The International Journal of Biostatistics

Volume 6, *Issue* 1 2010 *Article* 15

Evaluation of Incidence Rates in Pre-Clinical Studies Using a Williams-Type Procedure

Ludwig A. Hothorn, *Leibniz University Hannover*
Martin Sill, *German Cancer Research Center*
Frank Schaarschmidt, *Leibniz University Hannover*

Recommended Citation:
Hothorn, Ludwig A.; Sill, Martin; and Schaarschmidt, Frank (2010) "Evaluation of Incidence Rates in Pre-Clinical Studies Using a Williams-Type Procedure," *The International Journal of Biostatistics*: Vol. 6: Iss. 1, Article 15.
DOI: 10.2202/1557-4679.1180

Evaluation of Incidence Rates in Pre-Clinical Studies Using a Williams-Type Procedure

Ludwig A. Hothorn, Martin Sill, and Frank Schaarschmidt

Abstract

The analysis of dose-response relationships is a common problem in pre-clinical studies. For example, proportions such as mortality rates and histopathological findings are of particular interest in repeated toxicity studies. Commonly applied designs consist of an untreated control group and several, possibly unequally spaced, dosage groups. The Williams test can be formulated as a multiple contrast test and is a powerful option to evaluate such data. In this paper, we consider simultaneous inference for Williams-type multiple contrasts when the response variable is binomial and sample sizes are only moderate. Approximate simultaneous confidence limits can be constructed using the quantiles of a multivariate normal distribution taking the correlation into account. Alternatively, multiplicity-adjusted p-values can be calculated as well. A simulation study shows that a simple correction based on adding pseudo observations leads to acceptable performance for moderate sample sizes, such as 40 per group. In addition, the calculation of adjusted p-values and approximate power is presented. Finally, the proposed methods are applied to example data from two toxicological studies; the methods are available in an R-package.

KEYWORDS: binomial, ordered proportions, simultaneous confidence intervals, toxicology

Author Notes: Part of the work of LAH was performed during a sabbatical stay at TSRI, La Jolla. I like to thank the group leader, Dr. Nickolas Schork, for generous local support. The authors wish to thank the referees for their constructive comments, especially the editor's careful reading of the paper and helpful suggestion, which led to several substantial improvements.

1 Introduction

Standard experimental design in repeated toxicity studies consists of a negative control and several (commonly three) treatment groups. The frequently used parametric Dunnett (1955) and Williams (1971) procedures as well as the nonparametric Dunn (1964) and Shirley (1977) procedures for skewed data, are the recommended statistical procedures for toxicological studies involving continuous data (National Toxicology Program, 2009). In comparing particular treatments with the control group, Dunnett and Dunn procedures do not exploit the potential monotonicity of effects, which occurs when treatments represent increasing dosage. Conversely, the methods of Williams and Shirley assume the monotonicity of effects in comparing the dose and control groups.

Various types of response variables (e.g., continuous, proportions (rates), ordered categorical and multinomial data) are common in toxicological studies (OECD408, 1998). However, most recommendations concerning statistical methodology focus on continuous endpoints (see, e.g. National Toxicology Program, 2009). For example, the NTP provides the following recommendation regarding proportional data: 'Because vaginal cytology data are proportions (the proportion of the observation period that an animal was in a given estrous stage), an arcsine transformation is used to bring the data into closer conformance with a normality assumption' (National Toxicology Program, 2009). However, simulation studies involving small samples (Carriere, 2001) and zero cells (Rucker et al., 2009) showed that the arcsine transformation is inappropriate. A number of specific issues need to be addressed in toxicological studies that involve binomial proportions: 1. sample size may be too small to allow the application of simple asymptotic methods; 2. the direction of the toxic effect is usually known, (e.g., we are interested in only increasing mortality or incidence rates); and 3. no, or very few, critical events may be observed in the control group.

We propose using trend tests based on a totally ordered alternative hypothesis to evaluate dose-response relationships. A large number of such tests have been published. The most commonly used approach for proportional data - Cochran-Armitage test (Armitage, 1955) - rejects the null hypothesis of equal proportions in favor of a linear trend. However, this test is underpowered when the true trend is not linear, i.e., when the dose and response exhibit a convex or concave relationship (Bretz and Hothorn, 2002). Also, assuming a linear trend may be inappropriate in many studies, e.g., when the study design involves unequally spaced dosage levels between treatment groups. Additionally, the power of this test is strongly influenced by the choice of dosage scores used in the calculation of the test statistic, and optimal dosage scores are often unknown a priori. For datasets with small sample sizes, using exact unconditional and conditional Cochran-Armitage trend tests has

been proposed, and may be appropriate under different models (Tang et al., 2006). Another approach involves testing whether a slope parameter of a logistic model is non-zero, but the assumption of a linear relationship may again be inappropriate. Finally, a likelihood ratio test based on isotonic regression (Leuraud and Benichou, 2006) can be applied.

