluded from a statistical analysis upon a detailed assessment of the data and not due to a single statistical test.

Variance heterogeneity

Variance heterogeneity, i.e. unequal variance in the groups, is often observed in toxicological bioassays. In fact, one might argue that some level of variance heterogeneity, correlating with the effect and/or treatment level, should be expected after a toxic insult, e. g. due to different susceptibilities to the effect in the investigated population or due to interfering with some physiological process. Usually the variance heterogeneity is not further investigated, because the focus is on the potentially treatment-induced mean shift. The concern is mainly related to the derivation of unreliable confidence intervals and *p* values, i.e. false-positive (or negative) rates depending on the pattern of the variance over the groups and their sample sizes.

Similarly to the tests on normality, and the associated issues as discussed above, there are several different tests available on variance homogeneity. Most common are Levene's (Levene 1960) or Bartlett's test (Bartlett 1937), which have different robustness against extreme values or skewed distributions (Conover et al. 1981).

In case of variance heterogeneity, there are several adjustments available that make the pre-tests unnecessary, e. g. sandwich estimator (Herberich et al. 2010; Zeileis 2006) or reduced degrees of freedom (df) (Hasler and Hothorn 2008; Satterthwaite 1946; Welch 1947). Using such adjustments on the main tests result in robust estimation, in the case of heterogeneous variances [particularly when high variances occur in the group with small sample size (and vice versa)] and are still acceptable in the case of homogeneous variances. Therefore, these adjusted approaches can be recommended routinely where the df-reduction works better for small sample size designs (Hasler 2016).

Since both the standard Dunnett and Williams test are based on a common means square error estimator, individual inference can be biased, e.g. when a toxic response is associated with the smallest variance. Notice, as stated above, the usual main test "non-parametric" alternatives are often not robust in case of variance heterogeneity (Zimmerman 1996, 2004).

Figure 3 shows typically occurring variance heterogeneity, here increasing with dose, in simulated data. The figure also cautions to use visualization because "superficial" variance heterogeneity may occur due to vastly different reasons, e.g. subgroups or non-normal distributions.

Sample size

One of the most prominent but ignored issues in the statistical assessment of toxicology studies is that the sample size is often very limited. For 28 day (OECD 2008a, b, c) rodent repeated-dose toxicity studies only five animals per sex and group are required and for 90 days ten animals per sex and group (OECD 2018a). While carcinogenicity studies (OECD 2018b, c) require at least 50 animals per sex and group, the observed events are rare. Due to ethical considerations repeated-dose dog studies (OECD 1998) require only 4 animals/sex. Also in vivo genotoxicity studies require only five animals per test (OECD 2016a). From a statistical power perspective such observation numbers are challenging and many statistical assumption and main tests perform badly (Bland and Altman 2009).

On the other hand, when biological/toxicological relevance thresholds are not established and tested against zero control, also tiny/negligible effects can become statistically significant, because *p* values and confidence intervals directly depend on sample size. This can be demonstrated by, for example, the formula for the *t* statistic of the two-sample *t* test for balanced designs and equal variance (Fisher 1925; Student 1908), which tests if a sample mean ($\bar{m}_1$) is statistically equal to another sample mean ($\bar{m}_2$).

$$t = \frac{\bar{m}_1 - \bar{m}_2}{\mathrm{SD_p} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Here, $\mathrm{SD_p}$ is the pooled standard deviation of the sample [for calculation refer to Fisher (1925)] and n_1 and n_2 are the sample sizes for the respective mean values.

If there really is no difference between the means, the probability to observe increasing *t* values becomes smaller, viz the *p* value becomes smaller with higher *t* values (the *p* value should be calculated considering $n_1 + n_2 - 2$ degrees of freedom). It is obvious that *p* values also become smaller (and t increases), when the sample sizes increase (for all non-zero differences) because the denominator of the ratio decreases.

Alternative approaches

In the following, we present three approaches that allow the statistical assessment of data without using statistical decision trees: visualization and the use of regression diagnostics, robust testing, i.e. using tests with very relaxed assumptions, and the definition of "normal ranges".

Visualization

Generally, it can be recommended to plot data before any statistical testing is applied, i.e. to conduct exploratory data

