Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA

**Federal Food Safety and
Veterinary Office FSVO**
Risk Assessment

Binary data such as tumours from regulatory animal toxicity studies – Selected critical issues in statistical testing and the use of historical control data

July 2019

Table of contents

1. Summary

To improve transparency by fostering adherence to statistical guidance, the document addresses statistical issues regarding binary data (e.g. tumours or malformations) that are recurrently under discussion, including limitations of the statistical analyses performed, parameters that influence statistical evaluations, the power of toxicological studies, pairwise comparison versus trend tests, and appropriate use of historical control data (HCD). The main goal was to collect and comment the available relevant literature and regulatory guidelines.

The following conclusions were drawn. Binary endpoints should be evaluated using both statistical trend and pairwise tests. Factoring in survival and body weight disparities between study groups could reduce bias. For statistical tests investigating a possible trend or dose-relation systemic rather than nominal doses should be applied. Regarding the use of HCD, the following conclusions were drawn. The most relevant control group is always the concurrent control group. Any inclusion of HCD in a toxicological evaluation needs a clear description of the objectives. The suitability of HCD and the justification for its use require detailed explanation based on a biological and statistical reasoning. Statistically significant findings should never be discounted simply because they are within the HCD range or because concurrent control incidences are below the HCD average. If a specific endpoint is assessed using HCD, for an unbiased assessment all other endpoints should also be assessed based on the same HCD.

Thus, it is imperative that the existing guidelines for these issues are transparently applied at all times and scientifically sound rules on the use of HCD will be agreed in order to harmonise the conclusions drawn by different peer review groups. These statistical aspects should be seen as part of a weight of evidence judgement when forming the conclusion on a study.

2. Introduction

Animal toxicity studies are an essential part of the pesticide authorisation dossier, as they form the basis from which to derive human health-based guidance values (HBGV) such as acceptable daily intake (ADI) and acute reference dose (ARfD). Evaluation of animal toxicity studies facilitates identification of treatment-related hazards and assessment of dose-response relationships. Apart from continuous data such as body weight or clinical chemistry data, categorical, especially binary data, such as the presence or absence of tumours or malformations, are collected, presented, and analysed for biological relevance and statistical significance. Binary endpoints such as the presence or absence of tumours or malformations are among the most critical observations, as the subsequent hazard classification of compounds could restrict authorisation conditions and affect public perception, irrespective of the actual risk from realistic exposure.

The data requirements and basic principles of the toxicological evaluation of animal studies submitted for the authorisation of pesticides have been largely harmonised worldwide (OECD, 1994). This also applies to statistical methods used for interpreting the results. Many guidance documents have been published, along with related literature (IARC, 2006, OECD, 2015, USEPA, 2005).

The practice of dichotomizing results of statistical evaluations into significant and non-significant effects by applying arbitrarily chosen significance thresholds has long been highly controversial (Wasserstein *et al.*, 2019). Statistically non-significant results may well be treatment-related while statistically significant results may be unrelated to treatment. Nevertheless, in regulatory toxicology the practice of dichotomizing statistical analysis outcomes in significant and non-significant is still widely applied and therefore this paper is based on this approach. However, it is very important to be aware of the limits inherent in decisions based on fixed significance thresholds (e.g. $p< 0.05$). For a comprehensive review on the subject and alternative procedures in the field of statistical inference (Wasserstein *et al.*, 2019) and the references cited therein.

In many cases where elevated proportions of neoplastic lesions or malformations are observed with increasing doses of the test substance, the underlying mechanisms of action (MOA) responsible for these observations are unknown. In such situations, biological plausibility for a treatment relation is actively discussed; however, there are no established standards to define the quantitative changes that designate biological relevance in a general context (EFSA Scientific Committee, 2011). Thus, this may lead to inconsistent conclusions that substantially depend on expert judgement. Conclusions on possible causal relations between the treatment and observations would benefit from transparent statistical analysis and thorough documentation of the weight of evidence argumentation leading to the conclusion on the biological relevance (ECHA, 2016, EFSA Scientific Committee, 2017).

In the realm of randomised, controlled animal studies, statistics aid in inferring the relationship between the treatment and the observed effects. Once a statistically significant effect has been deemed treatment-related and biologically relevant in animals, its relevance in humans must be assessed. As all known human carcinogens that have been studied adequately in experimental animals have produced positive results in one or more animal species, IARC considers it biologically plausible that agents for which there is sufficient evidence of carcinogenicity in experimental animals also represent a carcinogenic hazard to humans (IARC, 2006). Accordingly, in the absence of MOA data that would allow to exclude human relevance, these agents are considered hazardous to humans (ECHA, 2017, USEPA, 2005).

When conducting an animal study, the statistical procedure should be established during the phase of the study design (OECD, 2015). However, sometimes also due to the age of the reports, statistical analyses do not comply with current expectations. In such situations, re-analysis of the data - either by the laboratory or by the regulatory assessor - might be considered (OECD, 2002). Providing the experimental results to the assessors in a structured electronic format compatible with statistical software packages, would facilitate re-analysis of the data.

In view of the large amount of data generated for each pesticide, coherence and transparency in the statistical evaluation would facilitate the comparison of the toxicological properties of different compounds and of evaluations performed by different authorities. The resultant increase in coherence and transparency in evaluations would in turn improve the reliability of read-across and *in silico* evaluation methods.

Evaluating the findings of tumours and malformations often stimulate discussions in peer review groups concerning the limitations of statistics, parameters that bias evaluation, the power of regular animal studies, pairwise comparisons versus trend tests, and the appropriate use of historical control data. The aim of the document is to emphasise that guidance is available for all these issues and to endorse its transparent application.

3. Data types in toxicity studies

The data generated in toxicological studies fall into the following broad categories:

1. Continuous data (measured values, e.g. body weight of animals)
2. Categorical data
    a. Binary data (e.g. tumour: yes or no)
    b. Ordinal scale (ordered categories, e.g. “mild”, “moderate”, or “severe” effects)
    c. Nominal scale (categories without order, e.g. reason of death such as “intercurrent death”, “interim sacrifice”, and “terminal sacrifice”)

3.1 Continuous data

The statistical procedures for intergroup comparisons of continuous data (e.g. body weights) and their interpretations are widely agreed and applied. If a parametric test such as the z or (Student) *t*-test is applied, the normality of the data (or sufficient sample sizes) and homoscedasticity (homogeneity of variance) should be considered. If the assumptions underlying the parametric tests are not fulfilled, non-parametric tests such as the Kruskal-Wallis test should be considered. However, non-parametric tests are slightly less powerful than parametric tests; while these are also not free from assumptions, they are more robust as they assess group medians.

For reasons discussed later, the statistical interpretation of continuous data will not be deepened here.

3.2 Binary data

Binary data in toxicological studies have two possible outcomes; the effect of interest is either present or absent. The number of animals with an effect of interest in relation to the total number of animals investigated in a group defines the proportion of affected animals. Typical binary data include histological findings such as the presence or absence of tumours or foetal anomalies in developmental studies. Assuming no treatment effect (null hypothesis is true), the probability of developing the effect in question is equal for all animals included in the study of a given sex. Hence, no statistically significant differences are expected in the proportion of affected animals between treatment groups. However, false positive results may occur by chance even in the absence of any treatment-related effect. Throughout the document, the terms “proportion” (of animals), “incidence”, and “rate” are used interchangeably. In principle, these terms all describe the ratio or percentage of animals with a certain outcome among a group of animals.

This document focuses on the statistical assessment of binary data such as the presence/absence of tumours and malformations, as conclusions based on these data often have far-reaching consequences for hazard and risk assessment. If the proportions of a tumour or a teratogenic response differ in a statistically significant manner between groups, the biological plausibility that this difference is related to a treatment must be assessed. If an observed adverse effect, such as the occurrence of a certain tumour, is considered treatment-related, then the compound of interest is considered to exhibit carcinogenic potential. The assignment of carcinogenic or teratogenic potential to a compound is independent of whether a threshold for triggering the effect exists or whether relevant human exposure occurs or not.

Regarding risk management and public perception, evaluation and interpretation of binary and continuous data have quite different implications. Continuous data such as changes in bodyweight nearly always allow for interpretation of the adversity, biological relevance, severity, and reversibility of the effect, even if the effect was undoubtedly treatment-related. In contrast, a treatment-induced incidence of 2%, corresponding to one single additional animal with tumour in the treatment group compared to control would certainly be considered an unacceptable public health concern.

Probably for such reasons, the effects described by binary data, especially if they are related to tumours or malformations, are usually scrutinised more rigorously during the review process than those described by continuous data. Therefore, this document primarily focuses on binary data. In the present document, tumour proportions are representatively used for binary outcomes. The principles discussed here pertain to other types of binary data such as malformations.

