JP Journal of Biostatistics

Published Online: August 2014
Available online at http://pphmj.com/journals/jpb.htm
Volume 11, Number 1, 2014, Pages 1-32

STATISTICAL EVALUATION OF TOXICOLOGICAL ASSAYS WITH ZERO OR NEAR-TO-ZERO PROPORTIONS OR COUNTS IN THE CONCURRENT NEGATIVE CONTROL GROUP: A TUTORIAL

Thomas Jaki[1], Andreas Kitsche[2] and Ludwig A. Hothorn[2]

[1]Medical and Pharmaceutical Statistics Research Unit
Lancaster University
UK
e-mail: jaki.thomas@gmail.com

[2]Institut für Biostatistik
Leibniz Universität Hannover
Herrenhäuser Str. 2, D-30419 Hannover
Germany

Abstract

In toxicological studies interest often lies in proportions or counts for which an increase in the dose group over control indicates a safety risk. Additionally, the control group observes values that are zero or near-to-zero for endpoints characterizing pathological processes. In such instances, the comparison of dose groups versus control requires special attention as inference for ratio-to-control is infeasible or unstable and inference for difference-to-control is highly sensitive to the number of zeros or near-to-zero values. In practice, assays are commonly performed multiple times in a laboratory so that data of some historical controls are available. When the concurrent control values fall within a corresponding normal range, the evaluation is

Received: November 25, 2013; Revised: January 24, 2014; Accepted: February 3, 2014
Keywords and phrases: count, proportion, near-to-zero, historical controls, poly-3 test, Williams trend test.

performed by comparing doses versus the concurrent control. If the data of the concurrent control are outside the normal range, a test versus the concurrent control has either an increased risk of a false-positive result or an increased risk of a false-negative result, depending on the direction of the deviation. In this work, we discuss a simple to use Williams-type approach for comparing against a mean historical value. The idea is illustrated on three examples and we show how the method can be implemented the statistical software package R.

1. Introduction

Methodologically, two different types of endpoints can be distinguished in toxicological studies: (i) outcomes of a general physiological process, e.g., the hemoglobin content, and (ii) outcomes of a specific pathological process, e.g., the number of micronuclei. The former are often continuous variables for which adverse reactions tend to cause either increasing or decreasing values (e.g., enlargement or inhibition of liver weight). The latter endpoints are often proportions or counts for which the direction of pathological counts or proportions is inherently increasing. A consequence of these directional changes is that one-sided tests are appropriate to investigate if a compound impacts the endpoint over a reference value or control.

For pathological outcome variables, the focus of this manuscript, the data in the untreated control group are often zero or near-to-zero, i.e., most of the randomized units (e.g., animals) reveal the value zero or only some small value, such as 1 or 2. In such instances, the common comparison of dose groups versus control requires special attention, as:

(i) inference for ratio-to-control, e.g., recommended for a *k*-fold interpretation [1], is either infeasible or unstable;

(ii) inference for difference-to-control is highly sensitive to the number of zeros or near-to-zero values, i.e., observing 0 tumors in 50 control animals versus 1 tumor in 50 controls changes the *p*-value of the test for treatment effect notably; and

(iii) the power of tests for the difference of proportions is largest when the proportion on control is zero since this minimizes the estimator of the common variance.

In practice, such assays are commonly performed multiple times in a laboratory and therefore data of some historical controls are typically available [2]. When the values of the concurrent control falls within a related normal range, the evaluation is performed by comparing doses versus the concurrent control. If the data of the concurrent control are, however, small compared to the normal range, a test versus the concurrent control has a false-positive tendency (i.e., an increased probability of wrongly claiming a change in the endpoint over control). Analogously, a false negative tendency occurs when the concurrent control data are large compared to the normal range. For example, the poly-3-trend test on histiocytic sarcomas in female rats in the long-term bioassay of benzophenone is not significant $(p = 0.074)$, but highly significant when taking the control data from six historical studies into account $(p = 0.004)$ [3].

It is therefore recommended to use tests that allow historical information to be incorporated when evaluating pathological outcome variables. For both, proportions and counts, modifications of the Cochran-Armitage-trend test have been proposed [4, 5], yet these approaches do not seem to be implemented routinely for the evaluation of toxicological assays. Reasons may include:

(i) a focus on current tumor evaluation in long-term carcinogenicity bioassays,

(ii) a complex approach that may not be transparent to toxicologists, published mainly in statistical journals (e.g., [3, 6-12]),

(iii) the US National Toxicological Program [13] recommends Dunnett and Williams-type approaches [14], while the Cochran-Armitage-trend test is only sensitive for near-to-linear shapes and not for any shape (see, e.g., [15]); and

(iv) asymptotic solutions are inappropriate for most toxicological assays which tend to have rather small sample sizes.

Recently, [16] proposed a two-step approach for the routine evaluation of the Ames fluctuation test which "sorts out" significant Williams-test results when the concurrent control is below a threshold in order to reduce the number of false positives. Wolf et al. [17] on the other hand proposed a method for the HET-MN assay that uses a threshold to decide whether a test versus all historical raw data should be used. Both of these approaches follow the US-FDA recommendation: *The concurrent control group is always the most appropriate and important in testing drug related increases in tumor rates... historical control data can be very valuable in the final interpretation of the study results* [18]. A potential problem for both approaches is, however, that they may be biased. The former may sort out real positive trends while the latter depends on the sample size of the historical controls.

Hayashi et al. [2] discussed the difficulties with the statistical analysis of rodent cancer bioassays using historical control data directly. In this paper, we propose and discuss, based on three examples, a simplified approach which starts by evaluating if the concurrent control is within a normal range. If it is, the concurrent control is used for comparisons while the arithmetic mean of the historical assays is used as a standard value if it is not. The actual decision for or against a trend is performed with a Williams test, modified for comparison against a fixed default value instead of an estimated mean value of the current control. The advantage of this approach is that its application is relatively simple and will result in smaller false decision rates than the two ideas discussed above, namely the Williams-test with only the concurrent control data and to model the historical and concurrent control data (see the simulation study in the Appendix for details). Note that other approaches which may result in even better error rates, for example based on Bayesian ideas [19] or mixture distributions (e.g., zero-inflated Poisson models or hurdle models, [20]), are also available. These are not discussed here as they tend to be rather complex and hence not suitable for day-to-day analysis. Here we describe a Williams-type procedure for the analysis

of studies including several dose groups and a zero-dose control to claim a dose-dependent trend when rejecting the null-hypothesis. To support implementation of the ideas discussed we also present code for the statistical software R [21] in the Appendix of this work. An evaluation of the statistical properties and comparison against alternative approaches of the method is also provided in the Appendix.

2. Three Examples

In this section, we will introduce three examples which will be used to illustrate the usefulness of the new approach. In the first example, the number of micronucleated erythrocytes measured by the hen's egg genotoxicity assay for micronucleus induction [HET-MN, 17] is the primary endpoint. This endpoint can be considered as count data, as the number of scored polychromatic and normochromatic erythrocytes is constant. Commonly a one-way layout with three or four doses of the test compound, a negative control (NC), and a positive control, with six eggs randomly assigned to each group is used.

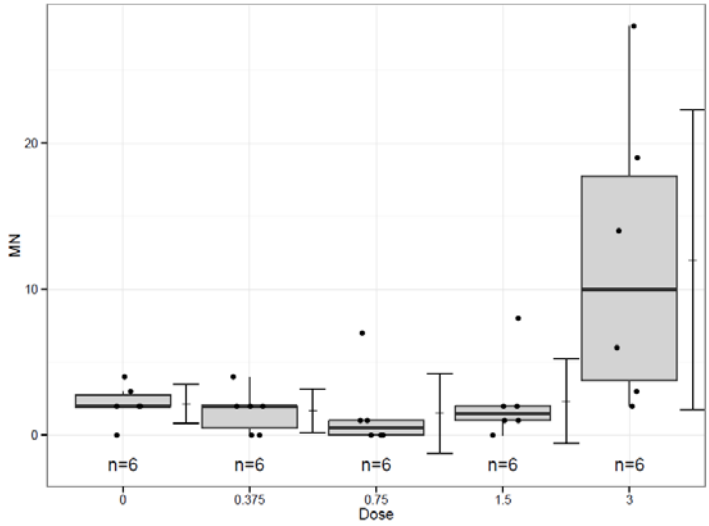

Figure 1. Boxplot of HET-MN assay data.