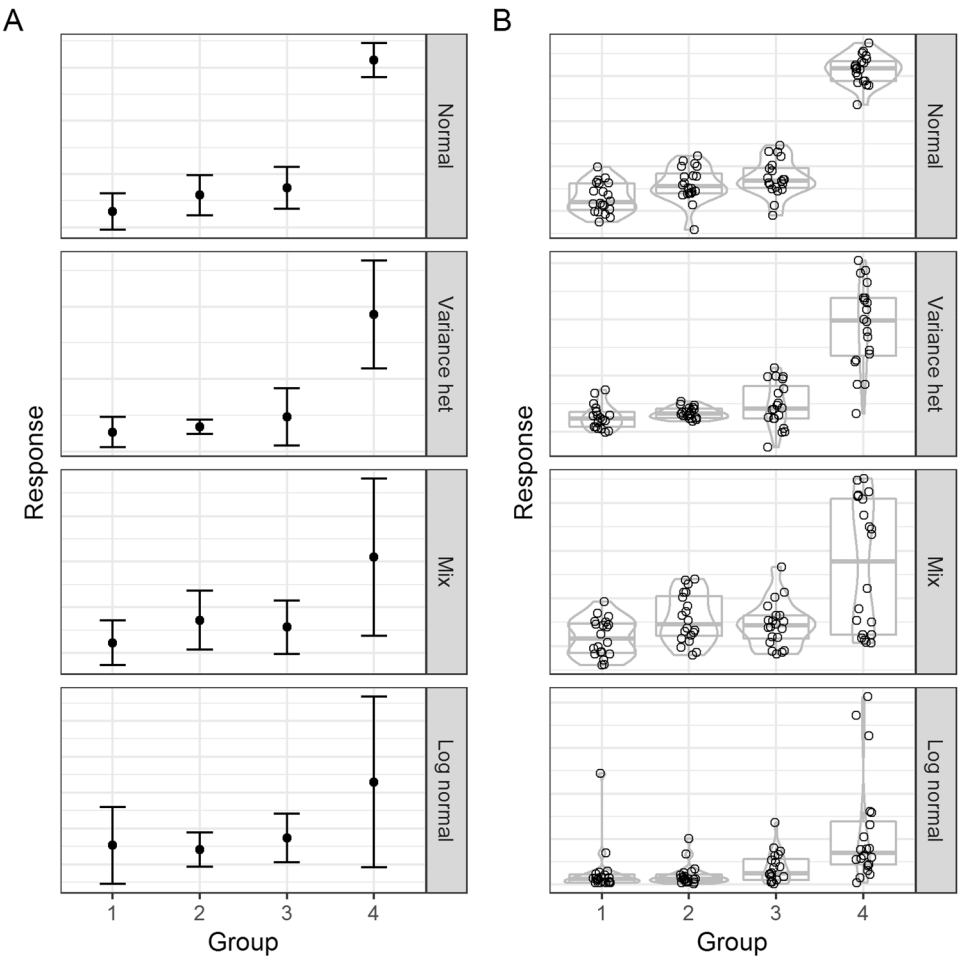

Fig. 3 The use or graphs for reviewing statistical test assumptions. Four data sets of four groups each were generated from different known distributions: normal distribution (Normal), normal distribution with increasing variance heterogeneity in the two higher dose groups (Variance het), a mixture of two normal distributions with variance homogeneity but with only 50% of the observations in the dose groups being susceptible of a treatment effect (Mix) and a log normal distributed response (Log normal). **a** Shows means and standard deviations. Except in the first plots there is an obvious variance heterogeneity between the groups for all plots except "Normal", which affect p-values and confidence intervals, i.e. decisions. **b** Shows individual values, box plots and violin plots, which can be used to assess the sample distribution (Wickham and Stryjewski 2011). For Group 4, the variance heterogeneity in "Mix" is different to be different to the one in "Variance het", which illustrates the use of plotting individual values. How to account for such a mixture of responders and non-responders in the statistical analysis may be a different discussion. "Log normal" obviously differs from the others plots in that we observe a large fraction of high values. This might be investigated further, see Fig. 4

analysis (Tukey 1977). This gives an indication of variation, grouping of individual responses, effect size and actual dose–response pattern. Such insight is important for a toxicological assessment but may be potentially obscured by single point value representation in result tables, common bar plots or statistical tests as represented by an asterisk in the data displays.

There have been many calls for the end of the typical mean bar graph and the plotting of individual values for data presentation (Cleveland 1993; Fosang and Colbran 2015; Nature Methods Editorial 2014; Pallmann and Hothorn 2016; Weissgerber et al. 2019). Similarly, the dangers of incorrect assessments by conducting a statistical analyses without plotting the data have been repeatedly voiced (Anscombe 1973; Kluxen 2019b; Matejka and Fitzmaurice 2017; Tukey 1977). However, in regulatory bioassays, there are usually few if any graphs. And one can assume that the reasoning behind the statistical analysis relies exclusively on assumption tests (which is accordingly also detailed in the methods sections of the respective study reports).

Variance heterogeneity and trend can be easily assessed by plotting means and standard deviations. Additionally plotting individual values reveals to some extent a deviation from the normal assumption and potential grouping of values. This is demonstrated in Fig. 3, which shows randomly generated example data with different properties.

Most notable is the variance heterogeneity in three of the four data sets, however, with different aetiologies that cannot be detected in detail by applying the usual assumption tests: mixture distribution and log-normality.

ANOVA-type tests are regression analyses where the predictor is not continuous but the group variable. In a regression analysis, a line is fit through the data, reduces the amount of error (smallest residuals) as compared to all other possible lines. For standard ANOVA-type tests, the best "line" is through the group means, because of its properties. The response of any value in the treatment part of the dataset can be described as the mean response of the control group plus/minus the mean response of the group the value is in; we assume that the difference of treatment group response and control response can be solely explained by a location shift, compare Fig. 2a. This is associated with some error; and in a standard ANOVA or Dunnett test, the error is assumed to be similar for all groups, which is the variance homogeneity assumption. We can also say that a random data point in a group is best predicted by its group mean and some uncertainty.

Hence, when a linear model is fitted to the data, the model can be investigated by regression diagnostics (Kozak and Piepho 2018), e.g. the distribution of residuals investigated for deviations from the model assumptions or extreme values. An example of one kind of such a diagnostic plot, a Quantile–Quantile (QQ) plot, is given in Fig. 4, which explores the normality assumption of two data sets already shown in Fig. 3. Here, the quantiles of two distributions are plotted against each other, the residuals of the model fit and the theoretical quantiles of a normal distribution. If the distributions are similar, the data will follow a line and we can assume that the residuals are normal distributed, which is one of the ANOVA-type test assumptions. Further, the leverage and influence of single values, i.e. values that tremendously affect model fit, can be explored by various methods (Cook 1977). If the "statistical significance" relies on a single value in a data set, the actual significance of this observation can be appropriately discussed within the regression analysis framework. An illustrating example for using "Cook's distance" is shown in Fig. 5, where the statistical significance is driven by two values. Such plots allow the identification of outliers, i.e. values that may be errors in the data record or observations subjected to methodical differences during the experiment as compared to the other observations.

Using residual plots instead of using statistical tests has several benefits (Kozak and Piepho 2018). We do not have to rely on the data-dependent performance of assumption test variety used in the decision tree and we do not have to rely on a strict binary rejection criterion but can assess for approximate compliance or clear deviation. Most importantly, the treatment factor is taken into account when investigating residuals from a model fit. If a graphical assessment appears "too subjective", the reader is encouraged to read-up the history behind the common 5% alpha threshold (e.g. in Salsburg 2002). Further, methods were developed that allow a more objective assessment by graphical model validation techniques (Ekstrøm 2014).

Such a detailed analysis of model fit may require a level of statistical literacy (or knowledge of the appropriate software) that cannot be necessarily expected for all researchers with their expertise concentrated in other fields. Hence, a graphical regression analysis may be suitable to investigate single responses (that are considered to be potentially most adverse) but may be inconvenient as a generic approach in

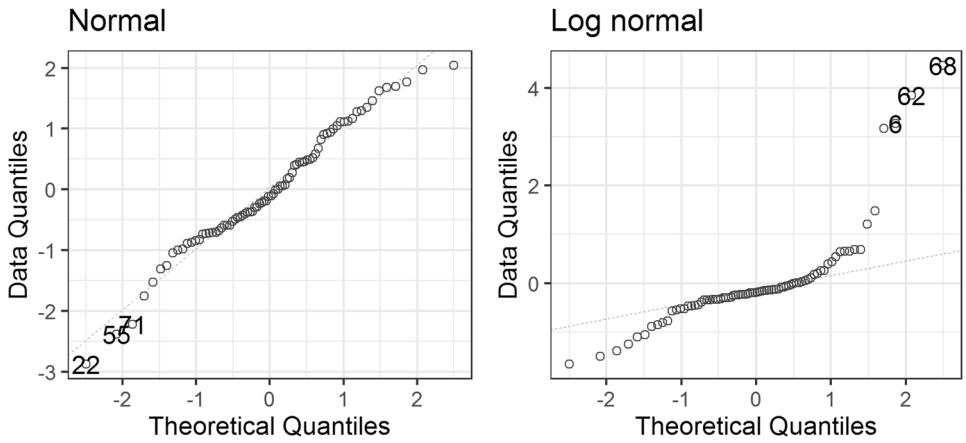

Fig. 4 The use of diagnostic plots for reviewing statistical test assumptions. For linear models, such as ANOVA/Dunnett-type tests, the distribution of the residuals from the model fit should be approximately normal (here, the difference of the individual observations from the respective group means). This assumption can be tested by Quantile–Quantile (Q–Q) plots, where the quantiles of one distribution are plotted against another, here normal distribution. When the distributions are similar, the individual values follow a line. It is obvious for the "Log normal" data deviate from this assumption. Please refer to (Kozak and Piepho 2018)