3.3 Ordinal data

Ordinal data such as the grading of histological findings (e.g. hepatocellular hypertrophy) into “mild”, “moderate”, or “severe” categories are rarely evaluated by statistical tests. One tool available for the analysis of categorical data by regression tools is the BMDS add-on CatReg (USEPA, 2017). Although the presence of the effect may be a clear binary outcome (yes or no for each animal), the grading may be somewhat arbitrary and dependent on the experience of the examining pathologist and the consistent application of the grading criteria. Moreover, the grading criteria used may be subject to changes (Gibson-Corley *et al.*, 2013). Therefore, a statistical evaluation of ordinal data may suggest a precision of evaluation, which does not exist and may even entail bias by leading to firm statistical conclusions based on imprecise categorisations. Therefore, in toxicological risk assessment statistical analysis of ordinal data should always be considered with care. Possibly for such reasons, ordinal data are rarely evaluated statistically.

4. Typical statistical testing of binary data

In view of the plethora of data generated in toxicological studies and presented to evaluating bodies, statistical analyses are applied for two purposes. First, statistical analyses assist in obtaining an overview of the data by flagging endpoints that differ significantly between groups for further scrutiny. Second, statistical arguments help to assess whether an (apparent) observed effect is (truly) associated with the treatment. Random assignment of animals to groups and proper statistical analyses ensure that statistically significant results are unlikely to have arisen by chance (OECD, 2002).

Although most currently available statistical evaluations of animal toxicity studies are based on frequentist hypothesis testing, alternative methods such as point estimations with confidence interval interpretation or Bayesian approaches have been proposed (OECD, 2015). However, frequentist hypothesis testing is still by far the most popular methodology. Therefore, this document focuses on frequentist hypothesis testing.

In its Guidelines for Carcinogen Risk Assessment, the US EPA recommends trend tests as well as pairwise comparison tests for determining whether chance, rather than a treatment-related effect, is a plausible explanation for an apparent increase in tumour incidence (USEPA, 2005). Since US EPA considers significance in either type of test is sufficient to reject the hypothesis that chance accounts

for the result, consistent with this policy it recommends that both pairwise comparisons and trend tests should always be performed (USEPA, 2005).
According to the methodological publications of the International Agency for Research on Cancer (IARC), priority should be given to trend tests, i.e., methods that are more powerful (and therefore more likely to indicate statistical significance, with lower false negative rates), when observed tumour rates increase monotonically with dose (IARC, 2006). The authors argue that a monotonic change in observed tumour rates with increasing dose should strengthen the inference that differences in tumour rates are due to exposure to the test substance, with steeper dose-response curves providing stronger evidence of an effect (Gart *et al.*, 1986).

4.1 Trend analyses

In trend analyses, the proportions of animals with a particular tumour type are analysed for a trend with increasing doses. The typically performed test for a linear dose-effect trend analysis is the Cochran-Armitage test (CA test) (Armitage, 1955, Cochran, 1954). In this test, an asymptotically chi-square-distributed test statistic accounts for the dose applied and the tumours observed in all the treatment groups. Under the null hypothesis, the absence of a linear trend with a slope greater than zero is tested, i.e. all groups are assumed to have equal tumour proportions. However, the CA test does not adjust for differential intercurrent mortality among the treatment and control groups, which can profoundly affect the conclusions of the carcinogenicity assessments. The tumour-initiating effects may be triggered shortly after the beginning of the treatment, but the tumours themselves may not develop until late in the study. Consider a compound with both the capacity to induce tumours and to shorten survival through non-tumour related general toxicity in treated animals. When no adjustment for differential survival is made in the analysis, the carcinogenic effect of such a compound may not be detected, because a considerable proportion of the treated animals was at tumour-risk for too short a time. In this case, the carcinogenic potential of the compound will be underestimated. Further, the carcinogenic potential of a compound may be overestimated if treatment increases survival compared to the controls, and this difference is not accounted for in the statistical analysis. Differences in intercurrent mortality that are not accounted for - irrespective of their statistical significance – could therefore bias statistical tumour evaluation. This is also true for the CA test, as mentioned earlier.
Several tests that adjust for differential intercurrent mortality have been developed. One test, which adjusts for differential mortality and prevents this possible bias, is the poly-k trend test. The poly-k trend test is an enhanced CA dose-effect trend test, where k (reflecting the tumour onset distribution) is usually set as k = 3, and is thus referred to as poly-3 trend test (Portier *et al.*, 1986, Portier and Bailer, 1989, Bailer and Portier, 1988, Bieler and Williams, 1993). In the test statistic, a risk weight, calculated as the fraction of the survival time of the whole study duration raised to the k^{th} power, is allocated to animals without tumours that die prematurely. The risk weight of an animal dying before study termination without tumour is therefore lower than that of animals dying prematurely with the tumour or that of animals sacrificed at study termination with or without tumours.
The Peto trend test is a dose-effect trend test that also adjusts for differential intercurrent mortality; it allocates higher weights to animals that died early, if the cause of death (COD) is considered causally linked to the tumour of interest (Peto *et al.*, 1980). Uncertainty about the classification of COD is addressed using a four-point scheme (STP, 2002, IARC, 2006). The Peto trend test relies on critical COD decisions that are not verifiable by evaluating authorities and that are not always apparent. Therefore, evaluating authorities may prefer the poly-3 trend test. However, both trend tests are recommended by the guidelines (OECD, 2015).

4.2 Pairwise analyses

For pairwise comparisons, the Fisher's exact test is typically used. In its basic form, it compares the binary outcome (tumour: yes/no) for two groups of different treatments, typically one treatment group and one control group. A 2 × 2 contingency table can be generated, with, for e.g. the variable "tumour" (yes/no) and the variable "treatment" (yes/no). Under the null hypothesis, the tumour proportions are assumed to be independent of the treatment (yes/no). The numbers in the four cells of the 2 × 2 contingency table follow a hypergeometric distribution, given the probability p for the observed outcome. Fisher's exact test does not adjust for differential survival. In a carcinogenicity study, the concern of the evaluating bodies lies rather in the increase than in the decrease in tumour incidences in the treat-

ed groups. Thus, a one-sided test may be considered more appropriate from a public health perspective because it tests for a difference in only one pre-specified direction and therefore has more power. The one-sided test increases false positive results and decreases false negative results by its increased power (1-β).
Furthermore, the CA and poly-3 tests (which account for differences in survival between the groups) can also be applied in a pairwise manner to support the identification of dose levels with no detectable increases in tumour incidence, compared to the controls (Peddada and Kissling, 2006).

5. Recurring issues in statistical evaluation of binary data

Besides treatment with a compound, other parameters influence the proportion of a particular tumour in a group, e.g. survival time, bodyweight. Some of these other factors are discussed in the following sections.

5.1 Intergroup disparities in survival

It is obvious that the survival time of animals is critical for identifying possible carcinogenic potential. Although the causative changes leading to tumour development might be induced very early in a study, the tumours themselves may not appear before the later stages of the study (see Section 4.1). Animals that die pre-term are therefore at risk of developing a tumour for a shorter time (Portier *et al.*, 1986). As differences in survival can lead to some degree of bias in the comparison of tumour rates, it is important to adjust for any difference, irrespective of statistical significance (Gart *et al.*, 1986). For both pairwise comparisons and trend analyses, methods such as the Peto and poly-3 tests are available, which account for differences in survival (Sections 4.1 and 4.2).

5.2 Intergroup disparities in body weight

Body weight of experimental animals correlates with incidences of certain tumour types (Haseman and Johnson, 1996, Haseman *et al.*, 1997) and other effects such as early onset of adverse metabolic events and endocrine-disruptive degenerative diseases (Keenan *et al.*, 1999, Keenan *et al.*, 1996). In studies employing *ad libitum* feeding regimes, animals usually become obese as the study progresses, and so, they may have an elevated risk for certain tumours, compared to less obese animals. It is not rare for animals in high-dose groups to have lower body weights than those in control groups because of impaired palatability of the feed supplemented with the compound of interest or because of the toxicity of the compound. If a tumour is body weight sensitive, the combined natural background and obesity-induced tumour incidence in a control group in total may exceed the gross tumour incidence in a less obese treatment group if the treatment-induced proportion is lower than the obesity-induced proportion. Hence, the carcinogenicity of a substance might remain unrevealed by an obesity-induced increase in tumour proportion in the control group (Seilkop, 1995).
Differences in survival and body weight between the control and treatment groups might act simultaneously to bias the observed tumour incidence. Increased body weight and lower mortality in the control group, compared to the treated groups, may lead to exaggerated background tumour proportions, concealing the true carcinogenic effect in the treated groups. Although body weight is often not accounted for quantitatively in the statistical analyses of toxicity studies, it should be considered at least qualitatively in the overall assessment as an additional mechanistic argument either supporting or questioning statistically significant and non-significant tumour proportions regarding their relatedness to chemical treatment. Alternatively, a logistic regression approach could be considered for adjusting for survival, body weight, and other confounding variables to reduce/eliminate bias.