The boxplot in Figure 1 shows the data for a selected assay without the positive control (see `MNassay` in the Appendix for full data). In this

example, a relatively large numbers of micronucleated erythrocytes (MN's) in the control $(\bar{x}_{NC} = 2.17)$, low numbers in the low doses and markedly increased number of MN's in the highest dose can be seen. In the same laboratory, 24 historical assays are available where the mean on the negative control at 1.77 is much smaller than the mean of the concurrent control [22] (see `histMN` in the Appendix). The question therefore arises, whether the *p*-values of a common Williams-type trend test for the comparison against the concurrent control (on the Freeman-Tukey transformed endpoint) are too large, due to an unusually large number of MN's in the concurrent control, and hence wrongly indicating no toxic effect. For the above example, we selected a real data example with only one of six MN-count being zero in the control for illustration. Figure 2 (from Figure 5 in [22], see `MNassayZeroControl` in the Appendix for full data) gives an example with only zero values in the concurrent control group.

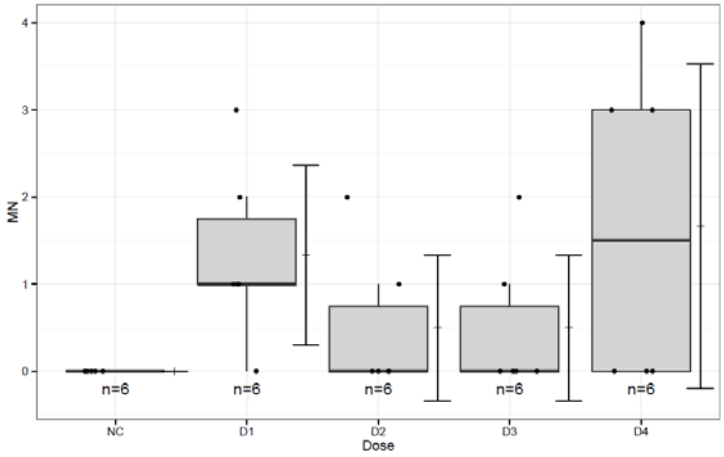

Figure 2. Boxplot of HET-MN with zero control data.

In the second example, alveolar-bronchial adenomas in a 102 weeks carcinogenicity bioassay treated with Pivalolactone are considered. These adenomas were analysed as (mortality-independent) crude tumors and historical control rates were available for 23 related bioassays [8]. The boxplot in Figure 3 shows a descriptive summary of the historical tumor rates as well as the tumor rates in the current study on different doses (stars).

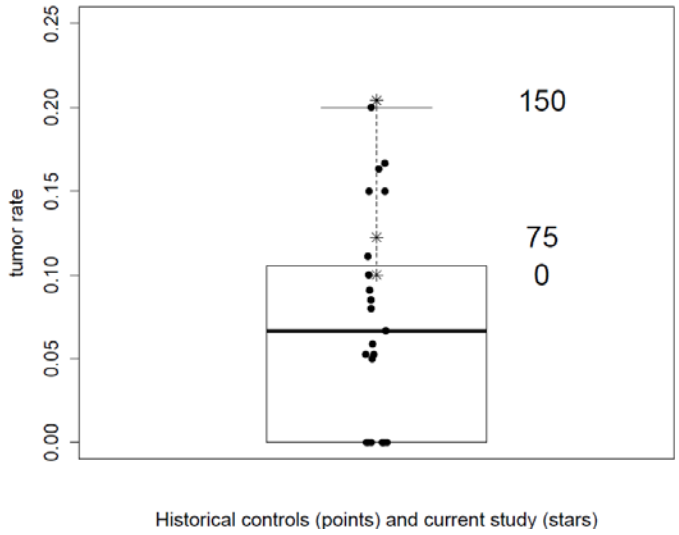

Figure 3. Boxplot of historical alveolar-bronchial adenomas data. Black points are the observed historical data and stars correspond to the concurrent data at doses 0, 75 and 150mg/kg.

Once more the concurrent control tumor rate is larger than the historical tumor rate, leaving the question whether the higher tumor rate of the concurrent control $(\hat{p}_{NC} = 0.1)$ compared to the historical tumor rate $(\hat{p}_{HC} = 0.07)$ causes a serious increase of the *p*-value of the common Williams-type test for comparison against the concurrent control.

The third example is focused on the mortality-adjusted analysis of tumors. The analysis of crude tumor rates is biased when mortality is different in the groups. Therefore, the poly-3 adjusted proportions are used for mortality adjustment when the cause of death is unknown. Here the individual data of tumor status (0=no tumor, 1=tumor) and day of death/sacrifice for the adenoma in Harderian glands in male mice of NTP long-term bioassay study C61621D are used (Table 1). The unexpected larger number of 4 tumors out of 50 animals in the concurrent control for this rare tumor leads to questions whether this may cause false negative decisions (i.e., a too large *p*-value) to occur. Historical control data from 13 studies including both, tumor status and day of death/scarifies, were collected from

the NTP database which suggest that 4 in 50 animals is untypically large (Table 2).

Table 1. Harderian gland tumors of current NTP study No. C61621D (Chemical Name: Triethanolamine; CASRN: 102-71-6)

Dose group	No. of animals	No. of tumors
0mg/kg	46	4
200mg/kg	45	5
630mg/kg	44	6
2000mg/kg	49	1

Table 2. Harderian gland tumor data from 13 historical and one concurrent (*it*) studies. Studies of 2 years length exposure rate of topical application. Animals represent male mice of strain B6C3F1. (CASRN: Chemical Abstracts Service Registry Number)

Chemical Name	CASRN	Study No.	No. of animals	No. without tumor	No. of tumors
Benzethonium chloride	121-54-0	C61494B	50	48	2
bis(2-Chloroethoxy)methane	111-91-1	C99028	50	50	0
Coconut oil acid diethanolamine condensate	68603-42-9	C55312B	50	48	2
1,2-Dibromo-2,4-dicyanobutane	35691-65-7	C97003D	50	43	7
Diesel fuel marine	DIESELFUEL	C54795B	49	46	3
Diethanolamine	111-42-2	C55174D	50	50	0
Diethyl phthalate	84-66-2	C60048B	50	49	1
Diisopropylcarbodiimide	693-13-0	C93020D	50	47	3
Lauric acid diethanolamine condensate	120-40-1	C55323B	50	49	1
Methyl trans-styryl ketone	1896-62-4	C95003C	50	50	0
Oleic acid diethanolamine condensate	93-83-4	C91014	49	47	2
Sodium xylenesulfonate	1300-72-7	C55403D	50	49	1
Triethanolamine	102-71-6	C61621D	50	46	4
4-Vinyl-1-cyclohexene diepoxide	106-87-6	C60139A	50	50	0

3. Williams-type Procedures for the Comparison Against a Standard Value

In this section, we will discuss straightforward statistical approaches that are applicable to the three examples introduced above. We will first introduce the methodology in generic terms and then illustrate their application in the subsequent section.

The main assumption we will use is that the mean value of the controls of historical assays is approximating the true value, denoted as standard, of the particular endpoint (e.g., the tumor rate or number of MN). The standard represents a constant value and hence will be treated to have no variance and is independent of sample size. From the Bayesian viewpoint, this assumption is rather limiting, but does simplify the problem substantially while making the approach numerically feasible and independent of the number of available assays. Therefore, multiple comparison procedures versus a standard are proposed. A modified Dunnett procedure for comparison against a standard was previously described [23] and related simultaneous prediction intervals have also been discussed [24].

3.1. Estimation of the threshold

Since a threshold is routinely used to decide whether an assay is evaluated with the concurrent control or with historical controls, the estimation of said threshold should be clear and transparent and making this decision is typically the first step of the analysis. It starts by defining a normal range as a prediction interval for k values in the concurrent control based on ζ_r values in R historical assays. Parametric approaches depend heavily on the underlying normal distribution while non-parametric ideas exhibit problems with tied data (e.g., when using counts). Moreover, the related phase II quality control chart approaches are complicated [25].