As mentioned, assuming a linear dose-response relationship may be inappropriate or at least suboptimal in terms of power in cases of sub- and supra-linear relationships. Also, one may often be interested in comparing several dose groups with the control group explicitly. Adequate methods should allow the definition of one-sided tests or confidence intervals when the direction of the effect of interest is known a priori. For these reasons, the Williams test (Williams, 1971, 1972) provides a good option to use because it was constructed for both a total order alternative as well as specific comparisons versus control. For a general unbalanced design and normally distributed variables, Bretz (2006) showed that Williams trend test can be approximately formulated as a multiple contrast test. For large sample sizes, similar asymptotic multiple contrast tests for binomial proportions have also been proposed (Bretz and Hothorn, 2002). A Dunnett-type procedure for one- or two-sided comparisons for the difference of proportions between several treatments and a control was proposed (Schaarschmidt et al., 2009), using the analogous approach of multiple contrasts, without an order restricted alternative. Recently, the small sample performance of simultaneous confidence intervals for contrasts of tumor proportions, confounded by mortality without cause-of-death information, has been investigated (Schaarschmidt et al., 2008a), including Williams-type contrasts as a special case. In contrast to the problem considered here, such observations are assumed to originate from realizations of two competing events assumed to follow two independent Weibull distributions.

In this paper, we investigate simultaneous confidence limits as well as multiplicity-adjusted p-values for Williams-type contrasts of binomial proportions. We assume that using binomial proportions as an effect measure is toxicologically meaningful and not biased by competing risks. Further, we explore ways to adjust for small sample size, and pay special attention to one-sided alternatives (i.e., settings, in which the direction of an interesting effect is already known). We achieve multiplicity adjustment by using appropriate quantiles of the multivariate normal distribution, taking the correlation between the contrasts into account. The confidence limits proposed in this paper can be used in inference for a global trend hypothesis, and to display the differences between pooled proportions of the dose groups and the proportion of the control group. The described procedure is validated in an extensive simulation study, comparing different small sample adjustments for balanced and unbalanced sample sizes, and large sets of binomial parameters sampled from the parameter space. We describe the appropriate test and

DOI: 10.2202/1557-4679.1180

adjusted p-values for single contrasts, and present an approximate power calculation for the global test on trend. The use of approximate power calculation is illustrated for situations related to the example and compared to results of simulated power. An R-software package implementing the proposed methods has been made publicly available.

2 Motivating examples

In a toxicological study, mice were exposed to a control treatment and three doses of a compound. After 6 months, mortality rates were assessed (Hothorn, 1994, see Tab. 1).

In the second study, the incidences of tubular epithelial hyaline droplet generation in male rats were reported for a 28-day oral dose toxicity study of nonylphenol (Woo et al., 2007, see Tab. 2).

Table 1: Chronic toxicity study on mice

Treatment	Control	10 mg/kg	50 mg/kg	100 mg/kg
No. of dead mice	4	1	6	8
Total no. of mice	40	20	20	20
Proportion of dead mice	0.10	0.05	0.30	0.40

Table 2: 28-day toxicity study on male rats

Treatment	Control	10 mg/kg	50 mg/kg	250 mg/kg
No. with hyaline droplets	0	0	3	8
No. of rats under observation	10	10	10	10
Observed proportion	0	0	0.3	0.8

In both examples, the aim is to detect the global trend depending on the dosage levels. In addition, extracting information regarding the dosage levels that lead to the trend may be necessary. The sample sizes are clearly too small to allow the determination of asymptotic confidence limits.

3 A Williams-type procedure

Let us consider a completely randomized one-way layout with I groups, $i = 1, ..., I$, where n_i denotes the number of Bernoulli trials in the ith group, Y_i is the number of successes among the n_i trials, and $i = 1$ denotes the control group and $i = 2, ..., I$

denotes treatment groups ordered by increasing dosage. The Y_is are assumed to be independent binomial random variables $Y_i \sim Bin(n_i, \pi_i)$, with point estimators $p_i = Y_i/n_i$.

Let $C = (c_1, ..., c_I)$ be a vector of contrast coefficients fulfilling the constraint $\sum_{i=1}^{I} c_i = 0$. Then, the linear combination $L = \sum_{i=1}^{I} c_i \pi_i$ has expectation $E(L) = 0$ if all proportions are equal such that $\pi_1 = \pi_2 = ... = \pi_I$. If we also define C such that $\sum_{c_i \leq 0} c_i = -1$ and $\sum_{c_i > 0} c_i = 1$, the linear combination L can be interpreted as a difference of weighted averages of proportions, with the simple difference of two proportions as a special case.