5.3 Disproportionality between external and systemic dose

For many compounds, the systemically available dose in animals does not increase proportionally with the externally applied dose. Even if the systemically available dose increases over the whole dose range, it may not be proportional. The absorbed fraction may decrease with increasing doses, or the systemically available dose may reach a plateau. However, if enterohepatic re-circulation occurs or if excretion becomes saturated, the systemically available dose may increase over-proportionally, relative to the externally applied dose. Such deviations from linearity may have a wide range of causes, e.g. dose dependently inducible or saturable toxicokinetic and toxicodynamic processes. Examples of active substances in pesticides, whose systemically available doses are not proportional to the applied doses, are cyflumetofen, flufenoxuron, and metrafenone (FAO/WHO, 2014).

In pairwise statistical analyses based on Fisher's exact test, these kinetic variables are ignored, as the tests only discriminate between treatment and control. However, in the CA, Peto, and poly-k tests for trend, the dose is of pivotal importance in the test statistic. Therefore, if the external and systemic doses increase proportionally, it is not relevant which of the doses are applied in the trend tests, as the proportion between the doses remains unchanged. However, in all other cases, if the non-proportionality between external and systemic doses is not accounted for, the CA, Peto, and poly-k tests are biased.

In case the systemically available doses do not increase proportionally with the externally applied doses, the systemically available doses should be used in the statistical analysis, if the data are available. To derive oral absorption rates, specific toxicokinetic studies are usually performed with only two different dose levels, which are often not related to the doses tested in chronic studies (OECD, 2010). It is therefore desirable for the protocols of carcinogenicity studies to include provisions to estimate the systemically available doses of all groups at different times (ICH, 2017, ICH, 1994). This would not only allow for adjustments in the statistical analysis. Additionally, it would help to identify responses that possibly deviate from a linear relationship with systemic dose due to dose-related transitions in MOA responsible for the effect in question. It is conceivable that different MOAs responsible for the effect in question operate at different dose ranges and with different dose-relationships and hence result in deviation from a linear dose-response relationship (Slikker *et al.*, 2004a, Slikker *et al.*, 2004b).

5.4 Informative value of effects at high doses

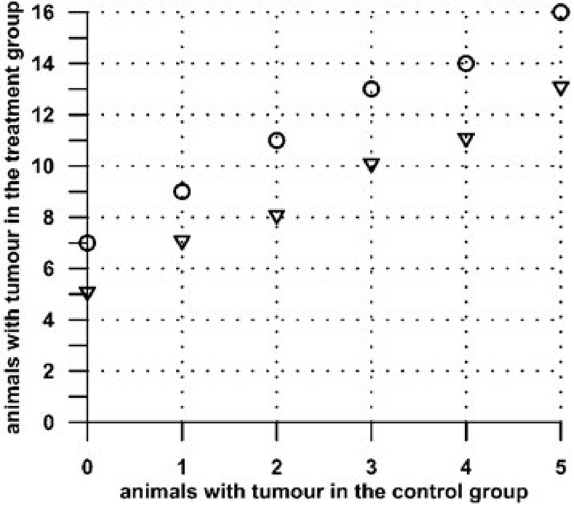
The majority of chemicals tested in carcinogenicity studies on rodents have shown carcinogenic potential at high doses (Gaylor, 2005, Johnson, 2002, Johnson, 2003, Gold *et al.*, 1989, Gold *et al.*, 2005). A high dose is defined as a dose above the maximum tolerated dose (MTD), which is defined by the OECD as the dose that is not lethal and that does not decrease bodyweight gain by more than 10% (OECD, 2002). It is important to note that laboratory animals are often overweight in toxicological studies and body weight reduction is not necessarily an indication of severe toxicity. However, doses are often claimed to exceed the MTD simply because the bodyweight gain is reduced by 10% or more, even though survival is not affected and no excessive general toxicity was found. Consequently, the relevance of the tumour responses found at these doses and their predictive value for lower doses are disputed. The argument to disregard high-dose tumour incidences is that, because toxicokinetic and toxicodynamic processes may change along the dose-response curve (Slikker *et al.*, 2004a, Slikker *et al.*, 2004b), high dose mechanisms of toxicity may sufficiently alter physiology to induce tumours but may not operate at lower doses that are of ultimate interest for risk assessment (Boobis *et al.*, 2016, Gaylor, 2005). Although this argument seems plausible, analysis of available rodent carcinogenicity assays supports the predictive power of high-dose findings. The vast majority of compounds exhibiting carcinogenicity above the MTD also showed increased site-specific tumour incidences at lower doses, either numerically or in a statistically significant manner (Haseman and Lockhart, 1994). These correlations justify the assumption that tumours observed at high doses might indicate carcinogenicity at lower doses. In the evaluation of the human relevance of effects observed in animals at high doses, the interdependence between low statistical power (resulting in high false negative rates) to identify the effects of low magnitude at low doses and high-dose effects in animal studies must always be considered: statistically insignificant increases in effects at lower doses may be attributed to the low power of the study. This reflects the fact that the magnitude of the effect at low doses may be too small to be identified as treatment-related, as the group sizes are too small and the statistical tests are underpowered. This lack of power should not be misinterpreted as proof of a threshold (Crump *et al.*, 1999).

5.5 Power of statistical tests

The statistical power of a study is defined as the probability of correctly identifying a treatment-related effect by rejecting the null hypothesis if it is false. The power of a statistical test essentially depends on the sample size and the effect size.

In toxicological evaluations of pesticides, tumour rates in the control and treated groups are most commonly assessed by pairwise comparisons using Fisher's exact test. However, as illustrated in the figure below (Figure 1), the power of Fisher's exact test to detect a treatment-related effect is limited, i.e., the probability of obtaining false negative conclusions (type II error) is high.

Figure 1: Fisher's exact test: significant tumour incidences in relation to background proportions



In a hypothetical experiment with 50 animals each in the control and treatment groups, the minimum number of tumour-bearing animals required in the treatment group to reach statistical significance (triangle: significance level α = 0.05; circle: significance level α = 0.01) was plotted as a function of the number of tumour-bearing animals in the control group.

Even if no tumours are detected in a control group of 50 animals, Fisher's exact test is only significant (at 5% level of significance) if 5 out of 50 (10%) animals in the treated group develop a tumour. Consequently, Fisher's exact test can only detect effects of at least 10% and even more if the background incidence is greater than 0%. Furthermore, Fisher's exact test does not adjust for differential survival. If the survival rate in the treatment groups is lower than that in the control group, the performance of Fisher's exact test is further impaired, because the time at risk for animals in the treatment groups is lower than that for the control animals. Using the poly-3 test in a pairwise manner is one possible way to attenuate this effect (see Section 4.2).

To increase the statistical power of a study, the sample size could be increased. Application of high doses may increase the effect magnitude and the chance to detect the carcinogenic effect of the test material in the study, which eases the correct identification of carcinogenic potential. Therefore, dose selection should be focused on maximising the chance to detect the carcinogenic effect of the test material in the study at the highest tested doses (Haseman, 1984, OECD, 2015) and hence high dose testing is scientifically defensible (Section 5.4).

Further, it would be desirable to include the performance of power analysis in the statistical evaluation of animal studies, which would allow us to assess the uncertainties in the statistical analysis performed, thus improving decision-making and interpretation of the statistical evaluation.

5.6 Background incidence and false positive and false negative rates

The background incidence has an impact on the false positive and false negative rate both in pairwise and trend tests (Fears *et al.*, 1977, Lin and Rahman, 1998). By tendency, high background incidences increase false positive rates but decrease false negative rates and low background incidences decrease false positive rates but increase false negative rates. Thus, tumour incidences found to be statistically significant in either pairwise or trend tests in a study with low background incidence can be considered reliably attributable to treatment and not be a false positive finding. For some more details see chapters later on.

5.7 Correction for multiple comparisons

Random assignment of animals to groups and proper statistical analyses maximally reduces the probability that statistically significant results have arisen by chance alone (OECD, 2002). However, the pool of animals from which the animals are randomly allotted to dose groups probably nearly always is composed of subpopulations of different background incidences and different sensitivity (Festing, 2010, Festing, 2016). Consequently, the randomized allocation of animals to dose groups stratified by body weight only may result in unevenly distributed sensitivities. Hence, false positive or false negative conclusions regarding a compound's potency to induce tumours cannot be excluded.