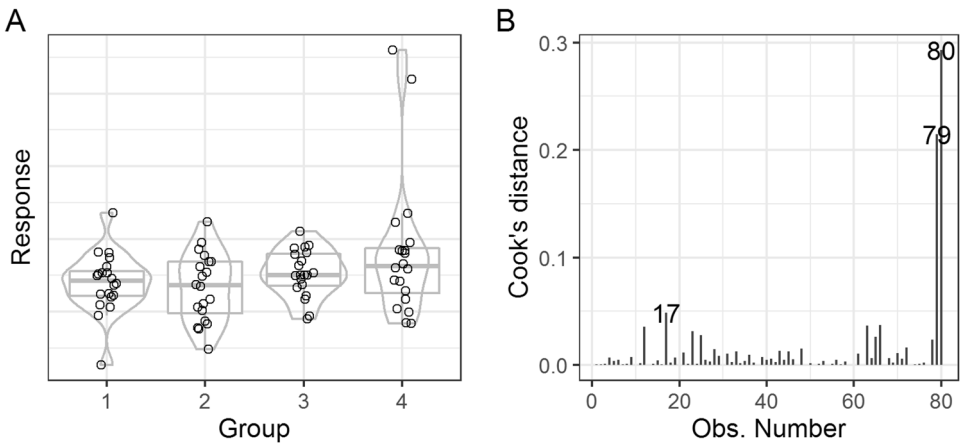

Fig. 5 The use of diagnostic plots for reviewing statistical test assumptions. **a** Shows a data set where the samples for Group 3 and 4 come from normal distributions with slightly increased means (0.5 units higher, all distributions have a standard distribution of 1) than the distributions of Group 1 and 2. Two extreme values where manually introduced in Group 4, which results in a statistical significant result of a fitted ANOVA model. **b** Plotting "Cook's distance" (Cook 1977) of the model fit reveals two values in the data set that drive the statistical significance of the ANOVA, values with the indices 79 and 80 in the data set (value 17 is also highlighted but has lesser influence/potency). Exclusion of the points from the statistical analysis result in no "statistical significance". In an actual experiment it may be questioned whether the observations are outliers, e.g. mistakes occurring in the data recording or observations with methodical deviations in the experimental conduct

bioassays with many endpoints. Graphical data presentation can however be realized for multiple endpoints as, for example, demonstrated for pair-wise standardized effect size plots (Schmidt et al. 2016). Those could be expanded by including multiple treatment groups in one plot.

Robust testing

Robust testing comprises several tests that perform well when deviations from the usual assumptions or extreme values occur (Wilcox 2012). If we want to estimate a typical data value in a data set, we can use point estimates. The data set [1, 1, 1, 1, 2] has a mean of 1.2 and a median of 1. Either value could be considered to be a good representation of a typical value. The data set [1, 1, 1, 1, 20] has a mean of 4.8, which is higher than all other values in the data set except the maximum, but still hast a median of 1. Thus, a median is can be considered to be more robust in case of extreme values in a data set than the mean. Similarly, statistical tests can be robust, i.e. less sensitive against deviations from their assumptions.

In case of a very skewed distribution and/or variance changes (variance heterogeneity) with dose level, an ANOVA gives incorrect p-values and confidence intervals because its assumptions are violated. That does not mean that an ANOVA cannot be conducted, it means that the results need to be questioned with regard to the deviations from its assumptions.

Therefore, if multiple comparisons for pair-wise testing are performed, it is useful to generically apply a test that is robust regarding deviations from the usual statistical tests' assumptions, so-called robust tests. Several of such tests have been proposed in the past. As indicated above, a Dunnett test can be adjusted for variance heterogeneity by sandwich estimator (Herberich and Hothorn 2012; Zeileis 2006) or by reduced degrees of freedom, Satterthwaite approximation (Hasler 2016; Satterthwaite 1946), or using robust linear models based on various methods to minimize the M-estimator (Koller and Stahel 2011). "Non-parametric" rank-sum tests, such as the Steel test (Steel 1959) or modifications (Konietschke et al. 2015), can be used without distributional assumptions but still require variance homogeneity and large sample sizes.

We recently compared (Hothorn and Kluxen 2019)[1] several modifications of the Dunnett procedure against the original test (Dunnett 1955) and against each other. In the same paper we presented new modifications of the Dunnett procedure for continuous data applying the novel most likely transformation (MLT) approach (Hothorn et al. 2018), the MLT-Dunnett and the modification for correlated multiple and differently-scaled endpoints, optimally dichotomized for continuous endpoints by continuous outcome logistic regression (COLR), the COLR-Dunnett.

Unfortunately, the MLT approach is mathematically complex, however, it can be easily applied with the available

[1] N.B. after peer-review another study investigated the robustness of the MLT-Dunnett for count data which is available as a pre-print: Hothorn and Kluxen (2020) Statistical analysis of no observed effect concentrations or levels in eco-toxicological assays with overdispersed count endpoints. https://doi.org/10.1101/2020.01.15.907881.

software (Hothorn 2018). It can also be seen as an extension of Box–Cox transformations (Box and Cox 1964), which is a family of power transformations (commonly log or square root, square, cubic etc. and their reciprocal values) that can be used to (approximately) normalize data. The MLT approach achieves the same but applies a cascade of increasingly complex transformation models and allows to choose the most appropriate one using a maximum likelihood framework. In this, the parameterization of the monotone increasing transformation function is achieved by Bernstein polynomials that can be used to approximate any continuous distribution (Farouki 2012). Where the usual regression models estimate the conditional mean as a function of dose and assume that distributional properties such as variance or skewness of the distribution can be ignored, they are considered in the MLT approach. Therefore, it is robust against any non-normal distributions (including discreteness), variance heterogeneity, extreme values, and even censored observations (values with a detection limit).

Continuous outcome logistic regression (Lohse et al. 2017) is similar to the most likely transformation approach but with the distribution function argument "logistic" for data-optimal dichotomization. It provides dimensionless odds ratios and their confidence intervals as effect size, which is optimally dichotomized for the endpoint specific distribution.