In bio-medical research, 2σ intervals are common where 2 is approximating the two-sided 95% quantile of the standard normal distribution. These intervals are also recommended by quality control

arguments [26]. The general form of these intervals is $\hat{\theta} \pm 2SD$, where $\hat{\theta}$ is the mean value of the historical controls and the SD is the standard deviation of the estimate used in the concurrent control data (e.g., the standard deviation of the mean of the negative controls). For count data, for example, the interval is constructed for Freeman-Tukey transformed variables and has been recommended based on a tradeoff between simplicity and validity [27]. The intervals are of the form: $\bar{x}_{HC} \pm 2SD_{HC}/\sqrt{n_{CC}}$, where $\bar{x}_{HC}$ and SD_{HC} are the mean transformed counts and estimated standard deviation of the historical control (HC) data, respectively, and n_{CC} is the sample size of the concurrent control. For tumor rates, the corresponding interval is $\hat{p}_{HC} \pm 2\sqrt{\frac{\hat{p}_{HC}(1-\hat{p}_{HC})}{n_{CC}}}$, where $\hat{p}_{HC}$ is the estimated tumor based on historical data and n_{CC} is the number of observations in the concurrent control group.

Besides the 2σ intervals described above, alternatives, such as using the difference between the maximum and the minimum of the historical control rates or the inter-quartile range (IQR) have been discussed [28].

3.2. Count data transformed to approximate normal distributed endpoints

Generalized linear models can be used to draw inference on count data. In particular, the variation between animals (or other experimental units) can be modeled by generalized Poisson distribution models, such as the quasi Poisson or negative binomial model [29]. The commonly used log-link function, however, causes unstable or infeasible ratio-to-control (odds ratio) estimates when zero or near-to-zero control data occur. A simple alternative is to transform the data and subsequently use common parametric tests on the transformed variable. For toxicological count data, such transformation approaches were used for the Cell Transformation Assay [30, 31], the Ames Fluctuation Assay [16], and the *in vitro* Chromosome Aberration Assay [32]. A common challenge encountered when using count data is a dose-dependent

increase in the variance so that the assumption of variance homogeneity is violated. According to Guan [33], the simple Freeman-Tukey root transformation [34], defined as $x_{ij}^{FT} = \sqrt{\kappa_{ij}} + \sqrt{\kappa_{ij}+1}$, where κ_{ij} is the count of animal j in treatment group $i = 0, 1, ..., k$ (0=concurrent control), can be recommended for transforming count data into approximate normal distributed variables which satisfy variance homogeneity approximately.

A Williams-type procedure

A Williams-type approach for comparing counts against the standard, ϑ, defined as the mean count in the historical control assays, can be used. The test statistic is defined as

$$t_{\text{Contrast}}^{\text{vs. Standard, normal distr.}} = \left(\sum_{i=1}^{k} c_{li}\bar{x}_i - \vartheta \right) / S_{i=1,\ldots,k} \sqrt{\sum_{i=1}^{k} c_{li}^2 / n_i}\,, \qquad (1)$$

where $\bar{x}_i$ and n_i are the mean transformed counts and the sample size in group i, respectively, and $S_{i=1,\ldots,k}$ is the pooled standard deviation of the transformed values. The contrast coefficients c_{li} of the standard Williams procedure [35] are modified for a k by k matrix (which does not include 0 for the concurrent control) as:

$$C_{\text{Williams}}^{\text{vs. Standard}} = \begin{pmatrix} 0 & \cdots & 0 & 0 & 1 \\ 0 & \cdots & 0 & \frac{n_{k-1}}{n_{k-1}+n_k} & \frac{n_k}{n_{k-1}+n_k} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \frac{n_1}{n_1+\cdots+n_k} & \cdots & \frac{n_{k-2}}{n_1+\cdots+n_k} & \frac{n_{k-1}}{n_1+\cdots+n_k} & \frac{n_k}{n_1+\cdots+n_k} \end{pmatrix}. \qquad (2)$$

This approach can be easily implemented using the argument `rhs` in the function `glht` of the R package `multcomp`. Notice, a simple Welch-type modification for the common occurring heteroscedasticity is available [36].

In the subsequent section, we will illustrate how adjusted p-values and simultaneous confidence intervals can be estimated for this Williams-type procedure.

3.3. Crude proportions

Proportions frequently occur in toxicology as pathological outcomes, such as mortality, crude tumor rate or incidences of histopathological findings. Traditionally cross-table analysis methods, such as chi-squared test, are used in this context. Small sample size observed in toxicological studies, however, leads to liberal results for table-based analysis methods when asymptotic approaches are used, while exact approaches tend to be conservative. Agresti and Caffo [37] proposed to overcome this issue by adding two pseudo success and two pseudo failure when computing the sample proportions. They showed that the corresponding Wald confidence intervals have coverage close to nominal level for small to medium sample sizes. This idea based on two-sample comparisons was subsequently extended to a one-way layout with $k > 2$ groups [38] and the Williams-type procedure [25]. Simulations for the typical design in toxicological studies with 3 to 4 groups and sample sizes below 20 indicate that a "Add1" approach is suitable for one-sided confidence intervals [25].

For $r = 1, ..., R$ historical control assays with Z_r events and m_r animals and a current assay with Y_i events and n_i animals ($i = 1, ..., k$ doses), the Williams-type approach can be defined using the following test statistic:

$$t_{\text{Contrast}}^{\text{vs. Standard, Proportions, Add1}} = \left(\sum_{i=1}^{k} c_{li} \hat{p}_i - \psi \right) \Big/ \sqrt{\sum_{i=1}^{k} c_{li}^2 \hat{V}ar(\hat{p}_i)}$$

with

$$\hat{p}_i = (Y_i + 0.5)/(n_i + 1),$$

$$\hat{V}ar(\hat{p}_i) = \hat{p}_i(1 - \hat{p}_i)/(n_i + 1),$$

$$\hat{q}_r = (Z_r + 0.5)/(m_r + 1),$$

$$\psi = \frac{1}{R} \sum_{r=1}^{R} \hat{q}_r.$$

This approach can be implemented using the R package `MCPAN` and both, adjusted *p*-values and simultaneous confidence intervals, can be found.

A Monte Carlo simulation study, available in the Appendix, shows that this approach is advantageous in situations where the expected value of the historical control data is smaller in comparison to the concurrent control group.

3.4. Poly-3-estimates

The primary endpoints in long-term carcinogenicity bioassay are tumor incidences. Due to a strong interaction between tumor formation and mortality, a simple evaluation of crude tumor proportions cannot be recommended. Instead the analysis of mortality-adjusted tumor rates by the poly-3-approach can be recommended [39, 40], in particular since cause of death information is rarely available. The idea is to use individual mortality-specific weights, $w_{ij} = (t_{ij}/t_{\max})^3$ (where t_{ij} is the time when a particular tumor in animal i of treatment group j was found), is particularly appealing due to its simplicity. The weight takes the value $w_{ij} = 1$, if an animal survives until the final sacrifice or dies during the study with a tumor of interest. The analysis of crude proportions can be easily replaced by poly-3 adjusted proportions for both concurrent bioassay and the historical control data by using adjusted proportions, $p_i^* = Y_i/n_i^*$ and adjusted sample sizes (animals at risk) $n_i^* = \sum_{j=1}^{n_i} w_{ij}$. A Williams-type procedure for the comparisons against the concurrent control is discussed in [41]. Analogously, this approach can be extended to a Williams-type procedure for the comparison against a standard, i.e., against the mean of the historical poly-3-adjusted proportions. This approach is much simpler than a recent proposal which requires resampling [6].

$$t_{\text{Contrast}}^{\text{vs. Standard, Poly3}} = \left(\sum_{i=1}^{k} c_{li} p_i^* - \xi \right) \Big/ \sqrt{\sum_{i=1}^{k} c_{li}^2 / \hat{V}ar(p_i^*)}$$

with

$$p_i^* = Y_i / n_i^*,$$

$$q_r^* = Z_r^* / m_r^*,$$

$$\xi = \frac{1}{R} \sum_{r=1}^{R} q_r^*.$$

4. Evaluation of the Examples

In this section, the three examples previously introduced are evaluated by the new Williams-type procedures for comparing against standard using the R packages `multcomp` and `MCPAN`. The process used to analyse the data is discussed here while step-by-step instructions using R are given in the Appendix. Furthermore, for direct comparison, the *p*-values or confidence limits of the standard Williams-type approach for comparisons against the concurrent control are reported to demonstrate the advantage of the new approach.