Correction methods are often used to control for the overall false positive rate (family-wise error rate) in carcinogenicity screening experiments with multiple tests (two species, two sexes, and 20-30 organs). Several correction methods have been described. One possibility is to use different significance levels based on the different background incidences of tumour types (Lin and Rahman, 1998). However, this procedure is highly dependent on *a priori* knowledge of the spontaneous background proportions for the strain at the time of the study. Usually, this information is not available, but could be derived from historical control data (HCD). However, the use of HCD for statistical purposes is associated with many uncertainties, as described in a separate document.

Another method to resolve the multiple comparisons issue and ensure that family-wise error rates do not exceed the significance level (α) is to adjust the significance level based on the number of tests performed. Different methods have been described for this purpose (Bonferroni, 1936, Sidak, 1967, Holm, 1979). Although the Sidak correction is slightly less conservative than the Bonferroni method, there are situations in which both are overly conservative, i.e., the actual family-wise error rate is considerably smaller than the predefined α. The approach described by Holm is more powerful than either of these methods (Holm, 1979).

Possible consequences of significance level corrections are illustrated by the following study (Table 1), in which acrylamide (a known carcinogen classified according to Annex VI of Regulation (EC) No 1272/2008 into category 1B (EU, 2008)) was administered to male F344 rats through drinking water (NTP, 2012). For tumour types given in the table below, only the proportion of mesothelioma in epididymis in the highest dose group is significantly increased, compared to the control, by Fisher's exact test (α = 0.05); all the other listed tumour types show a dose-related positive trend in tumour proportions with the CA test (α = 0.05). If the significance level was adjusted to α = 0.00042 by applying the Bonferroni method, none of the tests would be deemed significant.

Table 1: Tumour incidences in an acrylamide drinking water study in male F344 rats

Dose (mM in drinking water)	0	0.0875	0.175	0.35	0.70	p value	
Number of animals	48	48	48	48	48	F[a]	CA[b]
Epididymis: Mesothelioma	2	2	1	5	8	0.046	0.003
Heart: Schwannoma	1	2	3	4	6	0.056	0.028
Thyroid Follicular Cell Carcinoma	1/47	2	3/47	6	6	0.059	0.026
Pancreatic Islets: Adenoma	1/46	2	4	1	6	0.062	0.056
Testes: Mesothelioma	1	2	1	1	5	0.102	0.040

[a] *Fisher's exact test (high-dose versus control)*

[b] *Cochran-Armitage trend test*

This example illustrates that, to identify possible treatment-related binary effects such as neoplastic lesions, a correction of the significance level using Bonferroni or Sidak methods is questionable. Only extremely potent carcinogens would be identified by this approach.

Therefore, to screen for carcinogenic effects, it is more reasonable to assess the biological importance of these tumours, which are significantly different from the controls at the usually applied significance level of 0.05, instead of adjusting for multiple testing *a priori*. Furthermore, apart from the genotoxic mechanisms of action, different neoplastic lesions could develop through different (perhaps independent) mechanisms of action, and can therefore be regarded as different unrelated outcomes. It is therefore not reasonable to evaluate neoplastic lesions of different aetiology at decreased significance levels, just because the apical manifestations (tumours) are grouped together terminologically for convenience of classification. Significance level correction methods ignore the multiplicity of biological mechanisms of action that lead to different neoplastic lesions and do not consider different tumour types as individual lesions.

If evidence suggests that several neoplastic lesions were caused by common initiating mechanisms (e.g. mutagenicity) or represent different stages of tumour progression (adenoma → carcinoma), the correction for multiple testing becomes even more questionable. In such cases, an overly conservative significance level correction might cause type II errors (retaining the null hypothesis, although it is false), although biological plausibility for multi-site neoplastic lesions is given, but, for example, the mutagenic potency is not extreme (see acrylamide example above).

To decrease false negative rates and increase the power of the study, the sample size or the magnitude of the effect could be increased by increasing the doses. An alternative approach would be to apply higher significance levels for statistical tests, thus lowering the probability of falsely concluding that there is no effect. However, this would increase the false positive rate, and therefore requires thorough justification. Nonetheless, availability of data regarding the power to detect a certain tumour type could substantially improve decision-making.

Placing restrictions on pesticide products based on false positive conclusions regarding the carcinogenic potential of their active ingredients may have economic consequences, at least for the applicant. If the concerned pesticide has more desirable agronomic or environmental properties than those already in the market, these restrictions may also have food security and environmental implications.

The public health bodies' remit is to prevent possibly detrimental exposures of the population. Consequently, from this perspective the control of false negative rates is of central interest. Because any endeavour to reduce false positive rates increases false negative rates, eventually it is policy to decide which ratio of false negative to false positive conclusions is acceptable. Additionally, if carcinogenic potential is not identified until after the pesticide is marketed (either due to the low power of carcinogenicity studies or due to inappropriate statistical methods), the whole evaluation process of pesticides may be compromised, triggering concerns about its reliability. In summary, from a public health perspective, the false negative proportion is of more concern than the false positive proportion. In this regard, measures should primarily focus on the reduction on the false negative proportions and only secondary on the reduction of false positive proportions.

6. False positive and false negative rates

In statistical hypothesis testing, the null hypothesis is either rejected or retained. The decision to reject the null hypothesis or not is influenced by two types of error. One is to conclude that the treatment has an effect (rejection of null hypothesis), although there is none (type I error, false positive). The probability of a type I error in a test is denoted by α, the significance level. The second type of error is to conclude that the treatment has no effect (retention of null hypothesis), although there is one (type II error, false negative). The probability of retaining the null hypothesis although it is false is denoted by β. The power of a test, i.e., the probability to correctly reject the null hypothesis when in fact the alternative hypothesis is true, is 1 − β.

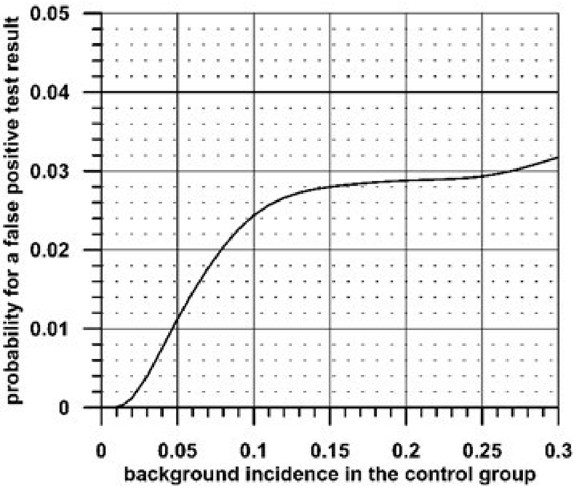
Type I and II error rates are determined by the strength of the biological effect, study design parameters (e.g. group sizes), and statistical methods used (e.g. pairwise or trend tests).

	Null hypothesis is true	Alternate hypothesis is true
The null hypothesis is retained	1- α	β (type II error)
The null hypothesis is rejected	α (type I error)	1-β (power of the analysis)

6.1 False positive rates in pairwise tests

Fears et al. calculated the probability of false positive results in one-sided Fisher's exact tests for a treatment group in relation to the tumour proportion in the control group (Fears *et al.*, 1977). For a given background proportion, the binomial probability of observing x tumour-bearing animals in the control group and at least y tumour-bearing animals in the treatment group was calculated. As mentioned earlier, at least five tumour-bearing animals in a treatment group of 50 animals are necessary to yield a positive result with Fisher's exact test, given that there are no tumour-bearing animals in the control group of 50 animals. If the background proportion of a tumour in a given tissue is, say 0.03, the binomial probability for the absence of tumours in the control is 0.2181 and that for the presence of at least 5 tumours in the treatment group is 0.0168. The product of these two probabilities is the weighted probability (P = 0.0037) of observing at least 5 tumours in the treatment group and none in the control (a false positive result), if the background proportion for a tumour in the given organ is 0.03 and the null hypothesis that all groups have equal tumour proportions holds true. For all possible distributions of tumours between a control and a treatment group yielding a significant Fisher's exact test result, the probabilities are calculated accordingly. The sum of all these probabilities would be the overall probability (P = 0.0039) of a false positive test result in a pairwise, one-sided Fisher's exact test for the specific tumour type, if the background proportion is 0.03. These overall probabilities for a tumour in an organ can be calculated for all background proportions between 0 and 1. For typical background proportions of many tumour types (≤0.06) (Haseman and Elwell, 1996), the false positive proportions show a maximum value of 0.016 or lower (Figure 2).