Based on a simulation study (Hothorn and Kluxen 2019)[2], the MLT-Dunnett can be recommended as being almost always appropriate for continuous data with having greater power and a smaller Type-I-Error than its other robust versions. If one prefers confidence intervals instead of *p* values, which is recommended (Amrhein et al. 2019; Wasserstein et al. 2019), which allows the interpretation of effect size and its uncertainty in toxicological units, this is considerably more difficult for the MLT-Dunnett test. The reason for this is that a retransformation into the measured scale is not available. A way out then is the odds ratio as effect size for COLR-Dunnett test, which is however not very common in toxicology. The COLR-Dunnett allows the simultaneous comparison of multiple differently scaled end-points within one bioassay with odds-ratio as the common effect size. (Hothorn and Kluxen 2019) This means that effects within one data set can be compared on a common scale and not by different effect sizes as it may be the case when following decision tree approaches.

However, all multiple comparison tests used in toxicology and also robust methods such as the MLT-Dunnett, do not perform optimally in small sample size designs (also see discussion).

Nevertheless, we believe that robust testing can replace decision tree-dependent testing in a toxicological bioassay as the standard approach.

Reference ranges

A different way to assess toxicologically concerning responses in regulatory assays is to define reference ranges of responses that are considered "normal", as commonly done for clinical haematological assessments for humans. Unfortunately, there are still many open questions before this approach can be generically implemented.

There are different ways to define reference ranges: they can be based on the concurrent assay or based on historical data or a combination of the two with the concurrent control being the most recent historical control data. Further, they should consider biological variance and random error.

The benefit for using historical values is clearly that, if the historical control data are from the same population, this will result in a better estimate of the population standard deviation because of an increased n and because the response of the concurrent control could be extreme. An issue is that due to genetic drift and potential background infections, time-dependent variation of some parameters might occur. This requires continuous monitoring and attention from the breeders or radically different factorial approaches with associated statistical challenges (Festing 1993). While historical control data needs to be submitted for human safety assessment, e.g. according to the data requirements in pesticide regulation (European Commission 2013), there is no agreed standard on form or detail. It is also not mandatory to provide control charts. Similarly, there is no agreed numerical range derivation for historical control data. In ecotoxicology, a more involved discussion about how to use historical control data only just started (Wheeler 2019).

There are at least three methods to defined historical control ranges: some OECD test guidelines require the derivation of confidence intervals (normal or Poisson-based depending on the response) for historical control means. Contrary to this, Igl et al. (2019) proposes to use tolerance intervals. A confidence interval contains the population mean of a concurrent population with a pre-specified confidence probability. A tolerance interval contains a specified proportion of future samples (with a pre-specified confidence probability), where the number of future samples is not specified. A tolerance interval may consider that 90% of future observations will fall into the interval with a 95% confidence. A prediction interval contains a single future observation within n future samples, with a pre-specified confidence probability (Hahn and Meeker 1991). Thus a confidence interval considers the population mean but both

[2] Hothorn and Kluxen (2020) has become available in the meantime as a preprint, which investigates the MLT-Dunnett's robustness for count data.

tolerance and prediction intervals future observations. This is the key distinction and indicates their value when using historical control data.

While it seems to be clear that the confidence interval is inappropriate, tolerance intervals suffer from a suitable choice of the proportion (90% or 50%?) in future samples. Hence, we propose prediction intervals for toxicological reasoning: a certain interval from historical controls of size n is estimated in such a way that a single future value is within the interval. If so the effect can be assessed as incidental, otherwise as treatment-related. Instead of a prediction interval for a future single value, a prediction interval for a future mean (of a certain observation number) can be estimated alternatively. Note all intervals depend on observation number, which seems however to be non-critical for regulatory studies with recommended/required sample sizes.

Figure 6 shows a hypothetical example of using reference ranges. Two of the shown group responses are within the reference range, which can be defined in various ways as described above. Two of the group responses are outside of the range and can be discussed to be of toxicological relevance. The uncertainty associated with their effect size can be described by confidence intervals, which can be interpreted as "compatibility intervals" (Greenland 2019). An application of such interpretation in toxicology is discussed elsewhere (Hothorn and Pirow 2019; Kluxen 2019a).

Case study

As a simple case study, the clinical chemistry endpoint creatine kinase was chosen from a 13-week study with sodium dichromate dihydrate administered to F344 rats (National Toxicology Program 2010). The data are plotted in Fig. 7.

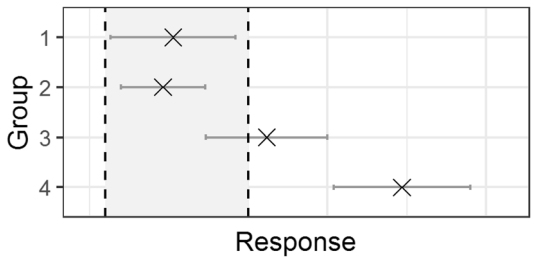

Fig. 6 The use of reference ranges to assess toxicological relevance. The graph shows the responses of a hypothetical data set. The axes of this graph have been flipped to distinguish it from Fig. 3a. The mean responses (×, multiplication sign) of group 1 and 2 are within the range of acceptable variation which can be estimated by different methods as discussed in "Reference Ranges". The responses of groups 3 and 4 are outside of the reference range. The mean responses can also be interpreted with regard to uncertainty by using confidence intervals

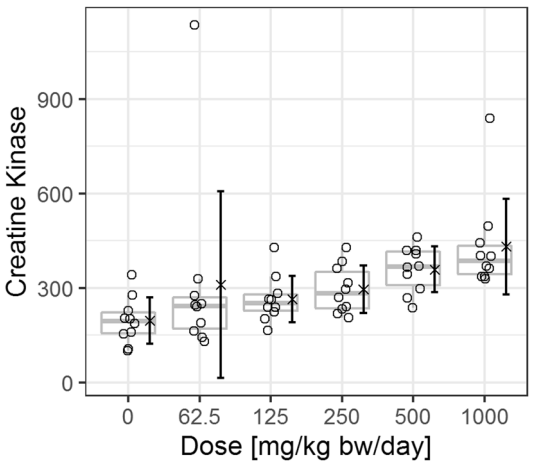

Fig. 7 Case study: Creatine Kinase induction in a 13-week study with sodium dichromate dihydrate administered to F344 rats (National Toxicology Program 2010). Plotted are individual values, box plots in grey and mean (×, multiplication sign) and standard deviations. The graphs show two high values that appear to introduce variance heterogeneity

The plot shows a dose response but due to the presence of two high values pronounced variance heterogeneity.

Table 1 shows the results of applying 3 different assumption tests for both the normality and the variance homogeneity assumption. While there is an unequivocal result for the non-normality decision, if the common $p < 0.05$ decision criterion is applied, the p-values differ tremendously. A similar observation can be made for the assumption tests on variance homogeneity. Here, it is quite obvious that the selected test in the decision tree could affect the further statistical analysis.