4.1. Evaluation of the HET-MN example

We start by transforming the count data (both concurrent and historical) using a Freeman-Tukey transformation to obtain an approximately normally distributed variable whose variance is homogeneous. The mean of the transformed concurrent control is 3.07 and outside the normal range of the transformed historical controls [1.05, 2.50]. Therefore, the Williams-type procedure against the mean of the historical controls is recommended to avoid a false negative decision, i.e., too large *p*-values.

Table 3. *P*-values of Williams-type contrasts against concurrent and standard control

Comparison	Concurrent	Standard
C vs. D_3	0.0038	< 0.0001
C vs. $\frac{D_3 + D_{1.5}}{2}$ 2	0.0741	0.0000
C vs. $\frac{D_3 + D_{1.5} + D_{0.75}}{3}$	0.2680	0.0003
C vs. $\frac{D_3 + D_{1.5} + D_{0.75} + D_{0.375}}{3}$	0.3942	0.0004

Table 3 provides the *p*-values for the trend contrasts as defined in equation (2) against both, the concurrent and the standard. A significant trend for the first contrast which compares the highest doses against control can be seen irrespective of using concurrent controls or a standard value. The *p*-value for the comparison against the historical control is, however, much smaller. Moreover, all contrasts are significant when comparing against the historical control whereas only the first contrast is significant when comparing against concurrent control, suggesting that a safety issue may be present that is undetected if only the concurrent control data are used.

4.2. Evaluation of crude tumor rate

We begin by estimating the proportion of tumors for both concurrent and historical data using the Add1 approach. The spontaneous rate of the concurrent control is with 0.12 inside the 2σ interval [–0.04, 0.23], so that the Williams-type procedure against the concurrent control is indicated. Note that the simple range of Add1 adjusted proportions is with [0.02 0.21] very similar. Both approaches using historical controls, namely the logistic model [8] and the Williams-type approach [19], show a significant increase in tumor rates (see Table 4), whereas the comparison against the concurrent control (which is the correct approach in this case) is not significant.

Table 4. Lower confidence limits for Williams-type approaches

Comparison	Concurrent	Standard	Logistic model
C vs. D_{150}	–0.0721	0.0076	0.10
C vs. $\frac{D_{150} + D_{75}}{2}$	–0.0920	0.0052	0.10

4.3. Evaluation of poly-3 tumor rates

First the poly-3 estimates for both concurrent and historical data are calculated from the raw data that contain the tumour status and the days on study for each individual. In this specific example, the mortality in animals with Harderian gland adenomas is not substantial so that the crude proportions and the poly-3 rates are quite similar. Notice, however, that this is not typically the case for other tumors and studies.

Table 5. Crude proportions and poly-3 estimates for historical studies

Study No.	No. Harderian Gl. Adenoma	No. Animals	Proportion	Poly3Proportion
C61494B	2	50	0.040	0.042
C99028	0	50	0.000	0.000
C55312B	2	50	0.040	0.043
C97003D	7	50	0.140	0.152
C54795B	3	49	0.061	0.083
C55174D	0	50	0.000	0.000
C60048B	1	50	0.020	0.021
C93020D	3	50	0.060	0.065
C55323B	1	50	0.020	0.022
C95003C	0	50	0.000	0.000
C91014	2	49	0.041	0.044
C55403D	1	50	0.020	0.022
C60139A	0	50	0.000	0.000

Table 6. Crude proportions and poly-3 estimates in current study

Treatment	Harderian Gl. Adenoma	No. animals	Proportion	Poly3Proportion
0mg/kg	4	50	0.080	0.083
200mg/kg	5	50	0.100	0.107
630mg/kg	6	50	0.120	0.132
2000mg/kg	1	50	0.020	0.021

Although the poly-3 rate of the concurrent control is at 0.083 within the 2σ interval of the historical poly-3 rates, [–0.016, 0.092], both Williams type approaches are compared here for illustration. Note that the Williams-approach used includes an "umbrella protection" as a downturn effect at the highest dose may occur. For all possible peak doses, namely 2000, 630 or 200mg/kg, separate Williams contrast tests are estimated, adjusted against multiple contrasts and multiple peak doses [42].

Table 7. *P*-values for umbrella protected Williams-type contrasts using a concurrent control and historical data

Comparison	Concurrent	Standard
C vs. D_{2000}	0.988	0.992
C vs. $\frac{D_{2000} + D_{630}}{2}$	0.790	0.217
C vs. $\frac{D_{2000} + D_{630} + D_{200}}{3}$	0.711	0.059
C vs. D_{630}	0.405	0.088
C vs. $\frac{D_{630} + D_{200}}{2}$	0.438	0.024
C vs. D_{200}	0.576	0.187

From the results in Table 7, one can see no trend when using the concurrent control only. Considering the historical control data, a trend for doses up to 630mg/kg as a plateau shape can be found.

5. Summary

The endpoints of a pathological process in toxicology, such as the number of micronuclei or tumor rate, are frequently evaluated as counts or proportions and often are zero or near-to-zero in the concurrent control group. For such data using an inadequate method of analysis will result in an increased risk of an incorrect decision. In the case when concurrent control data have a smaller observed value than the historical values, an increased risk of a false positive decision exists, while the risk of a false negative decision is inflated if the concurrent values are larger than the historical values. Several statistical approaches for a weighted analysis of concurrent and historical controls are available, although these are rarely used in routine analysis. One of the reasons is that these methods constitute a black box for toxicologists, while another reason is that guidelines recommend the primary and sole comparison against the concurrent control, provided they are within the range of historical controls. Furthermore, no software is publicly available for these specialist methods. Finally, these weighted approaches require, somewhat difficult to obtain, conditions on the historical controls, e.g., a larger number of included bioassays with a certain between-assay variability. The contrasting approach of using a Wilcoxon test comparing against all individual historical values on the other hand depends directly on the number of historical bioassays and can be therefore not recommended.

A William-type trend test comparing against a standard, namely the mean of the historical bioassays, is proposed. Its advantages are

(i) using Williams-tests recommended by the US-NTP as a general testing strategy,

(ii) robustness against data conditions of the historical data, in particular only a few historical bioassays can also be used,

(iii) its application is conditional on the concurrent control data being out-side the normal range of the historical data,

(iv) applicable to counts, crude proportions and poly-3 rates,

(v) independent of the number of available historical studies n_{HC}

(vi) availability of public software (R),

(vii) easy interpretability.

Of course, these advantages come at the disadvantage that this approach ignores the between-assay variability in the historical data.

When the use of the concurrent control instead of the historical controls is appropriate, Williams-type procedure against the zero-dose-control for counts, proportions and poly-3 estimates are available [22, 25, 41].

Acknowledgement

This work was supported in part by the German Science Foundation grant DfG-HO1687 for the last author (LAH).

References

[1] N. F. Cariello and W. W. Piegorsch, The Ames test: the two-fold rule revisited, Mutat. Res.-Genet. Tox. 369 (1-2) (1996), 23-31.

[2] M. Hayashi, K. Dearfield, P. Kasper, D. Lovell, H. J. Martus and V. Thybaud, Compilation and use of genetic toxicity historical control data, Mutat. Res.-Genet. Tox. En. 723 (2) (2011), 87-90.

[3] S. D. Peddada, G. E. Dinse and G. E. Kissling, Incorporating historical control data when comparing tumor incidence rates, J. Amer. Statist. Assoc. 102(480) (2007), 1212-1220.

[4] R. Tarone, The use of historical control information in testing for a trend in Poisson means, Biometrics 38 (2) (1982), 457-462.

[5] R. Tarone, The use of historical control information in testing for a trend in proportions, Biometrics 38 (1982), 215-220.

[6] G. E. Dinse and S. D. Peddada, Comparing Tumor rates in current and historical control groups in rodent cancer bioassays, Stat. Biopharm. Res. 3(1) (2011), 97-105.

[7] J. G. Ibrahim and L. M. Ryan, Use of historical controls in time-adjusted trend tests for carcinogenicity, Biometrics 52(4) (1996), 1478-1485.

[8] R. T. Smythe, D. Krewski and D. Murdoch, The use of historical control information in modeling dose-response relationships in carcinogenesis, Statist. Probab. Lett. 4(2) (1986), 87-93.

[9] D. G. Chen, Incorporating historical control information into quantal bioassay with Bayesian approach, Comput. Statist. Data Anal. 54(6) (2010), 1646-1656.