Figure 2: Probability for a false positive result in relation to a tumour's background incidence



The figure depicts the probability of obtaining a false positive test result in pairwise Fisher's exact test (50 animals/dose group) for a tumour type with a given background incidence.

If the proportion of one tumour type in a carcinogenicity study is shown to increase in a treatment-related manner, the tested compound would be considered carcinogenic. Therefore, the probability of having at least one false positive increase in a typical carcinogenicity data package is relevant. Approximately 20-30 organs in two sexes of two species (mouse and rat) are usually examined for tumours. Thus, 120 hypotheses are tested when comparing a control group to a treatment group. Table 2 shows the calculation of the upper boundaries for the false positive proportions of a hypothetical tumour background. In this virtual example, it is assumed that very rare, rare, common, and frequent tumours are distributed similarly between sexes and species. For given distributions of background incidences in 25 organs, the overall false positive rates are calculated using one-sided Fisher's exact tests. The results showed that, for one sex in a given species, the overall false positive proportion is 5%; for both sexes of one species, it is approximately 10%; and for both sexes of both species, it is approximately 20%; however, if the organ with the highest background proportion (0.5) is excluded, the proportions fall to approximately 2%, 5%, and 9%, respectively. Essentially, if a false positive rate per organ ≥ 1% is considered unacceptable, only tumours with background proportions > 0.045 (i.e., at least 3 out of 50 control animals bearing a specific tumour) are of concern, regarding their contribution to the overall false positives (see Figure 2). In the virtual example outlined in Table 2, this would only concern the three organs with the highest background incidence, while the others do not substantially contribute to the overall false positive rate.

Table 2: False positive results from Fisher’s exact test in a hypothetical carcinogenicity study

25 organs examined; organs categorised according to their spontaneous background tumour rates (in brackets)	False positive rates per one sex of one species	
	organ	category
16 organs ***very rarely*** (0.005) having a tumour	0.000004	0.000068
6 organs ***rarely*** (0.010) having a tumour	0.000088	0.000530
2 organs ***commonly*** (0.050) having a tumour	0.011231	0.022462
1 organ ***frequently*** (0.500) having a tumour	0.028670	0.028670
False positive rate per one sex of one species = sum of all organs		0.051730
False positive rate per two sexes of one species = sum of all organs × 2		0.103460
False positive rate per two sexes of two species = sum of all organs × 4		0.206920

Reading example: The theoretical probability in this virtual example of finding a statistically significant tumour incidence in the brain is set to be very low (0.005). Therefore, to find a statistically significant tumour incidence in the brain, if the treatment does not increase the probability of developing a brain tumour, is 0.000004 (false positive rate). The other 15 organs of this background incidence category will also have the same false positive rate (0.000004). Therefore, the overall false positive rate (to find at least one statistically significant tumour type in at least one organ of this category) for this category of organs is 16 × 0.000004 = 0.000068 (rounded). The sum of all the background incidence categories gives us the probability of a false positive tumour type in one sex of one species, which is 0.051730.

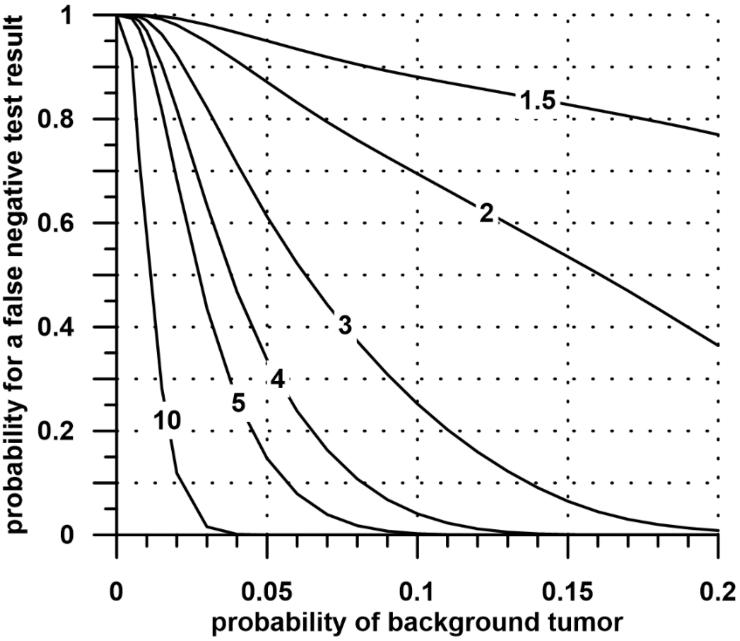
6.2 False positive rates in trend tests

Trend tests are usually more powerful than pairwise tests, as they integrate information from all treatment groups simultaneously and consider the dosage. The power of survival-unadjusted trend tests (e.g. CA test) has been estimated through simulations (Lin and Rahman, 1998). Based on such analyses, false positive rates for a given significance level with trend tests were found to be approximately twice as large as those with pairwise comparisons. Similar to pairwise comparisons, in carcinogenicity assays as well, tumour types with high spontaneous background proportions contribute disproportionally to the overall false positive proportion. Spontaneous background proportions of <1% do minimally increase the false positive proportion. Lin et al. found that the overall false positive proportion in a typical carcinogenicity screening (two species, two sexes, 20–30 organs) was approximately 10% when the significance levels for rare tumours (background incidence ≤1%) and common tumours (background incidence >1%) were adjusted to 0.025 and 0.005, respectively (Lin and Rahman, 1998).

6.3 False negative rates in pairwise tests

Fears et al. calculated the false negative proportion as the cumulative binomial probability for the maximum proportion yielding no statistical significance, in relation to the proportion in the control group, by applying Fisher’s exact test (Fears *et al.*, 1977). For example, if no tumour-bearing animals are observed in the control group, the presence of five tumour-bearing animals in the treatment group would be statistically significant in Fisher’s exact test. The false negative proportion reflects the binomial probability of observing a maximum of four tumour-bearing animals (Fisher’s exact test negative) in the treatment group, assuming that the null hypothesis that treatment and control groups have equal tumour incidence is false. Similar to the overall false positive proportion, the overall false negative proportion for one organ can be calculated as the sum of the probabilities for each combination of control and treatment proportions, but using different probabilities for tumour development in the control and treatment groups (i.e., the treatment has an effect). To exemplify this, it is assumed that the background probability for the development of a particular tumour in an organ in the control was 0.01; this probability was then increased to 0.05 (corresponding to an extra risk of 4.4% in a group of 50 animals) by the compound in the treatment group. The overall false negative proportion for this organ is ca. 0.93 (93%), and it increases to nearly 1 (100%) if the treatment increases the probability of developing a tumour by less than fivefold (see Figure 3), compared to the control. Thus, the probability of missing a true effect is very high if the effect has a low magnitude and the background probability of control animals developing tumour is low. The false negative proportion decreases with increasing background incidence in the control.

Figure 3: Probability for a false negative result in relation to a tumour's background incidence



The figure depicts the probability of obtaining a false negative test result (50 animals/dose group, Fisher's exact test) for a tumour type with a given background incidence and a given x-fold (number in the curve) increase in tumour probability through treatment. The calculations and the assumptions based on are described elsewhere (Fears *et al.*, 1977). *Reading example: with a background tumour probability of 0.05, the probabilities for false negative results are 95%, 87%, 61%, 34%, 15%, and 0% if the treatment increased the probability of developing a tumour by 1.5-, 2-, 3-, 4-, 5-, or 10-fold, compared to the background tumour probability.*

6.4 False negative rates in trend tests

Gaylor analysed the power of trend tests such as the CA trend test, which was applied by NTP to evaluate carcinogenicity studies, using 50 animals/dose group (Gaylor, 2005). He found that increasing group size from 50 to 200 animals also increased the percentage of positively tested compounds from 62% to 92%. Thus, the author concluded that the false negative rate of trend tests in typical carcinogenicity assays including high doses at the MTD might be around 30%.