If we consider the Lilliefors-version of the Kolmogorov test for small observation numbers (Dallal and Wilkinson 1986), the p value of 0.003 speaks against the assumption of normal distribution (Note that the group structure of the design was not taken into account). According to the decision tree in Fig. 1, this could then lead to a global rank test. Since the p value of the corresponding Kruskal–Wallis test is very small ($p = 0.00005$), Wilcoxon tests in pairs for comparison against control would follow. Its Bonferroni-adjusted (Shaffer 1995) two-sided p values and the corresponding tests after simply using a MLT-Dunnett are presented in Table 2.

For the Wilcoxon test, the no observed effect dose is 125 mg/kg bw/day. The robust alternative without any pre-tests, the MLT-Dunnett yields monotonously smaller adjusted p-values and a smaller no observed effect dose, when a p value of < 0.05 is considered appropriate (and biological relevance is not taken into account). The robust approach does not require a decision tree and achieves qualitative other, more sensitive significance decisions. The

Table 1 Assumption tests conducted on the case study data result in vastly different *p*-values

Assumption tested	Assumption tests with corresponding *p* values		
Normality	Shapiro–Wilk	Anderson–Darling	Lilliefors (Kolmogorov–Smirnov)
	1.121865e−08	5.4083e-07	0.002935813
Variance homogeneity	Levene	Bartlett	Fligner–Killeen
	0.1665637	1.004338e−06	0.9712356

This suggests that when different decision trees are used on the same data the choice of the assumption test present in the decision tree can affect in different main tests and subsequently the results

Table 2 One-sided *p* values of three different statistical evaluations of the case study

Dose group comparison against control (0)	Wilcoxon test (Bonferroni-adjusted)	MLT-Dunnett	Standard Dunnett
62.5–0	0.79	0.13	0.15
125–0	0.11	0.04	0.41
250–0	0.013	0.007	0.22
500–0	0.0005	< 0.0001	0.04
1000–0	0.0002	< 0.0001	0.002

p-values of the standard Dunnett tests are also shown for comparison and illustrate the effect of ignoring variance heterogeneity. Note only one-sided p-values were calculated under the assumption that only increasing Creatine Kinase would be toxicologically relevant. The R-code for this example is available in the supplementary material.

Discussion

The two currently used methods for hazard characterization of chemicals are either to identify a treatment level that is statistically and ideally biologically different from control or deriving a dose that elicits a response equal to or exceeding a pre-defined threshold. The prior establishes a No Observed Adverse Effect Level (NOAEL, or NOAEC for concentration) with the aid of Dunnett or pair-wise statistical testing and the latter a benchmark dose level (BMD) or an effective concentration (EC), by statistical dose–response modelling.

For BMD derivation, the state-of-the-art is model averaging (Hardy et al. 2017; Jensen et al. 2019; Wheeler and Bailer 2007), because it reduces the impact of selecting an incorrect model and it performs better when the biological response cannot be derived due to the common limited number of dose groups. Model-averaging ameliorates the larger error of choosing an incorrect dose–response model. Conversely, the decision tree approach does not offer a similar protection and may due to the insufficiency of the pre-tests result in sub-optimal analysis with different tests for different endpoints, which makes their results incomparable. Using decision trees also increases the overall false positive rate, which is not desirable.

So why use testing against control at all to derive NOAELs if the generic decision tree method seems inadequate and dose–response modelling is available? On the one hand, pair-wise or Dunnett testing is a current regulatory requirement, e.g. for hazard identification. On the other hand, it is possible to derive NOAELs by using assumption-free/robust tests without relying on statistical decision trees.

In principle, pair-wise or Dunnett testing perform well when the tests' assumptions are approximately satisfied and many tests have a certain robustness against slight deviations from their assumptions, when the group observations are balanced. Hothorn (2014) previously proposed to use a minimal decision tree based on certain data properties, due to the deficiencies with the decision tree approach. However, it is considered unlikely that a refined statistical assessment of data and model properties, such as regression diagnostics, can be performed on every endpoint investigated within one study under realistic conditions. Further, it may be unreasonable to assume that all (eco-) toxicologists have capacity to acquire the level of statistical literacy or practically conduct the refined assessments.

(Eco-) toxicologists need to identify parameters in a bioassay that may raise concern, and this is often out of a very high number of parameters that are investigated within one study, e.g. clinical chemistry. Here a statistically significant result often acts as a first clue, and triggers further investigation, i.e. whether other related parameters, such as other clinical chemistry changes, organ weights or histopathological observations, are dose-dependently affected.

Thus, some generic approach is needed and we believe that robust testing can be used as a better and simpler alternative to decision trees.

Generically, robust testing seems to be most appropriate and easily applicable for pair-wise testing and is a suitable replacement of the commonly used decision trees, because the generic application of decision trees may result in sub-optimal test choices. An a priori-defined and options-restricted decision tree is counter-productive, because it may result in a flawed assessment based on a single incorrect decision within a batch of several subsequent binary decisions. Subsequent application of dependent statistical tests propagate error. It seems more appropriate to use a test that performs best in most situations, such as the MLT-Dunnett for continuous data.

We also briefly introduced the concept of using reference ranges. There are currently however still some unresolved issues with the approach. For example, it needs to be established whether the group mean, or an individual value within a group or the difference to control would be relevant and whether between assay variability should be considered. Currently and usually the between assay variability is not considered. Historical assays can be modelled as a random factor to estimate prediction intervals for random effects models (Hoffman and Berger 2011).

Decision trees have the superficial merit of describing a statistical analysis in well-documented GLP-style. However, using e.g. the platform-independent, free-statistics software "R" (R Core Team 2017) allows documentation of the statistical analysis by design because the analysis is conducted by running software code. The code can be easily reported, quality checked and peer-reviewed; i.e. the data analysis satisfies the "reproducible research" concept (Gandrud 2015). Further, all presented robust methods are available in R, for example code refer to Hothorn and Kluxen (2019).

The authors would like to caution the use of statistical tests, also the use of robust tests, in isolation to identify a supposedly relevant (eco-) toxicological response. Statistical significance as the sole relevance criterion has been repeatedly criticized (Amrhein et al. 2017; Nuzzo 2014; Wasserstein and Lazar 2016) and recently a complete special issue of *The American Statistician* was dedicated to the problem (Wasserstein et al. 2019). All statistical tests depend on some fundamental assumptions, i.e. what is an appropriate threshold for statistical significance/an appropriate and acceptable error rate (which is usually set arbitrarily at 5%; as discussed in Salsburg (2002)). If the results from a statistical analysis clearly deviate from the general pattern observed by exploratory data analysis, there might be an issue with the applied inference model. If there is no consistent and repeatable response, also if there is a single isolated statistical response within a set of related parameters as found in clinical chemistry, the chances for a Type-II error are slight ("false negative"). This is consistent with the toxicological method of establishing dose–response being clearly different from historical control variation. A repetition of the assay or a weight-of-evidence assessment also reduces the Type-I error rate ("false positive"). A statistically significant result should always be submitted to a "plausibility check". If the result of a statistical analysis seems implausible with regard to the general pattern in the data (also compare the "common sense test", Fox and Landis (2016)), independent of whether or not statistical significance was established, either more data needs to be produced or other data/information are needed for a weight-of-evidence assessment. In case of vertebrate testing, the generation of new data might, however, be restricted due to ethical considerations.