[10] Y. P.Ma, J. H. Guo, N. Z. Shi and M. L. Tang, On the use of historical control information for trend test in carcinogenesis, Biometrics 58(4) (2002), 917-927.

[11] L. Ryan, Using historical controls in the analysis of developmental toxicity data, Biometrics 49(4) (1993), 1126-1135.

[12] D. Dunson and G. Dinse, Bayesian incidence analysis of animal tumorigenicity data, J. Roy. Stat. Soc. C-App. 50(2) (2001), 125-141.

[13] National Toxicology Program (U.S.), Testing information - Descriptions of NTP study types-Statistical procedures-Expanded overview, available from <http://ntp.niehs.nih.gov> (accessed on 19 Oct 2011) (July 2011).

[14] D. A. Williams, A test for differences between treatment means when several dose levels are compared with a zero dose control, Biometrics 27 (1) (1971), 103-117.

[15] R. Tarone and J. Gart, On the robustness of combined tests for trend in proportions, J. Amer. Statist. Assoc. 75(369) (1980), 110-116.

[16] G. Reifferscheid, H. M. Maes, B. Allner, J. Badurova, S. Belkin, K. Bluhm, F. Brauer, J. Bressling, S. Domeneghetti, T. Elad, S. Flueckiger-Isler, H. S. Grummt, R. Guertler, A. Hecht, M. B. Heringa, H. Hollert, S. Huber, M. Kramer, A. Magde burg, H. T. Ratte, R. Sauerborn-Klobucar, A. Sokolowski, P. Soldan, T. Smital, D. Stalter, P. Venier, C. Ziemann, J. Zipperle and S. Buchinger, International round robin study on the Ames fluctuation test, Environ. Mol. Mutagen. 53(3) (2012), 185-197.

[17] T. Wolf, C. Niehaus-Rolf, N. Banduhn, D. Eschrich, J. Scheel and N.-P. Luepke, The hen's egg test for micronucleus induction (HET-MN): novel analyses with a series of well-characterized substances support the further evaluation of the test system, Mutat. Res.-Gen. Tox. En. 650(2) (2008), 150-164.

[18] Center for Drug Evaluation and Research, Guidance for industry: statistical aspects of the design, analysis, and interpretation of chronic rodent carcinogenicity studies of pharmaceuticals, Tech. rep., US Food and Drug Administration, 2001.

[19] A. Kitsche, L. A. Hothorn and F. Schaarschmidt, The use of historical controls in estimation simultaneous confidence intervals for comparisons against a concurrent control, Comput. Statist. Data Anal. 56(12) (2012), 3865-3875.

[20] T. Loeys, B. Moerkerke, O. De Smet and A. Buysse, The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression, British J. Math. Statist. Psych. 65 (2012), 163-180.

[21] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org/

[22] L. A. Hothorn, K. Reisinger, T. Wolf, A. Poth, D. Fieblinger, M. Liebsch and R. Pirow, Statistical analysis of the hens egg test for micronucleus induction (HET-MN assay), Mutation Research 757 (2013), 68-78.

[23] D. Rasch, G. Herrendorfer, J. Bock, N. Victor and V. Guiard, Verfahrensbibliothek -Versuchsplanung und Auswertung, Oldenbourg, 1996.

[24] S. Cheung, K. Wu and S. Lim, Simultaneous prediction intervals for multiple comparisons with a standard, Statist. Papers 43(3) (2002), 337-347.

[25] L. A. Hothorn, M. Sill and F. Schaarschmidt, Evaluation of incidence rates in pre-clinical studies using a Williams-type procedure, Int. J. Biostat. 6(1) (2010), Article 15.

[26] L. S. Nelson, When should the limits on a Shewhart control chart be other than a center line +/- 3-sigma?, J. Qual. Technol. 35(4) (2003), 424-425.

[27] S. Aebtarm and N. Bouguila, An empirical evaluation of attribute control charts for monitoring defects, Expert Sys. Appl. 38(6) (2011), 7869-7880.

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

[29] L. A. Hothorn and D. Gerhard, Statistical evaluation of the in vivo micronucleus assay, Arch. Toxicol. 83(6) (2009), 625-634.

[30] S. Hoffmann, L. A. Hothorn, L. Edler, A. Kleensang, M. Suzuki, P. Phrakonkham and D. Gerhard, Two new approaches to improve the analysis of balb/c 3t3 cell transformation assay data, Mutat. Res.-Gen. Tox. En. 744 (1) (2012), 36-41.

[31] H. Nishiyama, T. Omori and I. Yoshimura, A composite statistical procedure for evaluating genotoxicity using cell transformation assay data, Environmetrics 14(2) (2003), 183-192.

[32] L. A. Hothorn, A robust statistical procedure for evaluating genotoxicity data, Environmetrics 15(6) (2004), 635-641.

[33] Y. Guan, Variance stabilizing transformations of Poisson, binomial and negative binomial distributions, Statist. Probab. Lett. 79(14) (2009), 1621-1629.

[34] M. F. Freeman and J. W. Tukey, Transformations related to the angular and the square root, Ann. Math. Stat. 21(4) (1950), 607-611.

[35] F. Bretz, An extension of the Williams trend test to general unbalanced linear models, Comput. Statist. Data Anal. 50(7) (2006) 1735-1748.

[36] M. Hasler and L. A. Hothorn, Multiple contrast tests in the presence of heteroscedas ticity, Biom. J. 50(5, SI) (2008), 793-800.

[37] A. Agresti and B. Caffo, Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures, Am. Stat. 54(4) (2000), 280-288.

[38] R. M. Price and D. G. Bonett, An improved confidence interval for a linear function of binomial proportions, Comput. Statist. Data Anal. 45(3) (2004), 449-456.

[39] A. Bailer and C. J. Portier, Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples, Biometrics 44(2) (1988), 417-431.

[40] G. S. Bieler and R. L. Williams, Ratio estimates, the delta method, and quantal response tests for increased carcinogenicity, Biometrics 49(3) (1993), 793-801.

[41] F. Schaarschmidt, M. Sill and L. A. Hothorn, Poly-k-trend tests for survival adjusted analysis of tumor rates formulated as approximate multiple contrast test, J. Biopharm. Statist. 18(5) (2008), 934-948.

[42] F. Bretz and L. Hothorn, Statistical analysis of monotone or non-monotone dose-response data from in vitro toxicological assays, ATLA-Altern. Lab. Anim. 31(Suppl. 1) (2003), 81-96.

Appendix A. Supplementary Material

R code

In this section, we provide simple R code used to analyse the examples. Comments are preceded by #.