7. Recurring issues in the use of historical control data (HCD)

7.1 Variability of tumour incidences has biological causes

The interpretation of toxicological experiments assumes that any effect results from a biological mechanism of action (MOA). MOAs are triggered either by interactions of animal-inherent biological factors with a substance administered to the animals or by substance-unrelated animal-inherent biological factors alone or by their interactions with the environment. Therefore, a key paradigm of toxicological experimentation is that any response, be it treatment-related or not, arises from an underlying but often unknown MOA possibly sensitive for many animal-inherent and environmental factors. Hence, sound toxicological experimentation requires the inclusion of a control group for which all conceivable factors are as similar as possible to the chemical-treated groups. Against this background, invariable and variable tumour incidences in a broad set of control groups provide opposite information regarding the MOAs inducing the tumours. Essentially invariable incidences for a specific tumour in a broader set of control groups conducted under varying conditions indicates that the respective MOA is insensitive for changes in environmental factors. Variable incidences for a specific tumour, however, indicate high sensitivity of the respective MOA regarding changes in environmental factors. HCD are a collection of data derived from control groups of other studies considered helpful for interpreting the findings of the study under evaluation. Consequently, an HCD of essentially invariable tumour incidences is more reliable than one of variable tumour incidences because the former apparently is insensitive to factors that may change unrecognisably from study to study.

7.2 Background incidence, randomization of animals and HCD

If the causes are unknown, the tumours observed in untreated animals are referred to as "chance" observations and summarised as "spontaneous background incidence", related to unknown biological reasons. Large differences in tumour incidences are known to result from disparities in factors such as pathology nomenclature, test animal strain, husbandry, the investigating pathologist, and other differences in experimental conditions (OECD, 2015). Thus, unrecognised factors substantially contribute to what is summarised as "spontaneous background incidence" for any given setting. So it's important to be aware that the notions "spontaneous background" and "chance" do not mean that there is no reason for an observation. Rather, they mean that the causes of tumours and their interactions (summarized as MOA) are complex and unknown. For this very reason, the use of historical control data is extremely sensitive. Therefore, randomisation of largely genetically uniform animals and uniform study conditions to avoid bias as much as possible are crucial to ensure that the differences between dose groups can be solely attributed to the substance under investigation. Problems resulting from the use of genetically insufficiently uniform animal strains that include subpopulations of unknown sensitivities, e.g. outbred strains, have been discussed elsewhere and the use of panels of inbred strains was recommended (Festing, 2016).

The implicit, but rarely further substantiated, justification for the use of HCD is the assumption that the experimental frame (including all biological, study design, evaluation, and reporting aspects) in which the HCD were generated is comparable with that of the present study. Guidance documents emphasise the importance of the concurrent control group for testing for increased tumour rates (USEPA, 2005, Peddada *et al.*, 2007, OECD, 2002, OECD, 2015, Gart *et al.*, 1986), implying that the use of HCD to reach conclusions should require detailed scientific justification. Given that HCD were derived from studies that are sufficiently comparable with the study being investigated, they might be useful under certain circumstances. However, the concerned study and the studies of the HCD were conducted at different points of time, and so, the HCD animals were not assigned from the same pool as the ones in the investigated study. In addition, other parameters such as investigating pathologist and feed usually do not match between the studies comprising the HCD and the study under investigation. Thus, it is questionable if HCD is sufficiently comparable to the study under investigation. At any rate, clear reasoning should be presented regarding why the HCD is considered sufficiently similar to the concurrent groups. This reasoning must reach beyond the assertion that the HCD studies were conducted at the same laboratory, in the same animal strain, and within a 5 year period centred as closely as possible on the date of the study under review (see Chapter 7.4.1).

In view of the enormous significance that HCD gain when used to evaluate a study, its reporting should be as detailed as that of the study of interest. A clear reasoning on why and how the HCD will be used to evaluate the study under review has to be provided through informal comparison or statistical analysis.

7.3 Critical points in the use of HCD

In practice, the use of HCD typically is triggered simply by the presence of statistically significant findings, particularly if MOA considerations neither plausibly support nor challenge the treatment-relatedness of the finding.

If HCD for tumour incidences comprise only a limited number of recent control groups, e.g. 2-5, firm conclusions on the tumour incidence distribution are precluded, and therefore no real gain in power for estimating background tumour proportions can be achieved. Thus, the scientifically defensible benefit of using HCD consisting of a limited database for the assessment of a study needs to be clarified. Moreover, when an attempt is made to assess whether the concurrent control data are appreciably "out of line" with respect to the HCD, a relationship between the treatment and the increased tumour proportion in the study is often questioned, if the tumour proportions observed are within the range of HCD proportions. Without further specifications, such statements remain vague. Reference to the range could imply that the observed incidences in the study under review are close to either end of the distribution of historical control proportions or that they are somehow close to the mean, median, or another measure of its distribution. However, the use of the maximum proportion (upper bound) of HCD suggests that the single maximum proportion found in the HCD is the most appropriate value to compare with the proportion of the concurrent control group. Clearly, this is fallacious, as this reasoning confers all the interpretative weight on the maximum value found in HCD, which might well be an outlier. Furthermore, the use of the upper bound of the HCD as a reference disregards the distribution of values in the HCD. Further, each study included in the HCD potentially widens the range between the upper and lower bounds, but never reduces it. The situation is aggravated if the upper bound represents a poorly characterised data point. Moreover, if the concurrent control rate falls outside the historical range, there may be concern about whether the concurrent control and treatment groups (i.e. the study in general) are consistent with and comparable to the historical control groups (and studies) (Dinse and Peddada, 2011). This would indicate that either the study under evaluation is untypical for normal expectations (as defined by HCD) or vice versa. If the former conclusion is considered, the validity of the study would be doubtful.

Hence, the highest incidence found in a HCD, i.e., “the upper bound of incidences”, is a statistically meaningless value and should never be used. Additionally, mortality patterns and body weight distributions, parameters that may considerably impact tumour incidences, are disregarded when tumour incidences in the study of interest are simply compared to the HCD range. Accordingly, accepted guidelines state that statistically significant increases in tumours should not be discounted simply because incidence rates in treated groups are within the range of HCD, or because incidence rates in the concurrent controls are lower than average. Random assignment of animals to groups and proper statistical analysis should ensure that statistically significant results are unlikely to arise by chance alone (Dinse and Peddada, 2011, OECD, 2002, USEPA, 2005). Challenges and pitfalls in using HCD for study interpretation have been discussed by many authors (Keenan *et al.*, 2009, Haseman *et al.*, 1984, Elmore and Peddada, 2009), and along with guidance documents (OECD, 2002, USEPA, 2005), they discourage researchers from comparing tumour incidences of a study to HCD ranges.

7.4 Recommended use of HCD

7.4.1 Consider only HCD fully characterized according to agreed criteria

It has been established that tumour proportions in any given animal strain often are not stable and are influenced by many factors, including naturally occurring shifts in proportions with time and different aspects of husbandry such as chow type (Haseman *et al.*, 2003, Tennekes *et al.*, 2004a, Tennekes *et al.*, 2004b). Therefore, the EU (EU, 2011) requires the following specifications for the use of HCD, which is also endorsed by the Joint FAO/WHO Meeting on Pesticide Residues (JMPR) (WHO Core Assessment Group on Pesticide Residues, 2015), used for interpreting tumour proportions:

- identification of species and strain, name of the supplier, and specific colony identification, if the supplier has more than one geographical location;
- name of the laboratory and the dates of the studies performed;
- description of the general conditions under which animals were maintained, including the type or brand of diet and, where possible, the amount consumed;
- approximate age (in days) and weight of the control animals at the beginning of the study and at the time of sacrifice or death;
- description of the control group mortality pattern observed during or at the end of the study, along with other pertinent observations (such as diseases or infections);
- name of the laboratory and the examining scientists responsible for gathering and interpreting the pathological data from the study;
- a statement of the nature of the tumours that may have been combined to produce any of the rate data.

If these requirements are satisfied, the candidate HCD data should be examined to determine whether it is sufficiently similar to the concurrent control group (Haseman, 1995). Besides tumour incidence, body weight development, survival rate, and clinical chemistry and haematological data should be considered. Differences among the HCD or between the HCD and the concurrent control group in any of these parameters implies that the animals are not derived from the same population. As these control groups were not assorted through random assignment of animals out of a single common pool, such differences are to be expected. Thus, disparities suggesting that concurrent control and HCD groups are samples from different populations indicate that the HCD are not suitable for assessing the study being investigated.

The prerequisites for reliable HCD warrant that it is not biologically discernible from the concurrent control group. Hence, if the HCD is considered acceptable, it may be used to increase the power of the study by increasing the number of control animals and decreasing the variability of the true background values of measured variables such as tumour proportion. These prerequisites ensure that only very few studies provide eligible HCD. However, the extent of reporting on HCD data usually available to evaluating experts is limited, which severely hampers a sound evaluation of whether the concurrent control group and HCD are samples from the same population. With 2–5 groups typically comprising the HCD, it is not feasible to scrutinise whether the tumour proportion of the concurrent control derives from the tumour proportion distribution of the groups constituting the proposed HCD without considerable uncertainties.