It should be further considered that the current approach of deriving NOAELs in safety assessment, i.e. the proof-of-hazard assessment, can be scrutinized per se due to philosophical considerations: "absence of evidence is not evidence of absence" (Altman and Bland 1995). A proof-of-safety assessment would be more appropriate, as conducted in the testing for bioequivalence (U.S. Food and Drug Administration 2001), however, there is a lack of consensus on biological relevance thresholds, for which reference ranges could be helpful. Software for this is available (Dilba et al. 2004; Hothorn and Hasler 2008; Hothorn et al. 2008) and the application could be easily implemented (Delignette-Muller et al. 2011), also by using robust tests.

Only some of the endpoints in toxicology are continuous (and considered in most decision tree approaches), others are discrete (proportional, scores), ordered, counts or time-to-event (Szocs and Schafer 2015). For these endpoints and their appropriate effect sizes related Dunnett-type procedures are available, e.g. for ratio-to-control (Dilba et al. 2004), proportions (Schaarschmidt et al. 2009), relative effect size (Konietschke et al. 2015), poly-k estimates (Schaarschmidt et al. 2008) and time-to-event data (Herberich and Hothorn 2012).

Conclusion

The basic principle of toxicological testing, is to experimentally establish a dose–response and to identify a dose with a biologically relevant and probably non-random deviation from "normal". Statistical toxicology should be aiding and mirroring this. A binary yes-and-no answer, as propagated by the decision tree idea, is conceptually and statistically flawed, particularly in the commonly-used designs with small sample sizes.

While the decision tree approach has substantial deficiencies, it has the superficial merit of being easily plannable

and documentable and thus apparently fits well into a GLP environment. There are however tools available that make pre-testing unnecessary. Since manual model fitting/statistical analysis cannot be expected on a routine basis, robust testing can be recommended as a valid and better alternative to decision trees.

We refer to the generally better performance of the MLT-Dunnett for multiple comparisons as compared to current standard methodology (Hothorn and Kluxen 2019), which can be used to generically replace decision trees. However, using the robust Dunnett test instead of decision trees does not prevent assessment fallacies: also such tests are not able to identify (eco-) toxicological relevance in the current form and as a stand-alone statistical application.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Altman DG, Bland JM (1995) Statistics notes: absence of evidence is not evidence of absence. BMJ 311(7003):485. https://doi.org/10.1136/bmj.311.7003.485

Amrhein V, Korner-Nievergelt F, Roth T (2017) The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ 5:e3544. https://doi.org/10.7717/peerj.3544

Amrhein V, Greenland S, McShane B (2019) Retire statistical significance. Nature 567:305–307. https://doi.org/10.1038/d41586-019-00857-9

Anderson TW, Darling DA (1952) Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Ann Math Statist 23(2):193–212. https://doi.org/10.1214/aoms/1177729437

Anscombe FJ (1973) Graphs in statistical analysis. Am Stat 27(1):17–21. https://doi.org/10.1080/00031305.1973.10478966

Bartlett MS (1937) Properties of sufficiency and statistical tests. Proc R Soc Lond A Math Phys Sci 160(901):268–282. https://doi.org/10.1098/rspa.1937.0109

Bland JM, Altman DG (2009) Analysis of continuous data from small samples. BMJ 338:a3166. https://doi.org/10.1136/bmj.a3166

Box G, Cox D (1964) An analysis of transformations. Proc R Soc Lond A Math Phys Sci 26:211–252

Cleveland WS (1993) Visualizing data. At & T Bell Laboratories, Murray Hill

Conover WJ, Johnson ME, Johnson MM (1981) A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics 23(4):351–361. https://doi.org/10.2307/1268225

Cook RD (1977) Detection of influential observation in linear regression. Technometrics 19(1):15–18. https://doi.org/10.2307/1268249

Cumming G (2014) The new statistics: why and how. Psychol Sci 25(1):7–29. https://doi.org/10.1177/0956797613504966

Dallal GE, Wilkinson L (1986) An analytic approximation to the distribution of Lilliefors's test statistic for normality. Am Stat 40(4):294–296. https://doi.org/10.1080/00031305.1986.10475419

Dean RB, Dixon WJ (1951) Simplified statistics for small numbers of observations. Anal Chem 23(4):636–638. https://doi.org/10.1021/ac60052a025

Delignette-Muller ML, Forfait C, Billoir E, Charles S (2011) A new perspective on the Dunnett procedure: filling the gap between NOEC/LOEC and ECx concepts. Environ Toxicol Chem 30(12):2888–2891. https://doi.org/10.1002/etc.686

Dilba G, Bretz F, Guiard V, Hothorn LA (2004) Simultaneous confidence intervals for ratios with applications to the comparison of several treatments with a control. Methods Inf Med 43(5):465–469

Drezner Z, Turel O, Zerom D (2010) A modified Kolmogorov–Smirnov test for normality. Commun Stat Simul Comput 39(4):693–704. https://doi.org/10.1080/03610911003615816

Dunn OJ (1961) Multiple comparisons among means. J Am Stat Assoc 56:52–64. https://doi.org/10.1080/01621459.1961.10482090

Dunnett CW (1955) A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc 50(272):1096–1121. https://doi.org/10.2307/2281208

Ekstrøm CT (2014) Teaching 'instant experience' with graphical model validation techniques. Teach Stat 36(1):23–26. https://doi.org/10.1111/test.12027

European Commission (2013) COMMISSION REGULATION (EU) No 283/2013 of 1 March 2013 setting out the data requirements for active substances, in accordance with Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of plant protection products on the market. OJ L 93/1

Farouki RT (2012) The Bernstein polynomial basis: a centennial retrospective. Comput Aided Geom Des 29(6):379–419. https://doi.org/10.1016/j.cagd.2012.03.001

Festing M (1993) Genetic variation in outbred rats and mice and its implications for toxicological screening. J Exp Anim Sci 35(5–6):210–220

Fisher RA (1925) Statistical methods for research workers. Oliver & Boyd, Edinburgh

Fosang AJ, Colbran RJ (2015) Transparency is the key to quality. J Biol Chem 290(50):29692–29694. https://doi.org/10.1074/jbc.E115.000002