The HET-MN example

```
##read in data
histMN <-
structure (list(MN = c(0, 0, 1, 2, 1, 2, 0, 1, 1, 0, 1, 0, 1, 0,
1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 3, 0, 1, 0, 0, 0, 2, 0, 2, 3, 0,
0, 0, 0, 3, 1, 0, 0, 0, 0, 3, 1, 0, 0, 1, 2, 0, 0, 0, 1, 1, 0,
0, 1, 1, 2, 0, 0, 0, 2, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0,
1, 2, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 2, 1, 1, 0, 0, 3, 1, 2, 0,
3, 0, 1, 1, 1, 0, 1, 1, 0, 0, 2, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1,
1, 0, 0, 1), Run = structure(c(6L, 6L, 6L, 6L, 6L, 6L, 10L, 10L,
10L, 10L, 10L, 10L, 10L, 10L, 10L, 10L, 13L, 13L, 13L,
13L, 13L, 13L, 16L, 16L, 16L, 16L, 16L, 16L, 20L, 20L, 20L, 20L,
20L, 20L, 22L, 22L, 22L, 22L, 22L, 22L, 23L, 23L, 23L, 23L, 23L,
23L, 1L, 1L, 1L, 2L, 2L, 2L, 2L, 3L, 3L, 3L, 3L, 3L, 3L, 4L,
4L, 4L, 4L, 4L, 5L, 5L, 5L, 5L, 5L, 5L, 7L, 7L, 7L, 7L, 7L, 7L,
8L, 8L, 8L, 8L, 8L, 8L, 9L, 9L, 9L, 9L, 9L, 9L, 11L, 11L, 11L,
11L, 11L, 11L, 12L, 12L, 12L, 12L, 12L, 12L, 14L, 14L, 14L, 14L,
14L, 14L, 15L, 15L, 15L, 15L, 15L, 15L, 17L, 17L, 17L, 17L, 17L,
17L, 18L, 18L, 18L, 18L, 18L, 18L, 19L, 19L, 19L, 19L, 19L, 19L,
21L, 21L, 21L, 21L, 21L, 21L, 24L, 24L, 24L, 24L, 24L, 24L), .Label = c("01",
"02", "03", "04", "05", "06", "06-AK", "06-KM", "06-SK", "07",
"07-KM", "07-SK", "08", "08-KM", "08-SK", "09", "09-AK", "09-KM",
"09-SK", "10", "10-KM", "11-1", "11-2", "11-SK"), class = "factor")),.Names = c("MN",
"Run"), class = "data.frame", row.names = c(1L, 2L, 3L, 4L, 5L,
6L, 31L, 32L, 33L, 34L, 35L, 36L, 61L, 62L, 63L, 64L, 65L, 66L,
91L, 92L, 93L, 94L, 95L, 96L, 121L, 122L, 123L, 124L, 125L, 126L,
151L, 152L, 153L, 154L, 155L, 156L, 181L, 182L, 183L, 184L, 185L,
186L, 211L, 212L, 213L, 214L, 215L, 216L, 241L, 242L, 243L, 263L,
264L, 265L, 266L, 284L, 285L, 286L, 287L, 288L, 289L, 314L, 315L,
316L, 317L, 318L, 341L, 342L, 343L, 344L, 345L, 346L, 370L, 371L,
372L, 373L, 374L, 375L, 406L, 407L, 408L, 409L, 410L, 411L, 442L,
443L, 444L, 445L, 446L, 447L, 478L, 479L, 480L, 481L, 482L, 483L,
```

```
508L, 509L, 510L, 511L, 512L, 513L, 536L, 537L, 538L, 539L, 540L,
541L, 572L, 573L, 574L, 575L, 576L, 577L, 608L, 609L, 610L, 611L,
612L, 613L, 638L, 639L, 640L, 641L, 642L, 643L, 668L, 669L, 670L,
671L, 672L, 673L, 734L, 735L, 736L, 737L, 738L, 739L, 770L, 771L,
772L, 773L, 774L, 775L))

MNassay <-
structure (list(dose = c(0, 0, 0, 0, 0, 0, 0.375, 0.375, 0.375,
0.375, 0.375, 0.375, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 1.5,
1.5, 1.5, 1.5, 1.5, 1.5, 3, 3, 3, 3, 3, 3), MN = c(2, 2, 4, 0,
3, 2, 0, 2, 0, 2, 2, 4, 0, 1, 7, 0, 0, 1, 1, 8, 2, 2, 0, 1, 3,
14, 28, 19, 2, 6)), .Names = c("dose", "MN"), class = "data.frame", row. names = 698:727)

dose <- gl(5, 6, labels=c("NC", "D1", "D2", "D3", "D4"))
MN <- c(0, 0, 0, 0, 0, 0, 3, 1, 0, 1, 1, 2, 0, 1, 2, 0, 0, 0, 1, 0, 0, 2, 0, 0, 4, 0, 0,3, 3,
0)
MNassayZeroControl <- data.frame(dose, MN)

##Historical data
# Freeman-Tukey transformation
histMN$ft<-sqrt(histMN$MN)+ sqrt(histMN$MN+1)
#estimated p per historical study
phat_hist <- tapply(histMN$ft, histMN$Run, mean)
#obtrain standard value and its variance
MeanHist<-mean(phat_hist)
SDHist<-sd(phat_hist)
#find reference range
RefValue<-round(cbind(MeanHist-2*SDHist, MeanHist+2*SDHist), digits= 2)

##concurrent control data and analysis
#Freeman-Tukey transformation
MNassay$ft<-sqrt(MNassay$MN)+ sqrt(MNassay$MN+1)
#defining factor levels
MNassay$DOSE <- as.factor(MNassay$dose)
#load library multcomp which implements Williams-type procedures
ibrary (multcomp)
#Remove concurrent control from data
myMNassay <- droplevels(subset(MNassay, dose != 0))
# fit a reduced linear model
nam <- lm(ft ~ DOSE-1, data=myMNassay)
design<-summary(MNassay$DOSE)
```

```
#contrast matrix (without concurrent control)
cmatrix<-contrMat(design, type="Williams") [, -1]
#Find the Williams-type adjusted p-values against a standard
pValHist<-summary(glht(nam, linfct=cmatrix, rhs=MeanHist, alternative="greater"))$test$pvalues
#Print
print(pValHist)
#test against concurrent control
cmatrix_cc <-contrMat(design, type="Williams")
nam_cc <- lm(ft ~ DOSE-1, data=MNassay)
#Find the Williams-type adjusted p-values against concurrent control
pValCC<-summary(glht(nam_cc, linfct=cmatrix_cc, alternative=
"greater"))$test$pvalues
#Print
print(pValCC)
```

The crude tumor rate example

```
##read data
HIST <- data.frame(
     tumor = c(0,0,1,2,1,8,0,0,1,4,6,3,0,0,1,2,3,4,0,1,1,2,3),
     rats = c(20,12,19,25,10,49,20,12,19,47,54,18,19,10,17,22,20,20,17,20,15,20,20)
)
ACT <- data.frame(
     dose = c("0","75","150"),
     tumor = c(2,6,10),
     rats = c(20,49,49)
)

#Graphical illustration of the data
boxplot(HIST$tumor/HIST$rats, ylab="tumor rate", xlab="historical control and current study
(points)")
points(jitter(rep(1, 23)),HIST$tumor/HIST$rats, col="blue", pch=16)
points(c(1,1,1), ACT$tumor/ACT$rats, pch=19,type = "p",cex=1.3, col="red")
text(c(1,1,1)+0.3, ACT$tumor/ACT$rats, c("0","75","150"),cex=1.3)

#Using concurrent data
n <- ACT$rats # define a vector of number of experiments (n_i)
y <- ACT$tumor # define a vector of number of successes (y_i)
names(n)<-names(y)<-as.character(ACT$dose) # adding a names attribute
estp <- (y+0.5)/(n+1) # Add1 estimate for concurrent control proportion
varp <- estp*(1-estp)/(n+1) # Add1 variance estimate
# Creating an appropriate contrast matrix for Williams-type comparisons
```

```
library(MCPAN)

contrWill <- contrMat(n=n, type="Williams")

contrWill %*% estp # Estimator

# Obtain adjusted p-values

round(binomRDtest(y, n, type="Williams",method="ADD1", alternative="greater", dist="MVN")
$p.val.adj, digits=3)

#Simultaneous confidence intervals

Waldci(cmat=contrWill, estp=estp, varp=varp, varcor=varp, alternative="greater")$conf.int


#Using historical control data

HIST$prop <- with(HIST, (tumor+0.5)/(rats+1)) # Add1 estimate for hist control proportions

standard <- mean(HIST$prop) # calculating the standard value

#Reference interval

Ref     <-     round(c(standard-2*sqrt(standard*(1-standard)/n[1]),standard+2*sqrt(standard*(1-
standard)/n[1])),digits=2)

propCorrected <- estp[-1] - standard # substracting standard value from dose proportions

contrWill[,-1] %*% propCorrected # Estimator

# adjusted p-values

round(Waldtest(cmat=contrWill[,-1], estp=propCorrected, varp=varp[-1], alternative= "greater")
$p.val.adj, digits=3)

#Simultaneous confidence intervals

Waldci(cmat=contrWill[,-1], estp=propCorrected, varp=varp[-1], varcor=varp[-1], alternative
="greater")$conf.int
```