Accepting HCD and using them for study interpretation not only introduces uncertainties and arbitrariness, but also biases the assessment of animal studies that already have low statistical power. This ultimately leads to the conclusion that the concurrent control group should always be the most important consideration while testing for increased tumour rates (OECD, 2015).

7.4.2 Use appropriate HCD incidence distribution measures

Practical experience shows that tumour incidences in HCD often exhibit substantial heterogeneity of unknown distributions (Tarone, 1982, Tarone *et al.*, 1981, Tennekes *et al.*, 2004a) . Therefore, for comparing the concurrent control group with the HCD, the mean with standard deviation, median, and confidence interval of the tumour incidences should be considered, but never the range of the HCD. Which measure is most appropriate depends on the tumour incidence distribution in HCD. For example, if the incidence distribution of a rare tumour in HCD appears right-skewed log-normally distributed the median may be a more appropriate measure than the mean.

Furthermore, formal trend tests for comparing the HCD with the concurrent control group (Peddada *et al.*, 2007, Sun, 1999, Tarone, 1982) have been proposed. These tests depend on detailed analyses of tumour proportions in the HCD, compared to those in the concurrent control, and are therefore not completely straightforward. Currently, these tests are not routinely used in a regulatory context, and to date, regulatory bodies have not agreed upon appropriate methods.

Because the animals were randomised between control and treatment groups, differing concurrent control group and HCD tumour incidences are not an *a priori* indication that the concurrent control group is somewhat peculiar (USEPA, 2005). It should be emphasised that differences in tumour incidences in the concurrent control group and HCD only indicate that the concurrent control group and HCD are not from the same population and so, the HCD should be excluded as a reliable source of information (Dinse and Peddada, 2011).

7.4.3 Balanced evaluations require uniform control data bases for all endpoints

If HCD is considered appropriate for interpreting the incidence of a specific tumour in a study, it should be applied to all endpoints, including all tumour types (even those without apparent increase when compared to the concurrent control), to ensure a uniform assessment of all endpoints of the study under consideration using the same database. The use of different control databases to evaluate different endpoints in a study violates fundamental principles of biological experimentation, and therefore needs comprehensive scientific justification. Choosing only a single endpoint (e.g. type of tumour) to compare against the HCD based on the mere fact that its incidence is higher than in the concurrent control distorts the coherence of the evaluation. If the same database, i.e. the HCD deemed appropriate to evaluate a specific endpoint, is not likewise used to reconsider initially unremarkable endpoints the procedure might be understood as a deliberate effort to rationalise away initially suspiciously elevated tumour incidences. Such selective procedures distort the overall evaluation towards an underestimation of the toxicological profile of compounds.

7.4.4 Consider only reliable HCD with stable incidence distributions

Probably the scientifically most defensible contribution of HCD is in the assessment of very rare tumours in situations where, for example, one animal out of fifty in the highest dose group is observed with a tumour that is observed neither in the concurrent control nor in the lower dose groups. The proportions of such tumours will hardly be statistically significant in any test. However, if the tumour of interest is also extremely rare in the HCD, it would indicate that, although not statistically significant, the low-incidence tumours at the highest doses might be treatment-related. If the tumour of interest is also extremely rare or exhibits a very narrow distribution in a broader HCD that does not satisfy the strict requirements for its consideration, it would imply that the background incidence of the tumour in question is largely insensitive to experimental conditions, except chemical treatment. This information would strengthen the conclusion that the increased incidence of the tumour found in the study is treatment-related, even if it is not statistically significant. Accordingly, the US EPA guidelines emphasise that, while analysing uncommon tumours in a treated group that are not statistically significant compared to concurrent controls, the evaluation may benefit from the experience of HCD to conclude that the result is unlikely to be due to chance (USEPA, 2005).

8. Acknowledgment

We would like to thank the Bundesinstitut für Risikobewertung (BfR), Germany, for their helpful comments and suggestions.

9. References

Armitage, P. (1955). Tests for Linear Trends in Proportions and Frequencies. *Biometrics,* **11,** 375-386.

Bailer, A.J. and Portier, C.J. (1988). Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. *Biometrics,* **44,** 417-431.

Bieler, G.S. and Williams, R.L. (1993). Ratio Estimates, the Delta Method, and Quantal Response Tests for Increased Carcinogenicity. *Biometrics,* **49,** 793-801.

Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8,** 59.

Boobis, A.R., Cohen, S.M., Dellarco, V.L., Doe, J.E., Fenner-Crisp, P.A., Moretto, A., Pastoor, T.P., Schoeny, R.S., Seed, J.G. and Wolf, D.C. (2016). Classification schemes for carcinogenicity based on hazard-identification have become outmoded and serve neither science nor society. *Regul Toxicol Pharmacol,* **82,** 158-166.

Cochran, W.G. (1954). Some Methods for Strengthening the Common χ^2–Tests. *Biometrics,* **10,** 417-451.

Crump, K.S., Krewski, D. and Van Landingham, C. (1999). Estimates of the proportion of chemicals that were carcinogenic or anticarcinogenic in bioassays conducted by the National Toxicology Program. *Environ Health Perspect,* **107,** 83-88.

Dinse, G.E. and Peddada, S.D. (2011). Comparing tumor rates in current and historical control groups in rodent cancer bioassays. *Statistics in biopharmaceutical research,* **3,** 97-105.

ECHA. (2016). Practical Guide: How to use alternatives to animal testing to fulfil the information requirements for REACH registration. ECHA-16-B-25-EN, ECHA Publication No. 1831-6727 ECHA, Helsinki, Finland.

ECHA. (2017). Guidance on the Application of CLP Criteria. Guidance to Regulation (EC) No 1272/2008 on classification, labelling and packaging (CLP) of substances and mixtures. Publication No. ECHA-17-G-21-EN, Finland, Helsinki.

EFSA Scientific Committee. (2011). Statistical Significance and Biological Relevance. *EFSA Journal,* **9**.

EFSA Scientific Committee. (2017). Guidance on the use of the weight of evidence approach in scientific assessments. *EFSA Journal,* **15**.

Elmore, S.A. and Peddada, S.D. (2009). Points to consider on the statistical analysis of rodent cancer bioassay data when incorporating historical control data. *Toxicol Pathol,* **37,** 672-676.

EU (2008). Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006, L353.

EU (2011). COMMISSION REGULATION (EU) No 544/2011 of 10 June 2011 implementing Regulation (EC) No 1107/2009 of the European Parliament and of the Council as regards the data requirements for active substances., L 155.

FAO/WHO. (2014). Pesticide residues in food 2014. Evaluations Part II - Toxicological Retrieved 31 March 2017: http://apps.who.int/iris/bitstream/10665/164597/1/9789241665308_eng.pdf?ua=1.

Fears, T.R., Tarone, R.E. and Chu, K.C. (1977). False-positive and false-negative rates for carcinogenicity screens. *Cancer Res,* **37,** 1941-1945.

Festing, M.F. (2010). Inbred strains should replace outbred stocks in toxicology, safety testing, and drug development. *Toxicol Pathol,* **38,** 681-690.

Festing, M.F.W. (2016). Genetically Defined Strains in Drug Development and Toxicity Testing In *Mouse Models for Drug Discovery: Methods and Protocols* (Proetzel, G. & Wiles, M.V. eds.), pp. 1-17. Springer New York, New York, NY.

Gart, J.J., Krewski, P.N., Tarone, R.E. and Wahrendorf, J. (1986). STATISTICAL METHODS IN CANCER RESEARCH - VOLUME III - The design and analysis of long-term animal experiments, 40CFR Part 261., US Government Printing Office. Washington, DC.

Gaylor, D.W. (2005). Are tumor incidence rates from chronic bioassays telling us what we need to know about carcinogens? *Regul Toxicol Pharmacol,* **41,** 128-133.

Gibson-Corley, K.N., Olivier, A.K. and Meyerholz, D.K. (2013). Principles for valid histopathologic scoring in research. *Veterinary pathology,* **50,** 1007-1015.

Gold, L.S., Manley, N.B., Slone, T.H., Rohrbach, L. and Garfinkel, G.B. (2005). Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature through 1997 and by the National Toxicology Program in 1997-1998. *Toxicol Sci,* **85,** 747-808.

Gold, L.S., Slone, T.H. and Bernstein, L. (1989). Summary of carcinogenic potency and positivity for 492 rodent carcinogens in the carcinogenic potency database. *Environ Health Perspect,* **79,** 259-272.

Haseman, J.K. (1984). Statistical issues in the design, analysis and interpretation of animal carcinogenicity studies. *Environ Health Perspect,* **58,** 385-392.

Haseman, J.K. (1995). Data analysis: statistical analysis and use of historical control data. *Regul Toxicol Pharmacol,* **21,** 52-59; discussion 81-56.