Fox DR, Landis WG (2016) Don't be fooled—a no-observed-effect concentration is no substitute for a poor concentration–response experiment. Environ Toxicol Chem 35(9):2141–2148. https://doi.org/10.1002/etc.3459

Gandrud C (2015) Reproducible research with R and R studio. Chapman and Hall/CRC, New York

Greenland S (2019) Valid P-values behave exactly as they should: some misleading criticisms of P-values and their resolution with S-values. Am Stat 73(sup1):106–114. https://doi.org/10.1080/00031305.2018.1529625

Greenland S, Senn SJ, Rothman KJ et al (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 31(4):337–350. https://doi.org/10.1007/s10654-016-0149-3

Hahn GJ, Meeker WQ (1991) Statistical intervals—a guide for practitioners. Wiley, New York

Hamada C (2018) Statistical analysis for toxicity studies. J Toxicol Pathol 31(1):15–22. https://doi.org/10.1293/tox.2017-0050

Hardy A, Benford D, Halldorsson T et al (2017) Update: use of the benchmark dose approach in risk assessment. EFSA J 15(1):e04658. https://doi.org/10.2903/j.efsa.2017.4658

Hasler M (2016) Heteroscedasticity: multiple degrees of freedom vs sandwich estimation. Stat Pap 57(1):55–68. https://doi.org/10.1007/s00362-014-0640-4

Hasler M, Hothorn LA (2008) Multiple contrast tests in the presence of heteroscedasticity. Biom J 50(5):793–800. https://doi.org/10.1002/bimj.200710466

Hawkins DM (1980) Identification of outliers. Chapman and Hall, New York

Herberich E, Hothorn LA (2012) Statistical evaluation of mortality in long-term carcinogenicity bioassays using a Williams-type procedure. Regul Toxicol Pharmacol 64(1):26–34. https://doi.org/10.1016/j.yrtph.2012.06.014

Herberich E, Sikorski J, Hothorn T (2010) A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. PLoS ONE 5(3):e9788. https://doi.org/10.1371/journal.pone.0009788

Hoffman D, Berger M (2011) Statistical considerations for calculation of immunogenicity screening assay cut points. J Immunol Methods 373(1–2):200–208. https://doi.org/10.1016/j.jim.2011.08.019

Hothorn L (1989) Robustness study on Williams- and Shirley-procedure, with application in toxicology. Biom J 31(8):891–903. https://doi.org/10.1002/bimj.4710310802

Hothorn LA (2014) Statistical evaluation of toxicological bioassays—a review. Toxicol Res 3(6):418–432. https://doi.org/10.1039/c4tx00047a

Hothorn LA (2016a) Statistics in toxicology using R. CRC Press, Boca Raton

Hothorn LA (2016b) The two-step approach—a significant ANOVA F-test before Dunnett's comparisons against a control—is not recommended. Commun Stat Theory Methods 45(11):3332–3343. https://doi.org/10.1080/03610926.2014.902225

Hothorn T (2018) Most likely transformations: the mlt package. J Stat Softw

Hothorn LA, Hasler M (2008) Proof of hazard and proof of safety in toxicological studies using simultaneous confidence intervals for differences and ratios to control. J Biopharm Stat 18(5):915–933. https://doi.org/10.1080/10543400802287511

Hothorn LA, Kluxen FM (2019) Robust multiple comparisons against a control group with application in toxicology arXiv.

Hothorn LA, Pirow R (2019) Use compatibility intervals in regulatory toxicology [submitted to Regulatory Toxicology and Pharmacology].

Hothorn T, Bretz F, Westfall P (2008) Simultaneous inference in general parametric models. Biom J 50(3):346–363. https://doi.org/10.1002/bimj.200810425

Hothorn T, Möst L, Bühlmann P (2018) Most Likely Transformations. Scand J. Stat 45(1):110–134. https://doi.org/10.1111/sjos.12291

Igl B-W, Bitsch A, Bringezu F et al (2019) The rat bone marrow micronucleus test: statistical considerations on historical negative control data. Regul Toxicol Pharmacol 102:13–22. https://doi.org/10.1016/j.yrtph.2018.12.009

Jaki T, Hothorn LA (2013) Statistical evaluation of toxicological assays: Dunnett or Williams test-take both. Arch Toxicol 87(11):1901–1910. https://doi.org/10.1007/s00204-013-1065-x

Jarvis P, Saul J, Aylott M, Bate S, Geys H, Sherington J (2011) An assessment of the statistical methods used to analyse toxicology studies. Pharm Stat 10(6):477–484. https://doi.org/10.1002/pst.527

Jensen SM, Kluxen FM, Ritz C (2019) A review of recent advances in benchmark dose methodology. Risk Anal 39(10):2295–2315

Kluxen FM (2019a) "New Statistics" for regulatory toxicology? [submitted, preprint available https://doi.org/10.13140/RG.2.2.14639.48803]

Kluxen FM (2019b) Scatter plotting as a simple tool to analyse relative organ to body weight in toxicological bioassays. Arch Toxicol 93(8):2409–2420. https://doi.org/10.1007/s00204-019-02509-3

Kobayashi K, Pillai KS, Sakuratani Y, Abe T, Kamata E, Hayashi M (2008) Evaluation of statistical tools used in short-term repeated dose administration toxicity studies with rodents. J Toxicol Sci 33(1):97–104

Koller M, Stahel WA (2011) Sharpening Wald-type inference in robust regression for small samples. Comput Stat Data Anal 55(8):2504–2515. https://doi.org/10.1016/j.csda.2011.02.014

Konietschke F, Placzek M, Schaarschmidt F, Hothorn LA (2015) nparcomp: an R software package for nonparametric multiple comparisons and simultaneous confidence intervals. J Stat Softw 64(9):17. https://doi.org/10.18637/jss.v064.i09

Kozak M (2009) Analyzing one-way experiments: a piece of cake or pain in the neck? Sci Agric 66(4):556–562. https://doi.org/10.1590/S0103-90162009000400020

Kozak M, Piepho HP (2018) What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions. J Agron Crop Sci 204(1):86–98. https://doi.org/10.1111/jac.12220

Levene H (1960) Robust tests for equality of variances. In: Olkin I (ed) Contributions to probability and statistics; essays in honor of harold hotelling. Stanford University Press, Palo Alto, pp 278–292

Lohse T, Rohrmann S, Faeh D, Hothorn T (2017) Continuous outcome logistic regression for analyzing body mass index distributions [version 1; peer review: 3 approved]. F1000Res. https://doi.org/10.12688/f1000research.12934.1

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist 18(1):50–60. https://doi.org/10.1214/aoms/1177730491