The poly-3 tumor tumor rate example

```
#Data set of historical control data

HardGland_HistControl <-

data.frame(

     NTPStudy  =  c(rep("C61494B",  50),rep("C99028",  50),rep("C55312B",  50),rep("C97003D",
                  50),rep("C54795B",      49),rep("C55174D",      50),rep("C60048B",      50),
                  rep("C93020D",50),  rep("C55323B",  50),  rep("C95003C",50),  rep("C91014",
                  50),rep("C55403D", 49),rep("C60139A", 50)),

     Days = c(732, 732, 732, 732, 732, 732, 732, 732, 655, 732, 717, 732, 732, 733, 733, 613,
            733, 733, 603, 733, 733, 631 733, 733, 733, 733, 733, 733, 733, 733, 733, 637,
            733, 733, 733, 733, 719, 733, 733, 733, 733, 733, 733, 733, 733, 733, 733, 577,
            733, 733, 733, 730, 730, 730, 730, 729, 729, 539, 730, 729, 633, 728, 729, 729,
            558, 694, 730, 729, 729, 729, 729, 578, 729, 397, 730, 730, 730, 730, 544, 730,
            728, 729, 728, 728, 729, 499, 729, 729, 730, 730, 728, 707, 730, 730, 461, 533,
            730, 730, 495, 533, 728, 728, 490, 729, 729, 727, 612, 664, 729, 584, 729, 727,
            728, 729, 729, 729, 728, 728, 727, 729, 618, 729, 727, 729, 729, 729, 727, 728,
            729, 728, 645, 673, 729, 729, 728, 700, 729, 729, 728, 729, 729, 729, 729, 729,
            729, 728, 729, 728, 729, 729, 420, 731, 730, 710, 731, 732, 731, 732, 730, 701,
            731, 610, 731, 731, 732, 617, 674, 730, 665, 730, 732, 730, 730, 730, 730, 730,
            728, 574, 693, 731, 643, 731, 731, 616, 732, 732, 732, 731, 730, 732, 529, 732,
            731, 732, 546, 679, 732, 731, 662, 609, 731, 735, 735, 735, 686, 735, 735, 707,
            707, 735, 735, 126, 735, 413, 231, 735, 546, 735, 735, 518, 735, 735, 735, 630,
            735, 462, 735, 735, 735, 735, 735, 532, 735, 735, 735, 735, 252, 490, 735, 329,
            266, 441, 686, 735, 735, 70, 735, 224, 686, 735, 729, 730, 731, 596, 730, 731,
            729, 674, 411, 593, 730, 473, 731, 729, 730, 729, 730, 730, 730, 729, 716, 729,
            731, 730, 731, 731, 729, 729, 730, 730, 729, 730, 729, 678, 730, 729, 731, 730,
```

```
            729, 731, 730, 649, 729, 731, 730, 729, 485, 729, 611, 731, 451, 730, 635, 730,
            730, 730, 730, 730, 730, 730, 731, 731, 731, 731, 731, 732, 732, 732, 732, 620,
            737, 737, 737, 737, 737, 738, 738, 738, 738, 738, 732, 431, 732, 732, 732, 737,
            737, 716, 690, 737, 730, 558, 730, 730, 730, 731, 731, 731, 731, 731, 731, 730,
            730, 729, 731, 729, 730, 729, 729, 731, 731, 729, 731, 732, 731, 729, 732, 725,
            731, 732, 731, 702, 682, 730, 706, 731, 592, 730, 515, 730, 672, 731, 731, 732,
            732, 640, 721, 729, 730, 730, 730, 730, 730, 731, 731, 517, 732, 730, 730, 598,
            736, 735, 736, 735, 736, 735, 353, 736, 736, 736, 670, 736, 735, 736, 584, 736,
            735, 622, 735, 736, 736, 735, 736, 736, 736, 736, 736, 541, 735, 736, 735, 562,
            735, 735, 736, 735, 636, 735, 736, 736, 736, 735, 736, 646, 616, 735, 735, 735,
            735, 344, 521, 709, 729, 619, 730, 521, 731, 729, 729, 730, 730, 729, 730, 631,
            668, 729, 698, 503, 730, 731, 730, 729, 730, 691, 731, 730, 730, 730, 714, 729,
            730, 730, 730, 721, 585, 622, 729, 729, 402, 730, 730, 730, 730, 598, 712, 730,
            731, 729, 729, 730, 729, 730, 729, 730, 691, 667, 729, 730, 729, 729, 730, 729,
            729, 729, 730, 730, 729, 730, 729, 729, 456, 730, 729, 730, 729, 452, 730, 730,
            729, 730, 729, 730, 730, 730, 729, 730, 730, 485, 730, 729, 729, 695, 730, 561,
            729, 729, 699, 729, 730, 734, 691, 735, 651, 647, 735, 706, 734, 690, 575, 687,
            735, 680, 733, 734, 646, 735, 735, 735, 733, 697, 733, 734, 735, 705, 735, 564,
            733, 734, 733, 734, 734, 733, 733, 733, 686, 702, 735, 733, 575, 494, 691, 733,
            735, 734, 735, 735, 735, 735, 674, 730, 216, 588, 730, 730, 730, 730, 730, 730,
            730, 625, 731, 439, 731, 625, 731, 731, 731, 731, 731, 731, 601, 731, 731, 575,
            732, 732, 732, 732, 732, 613, 732, 640, 732, 732, 732, 732, 733, 733, 733, 6, 496,
            733, 733, 733, 733, 733, 5, 733, 733),

      Lesions = c(0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
                  0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0 ,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
                  0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                  0))

#Data set of concurrent study

HardGland_concurrent <-

data.frame(Treatment.Group = factor(rep(c("0 MG/KG","200 MG/KG","630 MG/KG","2000 MG/KG"),
each=50)),

      Days = c(731, 730, 679, 732, 617, 732, 730, 733, 733, 733, 733, 732, 732, 730, 730, 732,
             731, 731, 368, 733, 730, 732, 733, 733, 733, 730, 733, 730, 732, 732, 731, 731,
             732, 731, 732, 731, 732, 730, 732, 732, 730, 730, 732, 713, 731, 730, 730, 731,
             733, 732, 733, 733, 732, 733, 732, 732, 733, 516, 730, 668, 730, 667, 622, 731,
             731, 730, 732, 732, 732, 732, 732, 732, 732, 730, 732, 538, 732, 730, 731, 732,
             733, 730, 730, 731, 732, 731, 731, 683, 730, 733, 730, 666, 731, 731, 684, 717,
             689, 730, 733, 732, 730, 541, 733, 732, 733, 731, 671, 608, 733, 732, 494, 730,
             733, 730, 730, 733, 731, 731, 732, 733, 732, 732, 731, 695, 374, 731, 732, 732,
             730, 731, 653, 730, 732, 733, 508, 732, 698, 733, 731, 733, 732, 730, 678, 646,
             731, 730, 731, 733, 733, 732, 656, 733, 730, 730, 733, 732, 731, 730, 732, 730,
             730, 730, 730, 731, 731, 731, 713, 733, 731, 671, 730, 573, 732, 731, 732, 676,
```

```
            731, 731, 730, 607, 580, 730, 733, 725, 732, 732, 732, 732, 730, 730, 730, 732,
            731, 732, 730, 731, 733, 733, 679, 731),

     Lesions = c(0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0,
                 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0))

#Summary table for historical control data

HistControl <- aggregate(Lesions ~ NTPStudy,data=HardGland_HistControl, FUN=sum)

HistControl$Number <- aggregate(Lesions ~ NTPStudy,data= HardGland_HistControl, FUN=length)[,2]

HistControl$Proportion <- with(HistControl, Lesions/Number)

# Calculating Poly3 adjusted tumour rates for histircal control data using a for loop

library(MCPAN)

HistControl$Poly3Prop <- c()

for(i in 1: length(levels(HardGland_HistControl$NTPStudy))){

  Data <- subset(HardGland_HistControl, NTPStudy==levels(HardGland_HistControl$NTPStudy)[i])

  HistControl$Poly3Prop[i]   <-   poly3est(time=Data$Days,   status=   Data$Lesions,   tmax=
  max(Data$Days))$estp

}


#Summary table for consurrent study

#order the dose levels

HardGland_concurrent$Treatment.Group <- factor(HardGland_concurrent$Treatment.Group,

                                    levels=c("0   MG/KG","200   MG/KG","630   MG/KG","2000
                                    MG/KG"))

Current <- aggregate(Lesions ~ Treatment.Group,data=HardGland_concurrent, FUN=sum)

Current$Number    <-    aggregate(Lesions    ~    Treatment.Group,data=HardGland_concurrent,
FUN=length)[,2]

Current$Proportion <- with(Current, Lesions/Number)

# Calculating Poly3 adjusted tumour rates for concurrent study using a for loop

Current$Poly3Prop <- c()

for(i in 1:length(levels(HardGland_concurrent$Treatment.Group))){

  Data                              <-                              subset(HardGland_concurrent,
Treatment.Group==levels(HardGland_concurrent$Treatment.Group)[i])

  Current$Poly3Prop[i]        <-        poly3est(time=Data$Days,        status=Data$Lesions,
tmax=max(Data$Days))$estp

}

# Calculating the variance of Poly3 adjusted proportions

Current$VarPoly3 <- with(Current, Poly3Prop*(1-Poly3Prop)/(Number))


## Umbrella protected Williams type analysis using concurrent control

library(multcomp)

# Umbrella protected contrast matrix for Williams-type comparisons
```

```
contrUWill <- contrMat(n=Current$Number, type="UmbrellaWilliams")

Waldtest(cmat=contrUWill, estp=Current$Poly3Prop, varp=Current$VarPoly3, alternative="greater")


## Analysis using the mean of hist. control proportions as reference value

# mean of Poly3 adjusted proportions of historical control data

standard <- mean(HistControl$Poly3Prop)

# Reference interval

Ref <- round(c(standard-2*sqrt(standard*(1-standard)/Current$Number[1]),

          standard+2*sqrt(standard*(1-standard)/Current$Number[1])),digits=3)

# shifted Poly3 adjusted proportions

Current$Poly3PropShift <- with(Current, Poly3Prop-standard)

# calculating p-values

Waldtest(cmat=contrUWill[,-1], estp=Current$Poly3PropShift[-1], varp=
Current$VarPoly3[-1], alternative="greater")
```