Haseman, J.K. and Elwell, M.R. (1996). Evaluation of false positive and false negative outcomes in NTP long-term rodent carcinogenicity studies. *Risk Anal,* **16,** 813-820.

Haseman, J.K., Huff, J. and Boorman, G.A. (1984). Use of historical control data in carcinogenicity studies in rodents. *Toxicol Pathol,* **12,** 126-135.

Haseman, J.K. and Johnson, F.M. (1996). Analysis of National Toxicology Program rodent bioassay data for anticarcinogenic effects. *Mutat Res,* **350,** 131-141.

Haseman, J.K. and Lockhart, A. (1994). The relationship between use of the maximum tolerated dose and study sensitivity for detecting rodent carcinogenicity. *Fundamental and applied toxicology : official journal of the Society of Toxicology,* **22,** 382-391.

Haseman, J.K., Ney, E., Nyska, A. and Rao, G.N. (2003). Effect of diet and animal care/housing protocols on body weight, survival, tumor incidences, and nephropathy severity of F344 rats in chronic studies. *Toxicol Pathol,* **31,** 674-681.

Haseman, J.K., Young, E., Eustis, S.L. and Hailey, J.R. (1997). Body weight-tumor incidence correlations in long-term rodent carcinogenicity studies. *Toxicol Pathol,* **25,** 256-263.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Statist.,* **6,** 65-70.

IARC. (2006). IARC Monographs on the Evalaution of Carcinogenic Risks to Humans. PREAMBLE. Retrieved 22.08.2018 2017, from WHO: http://monographs.iarc.fr/ENG/Preamble/index.php.

ICH. (1994). Pharmacokinetics: Guidance for Repeated Dose Tissue Distribution Studies. S3B, Publication No. S3B.

ICH. (2017). Questions and Answers to ICH S3A: Note for Guidance on Toxicokinetics: The Assessment of Systemic Exposure in Toxicity Studies – Focus on Microsampling. S3A Q&As, Publication No. S3A Q&As.

Johnson, F.M. (2002). How many food additives are rodent carcinogens? *Environmental and molecular mutagenesis,* **39,** 69-80.

Johnson, F.M. (2003). How many high production chemicals are rodent carcinogens? Why should we care? What do we need to do about it? *Mutat Res,* **543,** 201-215.

Keenan, C., Elmore, S., Francke-Carroll, S., Kemp, R., Kerlin, R., Peddada, S., Pletcher, J., Rinke, M., Schmidt, S.P., Taylor, I. and Wolf, D.C. (2009). Best practices for use of historical control data of proliferative rodent lesions. *Toxicol Pathol,* **37,** 679-693.

Keenan, K.P., Ballam, G.C., Soper, K.A., Laroque, P., Coleman, J.B. and Dixit, R. (1999). Diet, caloric restriction, and the rodent bioassay. *Toxicol Sci,* **52,** 24-34.

Keenan, K.P., Laroque, P., Ballam, G.C., Soper, K.A., Dixit, R., Mattson, B.A., Adams, S.P. and Coleman, J.B. (1996). The effects of diet, ad libitum overfeeding, and moderate dietary restriction on the rodent bioassay: the uncontrolled variable in safety assessment. *Toxicol Pathol,* **24,** 757-768.

Lin, K.K. and Rahman, M.A. (1998). Overall false positive rates in tests for linear trend in tumor incidence in animal carcinogenicity studies of new drugs. *Journal of biopharmaceutical statistics,* **8,** 1-15; discussion 17-22.

NTP. (2012). Toxicology and Carcinogenesis Studies of Acrylamide (CAS No. 79-06-1).

OECD. (1994). Data Requirements for Pesticide Registration in OECD Member Countries: Survey Results. Publication No. OCDE/GD(94)47.

OECD. (2002). Guidance Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies. JT00130828, Publication No. ENV/JM/MONO(2002)19.

OECD. (2010). Test No. 417: Toxicokinetics. OECD Publishing.

OECD. (2015). Guidance Document 116 on the Conduct and Design of Chronic Toxicity and Carcinogenicity Studies, Supporting Test Guidelines 451, 452 and 453. JT03319769, Publication No. ENV/JM/MONO(2011)47.

Peddada, S.D., Dinse, G.E. and Kissling, G.E. (2007). Incorporating Historical Control Data When Comparing Tumor Incidence Rates. *Journal of the American Statistical Association,* **102,** 1212-1220.

Peddada, S.D. and Kissling, G.E. (2006). A survival-adjusted quantal-response test for analysis of tumor incidence rates in animal carcinogenicity studies. *Environ Health Perspect,* **114,** 537-541.

Peto, R., Pike, M.C., Day, N.E., Gray, R.G., Lee, P.N., Parish, S., Peto, J., Richards, S. and Wahrendorf, J. (1980). Guidelines for simple, sensitive signifcance tests for carcinogenic effects in long-term animal experiments. In: Long-term and Short-term Screening Assays for Carcinogens: A Critical Appraisal. International Agency for Research on Cancer Publication No. Supplement 2, Lyon.

Portier, C.J. and Bailer, A.J. (1989). Testing for increased carcinogenicity using a survival-adjusted quantal response test. *Fundamental and applied toxicology : official journal of the Society of Toxicology,* **12,** 731-737.

Portier, C.J., Hedges, J.C. and Hoel, D.G. (1986). Age-specific models of mortality and tumor onset for historical control animals in the National Toxicology Program's carcinogenicity experiments. *Cancer Res,* **46,** 4372-4378.

Seilkop, S.K. (1995). The effect of body weight on tumor incidence and carcinogenicity testing in B6C3F1 mice and F344 rats. *Fundamental and applied toxicology : official journal of the Society of Toxicology,* **24,** 247-259.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association,* **62,** 8.

Slikker, W., Jr., Andersen, M.E., Bogdanffy, M.S., Bus, J.S., Cohen, S.D., Conolly, R.B., David, R.M., Doerrer, N.G., Dorman, D.C., Gaylor, D.W., Hattis, D., Rogers, J.M., Setzer, R.W., Swenberg, J.A. and Wallace, K. (2004a). Dose-dependent transitions in mechanisms of toxicity: case studies. *Toxicol. Appl. Pharmacol,* **201,** 226-294.

Slikker, W., Jr., Andersen, M.E., Bogdanffy, M.S., Bus, J.S., Cohen, S.D., Conolly, R.B., David, R.M., Doerrer, N.G., Dorman, D.C., Gaylor, D.W., Hattis, D., Rogers, J.M., Woodrow, S.R., Swenberg, J.A. and Wallace, K. (2004b). Dose-dependent transitions in mechanisms of toxicity. *Toxicol. Appl. Pharmacol,* **201,** 203-225.

STP. (2002). Statistical methods for carcinogenicity studies. *Toxicol Pathol,* **30,** 403-414.

Sun, J. (1999). On the Use of Historical Control Data for Trend Test in Carcinogenicity Studies. *Biometrics,* **55,** 1273-1276.

Tarone, R.E. (1982). The use of historical control information in testing for a trend in Poisson means. *Biometrics,* **38,** 457-462.

Tarone, R.E., Chu, K.C. and Ward, J.M. (1981). Variability in the rates of some common naturally occurring tumors in Fischer 344 rats and (C57BL/6N x C3H/HeN)F1 (B6C3F1) mice. *J Natl Cancer Inst,* **66,** 1175-1181.

Tennekes, H., Gembardt, C., Dammann, M. and van Ravenzwaay, B. (2004a). The stability of historical control data for common neoplasms in laboratory rats: adrenal gland (medulla), mammary gland, liver, endocrine pancreas, and pituitary gland. *Regul Toxicol Pharmacol,* **40,** 18-27.

Tennekes, H., Kaufmann, W., Dammann, M. and van Ravenzwaay, B. (2004b). The stability of historical control data for common neoplasms in laboratory rats and the implications for carcinogenic risk assessment. *Regul Toxicol Pharmacol,* **40,** 293-304.

USEPA. (2005). Guidelines for Carcinogen Risk Assessment. Publication No. EPA/630/P-03/001F.

USEPA. (2017). Categorical Regression Analysis (CatReg). Retrieved 06.20.18: https://www.epa.gov/bmds/catreg.

Wasserstein, R.L., Schirm, A.L. and Lazar, N.A. (2019). Moving to a World Beyond “p < 0.05”. *The American Statistician,* **73,** 1-19.

WHO Core Assessment Group on Pesticide Residues. (2015). JMPR Guidance Document for WHO monographers and reviewers. Publication No. WHO/HSE/FOS/2015.1.