Matejka J, Fitzmaurice G (2017) Same stats, different graphs. Paper presented at the proceedings of the 2017 CHI conference on human factors in computing systems—CHI '17

Na J, Yang H, Bae S, Lim K-M (2014) Analysis of statistical methods currently used in toxicology journals. Toxicol Res 30(3):185–192. https://doi.org/10.5487/TR.2014.30.3.185

National Toxicology Program (2010) Toxicology and carcinogenesis studies of sodium dichromate dihydrate (CAS No. 7789-12-0) in F344/N rats and B6C3F1 mice (Drinking water studies). Technical report

Nature methods editorial (2014) Kick the bar chart habit. Nat Methods 11:113. https://doi.org/10.1038/nmeth.2837

Nuzzo R (2014) Scientific method: statistical errors – P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Nature 506:150–152

OECD (1998) Test no. 409: repeated dose 90-day oral toxicity study in non-rodents OECD guidelines for the testing of chemicals, section 4. OECD Publishing, Paris

OECD (2008) Test no. 407: repeated dose 28-day oral toxicity study in rodents. OECD Publishing, Paris

OECD (2010) Section 4: statistical and dose response analysis, including benchmark dose and linear extrapolation, NOAELS and NOELS, LOAELS and LOELS OECD guidance document for the design and conduct of chronic toxicity and carcinogenicity studies, supporting TG 451, 452 and 453. OECD Publishing, Paris

OECD (2014a) Current approaches in the statistical analysis of ecotoxicity data. OECD Publishing, Paris

OECD (2014b) Guidance document 116 on the conduct and design of chronic toxicity and carcinogenicity studies, supporting test guidelines 451, 452 and 453. OECD Publishing, Paris

OECD (2014c) No. 198 report on statistical issues related to OECD test guidelines (tgs) on genotoxicity. OECD Publishing, Paris

OECD (2016) Test no. 474: mammalian erythrocyte micronucleus test. OECD Publishing, Paris

OECD (2016) Test no.: in vitro mammalian cell micronucleus test 487. OECD Publishing, Paris

OECD (2018a) Test no. 408: repeated dose 90-day oral toxicity study in rodents. OECD Publishing, Paris

OECD (2018b) Test no. 451: carcinogenicity studies. OECD Publishing, Paris
OECD (2018c) Test no. 453: combined chronic toxicity/carcinogenicity studies. OECD Publishing, Paris
Pallmann P, Hothorn LA (2016) Boxplots for grouped and clustered data in toxicology. Arch Toxicol 90(7):1631–1638. https://doi.org/10.1007/s00204-015-1608-4
R Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
Ramsey FL, Schafer DW (2002) The statistical sleuth: a course in methods of data analysis. Thomson Learning, Duxbury
Salsburg D (2002) The lady tasting tea: how statistics revolutionized science in the twentieth century. Freeman, New York
Satterthwaite FE (1946) An approximate distribution of estimates of variance components. Biometr Bull 2(6):110–114. https://doi.org/10.2307/3002019
Schaarschmidt F, Biesheuvel E, Hothorn LA (2009) Asymptotic simultaneous confidence intervals for many-to-one comparisons of binary proportions in randomized clinical trials. J Biopharm Stat 19(2):292–310. https://doi.org/10.1080/10543400802622501
Schaarschmidt F, Sill M, Hothorn LA (2008) Poly-k-trend tests for survival adjusted analysis of tumor rates formulated as approximate multiple contrast test. J Biopharm Stat 18(5):934–948. https://doi.org/10.1080/10543400802294285
Schmidt K, Schmidtke J, Kohl C et al (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(3):731–751. https://doi.org/10.1007/s00204-015-1487-8
Schucany WR, Tony Ng HK (2006) Preliminary goodness-of-fit tests for normality do not validate the one-sample student t. Commun Stat Theory Methods 35(12):2275–2286. https://doi.org/10.1080/03610920600853308
Shaffer JP (1995) Multiple hypothesis testing. Annu Rev Psychol 46(1):561–584. https://doi.org/10.1146/annurev.ps.46.020195.003021
Steel RGD (1959) A multiple comparison rank sum test: treatments versus control. Biometrics 15(4):560–572. https://doi.org/10.2307/2527654
Student (1908) The probable error of the mean. Biometrika 6(1):1–25. https://doi.org/10.1093/biomet/6.1.1
Szocs E, Schafer RB (2015) Ecotoxicology is not normal: a comparison of statistical approaches for analysis of count and proportion data in ecotoxicology. Environ Sci Pollut Res Int 22(18):13990–13999. https://doi.org/10.1007/s11356-015-4579-3
Tukey JW (1977) Exploratory data analysis. Addison-Wesley Pub. Co, Reading
U.S. Food and Drug Administration (2001) Guidance for industry: statistical approaches to establishing bioequivalence
Wasserstein RL, Lazar NA (2016) The ASA's Statement on p-values: context, process, and purpose. Am Stat 70(2):129–133. https://doi.org/10.1080/00031305.2016.1154108
Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a world beyond $p < 0.05$. Am Stat 73(1):1–19. https://doi.org/10.1080/00031305.2019.1583913
Weissgerber TL, Winham SJ, Heinzen EP et al (2019) Reveal, don't conceal: transforming data visualization to improve transparency. Circulation 140(18):1506–1518. https://doi.org/10.1161/CIRCULATIONAHA.118.037777
Welch BL (1947) The generalization of `student's' problem when several different population variances are involved. Biometrika 34(1/2):28–35. https://doi.org/10.2307/2332510
Wheeler J (2019) Historical control data for the interpretation of ecotoxicity data: are we missing a trick? Ecotoxicology. https://doi.org/10.1007/s10646-019-02128-9
Wheeler MW, Bailer AJ (2007) Properties of model-averaged BMDLs: a study of model averaging in dichotomous response risk estimation. Risk Anal 27(3):659–670. https://doi.org/10.1111/j.1539-6924.2007.00920.x
Wickham H, Stryjewski L (2011) 40 years of boxplots. hadconz
Wilcox RR (2012) Introduction to robust estimation and hypothesis testing. Academic Press, Amsterdam
Wilk MB, Shapiro SS (1965) An analysis of variance test for normality (complete samples)†. Biometrika 52(3–4):591–611. https://doi.org/10.1093/biomet/52.3-4.591
Williams DA (1971) A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics 27(1):103–117. https://doi.org/10.2307/2528930
Zeileis A (2006) Object-oriented computation of sandwich estimators. J Stat Softw 16(9):16. https://doi.org/10.18637/jss.v016.i09
Zimmerman DW (1996) A note on homogeneity of variance of scores and ranks. J Exp Educ 64(4):351–362
Zimmerman DW (2004) A note on preliminary tests of equality of variances. Br J Math Stat Psychol 57(1):173–181. https://doi.org/10.1348/000711004849222