Simulation study

To evaluate the performance of the Williams-type procedure for the comparison against a standard value for crude proportions (further denoted as *Standard*), we compared it with a Williams-type procedure for crude proportions using the concurrent control only (modified for small sample sizes by adding pseudo-observations as described in 25, further denoted as *Add*1) and a method that incorporates the historical control information by fitting a beta-binomial model [19, further denoted as *BetaBin*] using Monte Carlo simulations. The generated datasets were designed according to the standard experimental design from the US National Toxicology Program. For the concurrent study we used a balanced design with one control group and three dose groups, where the sample size per group was set to 50. Further-more, we assumed a total number of 20 available historical control studies. The binomial proportions for the historical control groups were simulated from a beta distribution with parameters a and b. The parameters of the beta distribution were set to get an expected proportion of 0.1. Therefore, we selected three scenarios: Scenario A with *beta* $(a = 20, b = 180)$ representing a small between study variability (0.0004); Scenario B with *beta* $(a = 5, b = 45)$ reecting a moderate between study variance (0.0018);

Scenario C with *beta* $(a = 2, b = 18)$ that corresponds to a high between study variability (0.0043). Figure A.4 represents density curves for the three settings. The binomial proportions for the current study were set to $(\pi_0 = \chi, \pi_1 = \chi, \pi_2 = \chi + 0.05, \pi_3 = \chi + 0.05)$, representing an increasing dose-response pattern. To change the binomial proportion of the concurrent control group in comparison with the expected value of the historical control groups, we varied χ between 0 and 0.3 in increments of 0.01.

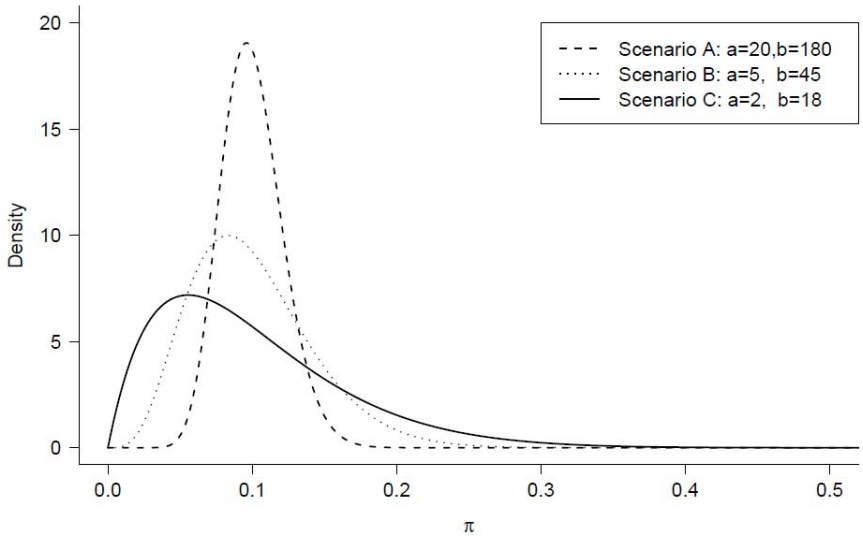

Figure A.4. Density curves of beta distributions used for Monte Carlo simulations.

Note that $\chi < 0.1$ reflects the situation where the proportion of the concurrent study is smaller than the expected proportion of the historical control groups and $\chi > 0.1$ corresponds to the case where it is greater. For each simulated data set and each method, we calculated Williams-type simultaneous confidence intervals. For each method, we computed the empirical power as the probability that the lower bound for any of the calculated confidence intervals is greater than zero. For each parameter setting, we simulated 10,000 data sets.

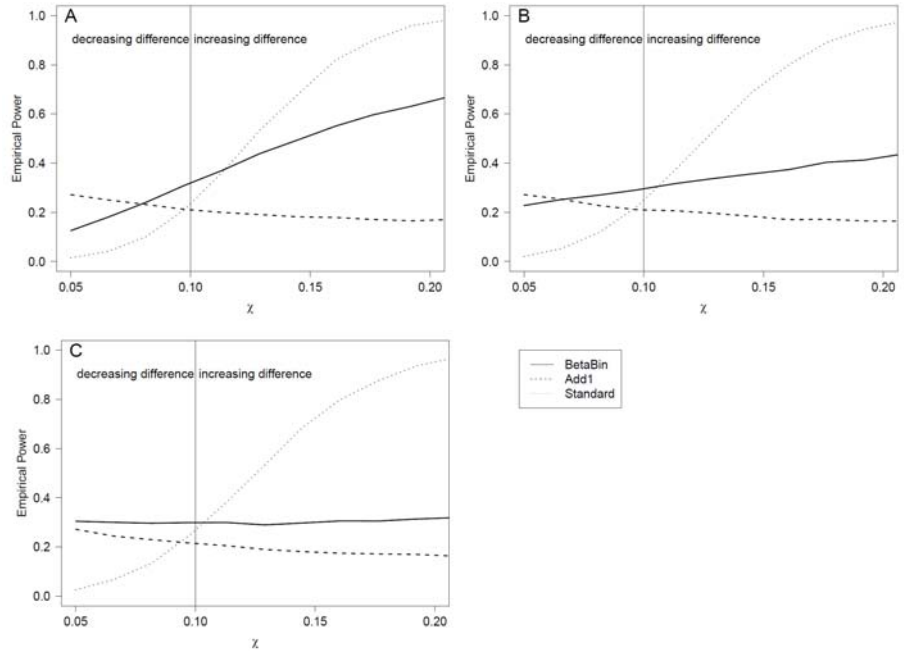

Figure A.5. Empirical power to detect an increasing dose effect. Values of χ below 0.1 represent a decreasing difference between the concurrent control and the expected value of the historical control data, while values above 0.1 represent an increasing difference.

Figure A.5 presents the empirical power for the three scenarios and increasing differences between the binomial proportion of the concurrent study and the expected value of the historical control groups. The power to detect an increasing difference between the control group and the remaining dose groups of both methods that incorporate the historical control information, *Standard* and *BetaBin*, monotonously increase with increasing χ. As opposed to this, the power of the *Add*1 method slightly decreases with increasing χ, because the variance estimator of the binomial proportions is also affected by the choice of χ.

In the chosen simulation setting the value $\chi = 0.1$ presents a cut point: if $\chi < 0.1$, the proportion of the concurrent control group is lower than the expected value of the historical control groups, and if $\chi > 0.1$, the concurrent control group is greater than the expected value of the historical control groups. Among the methods that incorporate historical control information, the *BetaBin* method shows a better performance by means of a greater power than the *Standard* method for $\chi < 0.1$. If the expected value of the historical control groups is greater than the concurrent control group, the *Standard* method shows a greater power than the *BetaBin* procedure.

The power of the *BetaBin* approach approximates the *Add*1 method with increasing between study variability. In contrast, the *Standard* approach performs equal in all scenarios, because this method does not take the between study variability into account. In summary, the *Standard* method is advantageous in situations where the expected value of the historical control data is smaller in comparison to the concurrent control group.