According to Bretz (2006), the Williams test (Williams, 1971, 1972) can be reformulated in terms of a test for $M = I - 1$ linear combinations L_m, $m = 1, ..., M$ of the proportions π_i. The contrast coefficients are defined in the following $M \times I$ matrix with elements c_{mi},

$$C_{M\times I} = \begin{pmatrix} -1 & 0 & \cdots & 0 & 0 & 1 \\ -1 & 0 & \cdots & 0 & \frac{n_{I-1}}{n_{I-1}+n_I} & \frac{n_I}{n_{I-1}+n_I} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ -1 & \frac{n_2}{n_2+\cdots+n_I} & \cdots & \frac{n_{I-2}}{n_2+\cdots+n_I} & \frac{n_{I-1}}{n_2+\cdots+n_I} & \frac{n_I}{n_2+\cdots+n_I} \end{pmatrix} \quad (1)$$

In this matrix, each row corresponds to one linear combination

$$L_m = \sum_{i=1}^{I} c_{mi} \pi_i$$

To test for an increasing trend, we test the intersection of the elementary null hypotheses $H_0 : \bigcap_{m=1}^{M} L_m \leq 0$, versus the union of the elementary alternative hypotheses $H_1 : \bigcup_{m=1}^{M} L_m > 0$. That is, the objective is to perform a one-sided Union-Intersection-Test on the M linear combinations (Bretz and Hothorn, 2002, Bretz, 2006), while controlling the family-wise error rate over all M contrasts.

Alternatively, lower simultaneous confidence limits for the linear combinations L_m can be used. In this case, one can conclude a presence of a trend in the proportions $\pi_1, ..., \pi_I$ if at least one of the M lower confidence limits excludes the value 0. In addition, the confidence limits describe the difference between the control group $i = 1$ and the weighted average of proportions of higher dose groups, and allow one to interpret the relevance of the effect size. Another advantage of confidence intervals lies in their ability to summarize the uncertainty of the estimates using an easily interpretable scale of differences in proportions.

In other words, the above procedure tests the null-hypothesis of no difference among the proportions,

$$H_0 : \pi_1 = \pi_2 = ... = \pi_I, \quad (2)$$

DOI: 10.2202/1557-4679.1180

against the alternative that the binomial proportions are increasing with increasing dosage compared to the control group. To achieve this, higher dose groups are successively pooled and compared to the control. This approach leads to multiple comparisons regarding a single overall question, and, therefore, an adjustment is needed to ensure that the overall hypothesis is tested at level α. The M local hypotheses are positively correlated, and this correlation is included in the test procedure in order to avoid the overall test becoming too conservative. By testing M different local hypotheses, the procedure is sensitive to a number of different dose-response shapes, and allows one to gain a more detailed knowledge of the dose-response shape than would be possible using a simple p-value for a global test of trend.

3.1 Approximate confidence limits for a single linear combination of proportions

The point estimator for a single linear combination L is $\hat{L} = \sum_{i=1}^{I} c_i p_i$ and the lower $(1-\alpha)$ Wald confidence limit for L is:

$$\left[\sum_{i=1}^{I} c_i p_i - z_{1-\alpha} \sqrt{\sum_{i=1}^{I} c_i^2 \hat{V}(p_i)} \right] \tag{3}$$

with $\hat{V}(p_i) = p_i(1-p_i)/n_i$ and $z_{1-\alpha}$ denoting the $(1-\alpha)$ quantile of the standard normal distribution. Wald limits for binomial proportions are known to keep the $(1-\alpha)$ coverage probability only for asymptotically large sample sizes (Agresti and Caffo, 2000, Price and Bonett, 2004, Brown and Li, 2005). So far, no exact confidence limits are available for a linear combination of proportions. In a seminal paper, Agresti and Coull (1998) showed that adding a total of four pseudo-observations to the observed successes and failures yields approximate confidence intervals for one binomial proportion with good small sample performance. Motivated by approximating the posterior distribution following from a uniform beta prior, Agresti and Caffo (2000) proposed an improved Wald confidence interval for the difference of two proportions by adding two successes and two failures. Brown and Li (2005) recommended the use of this method based on a comparative simulation study due to its good coverage probability for moderate sample sizes. Price and Bonett (2004) further extended this interval to linear combinations of I proportions. In their approach, p_i in Equation (3) is replaced by $\tilde{p}_i = (Y_i + 2/g)/(n_i + 4/g)$, and $\hat{V}(p_i)$ by $\tilde{p}_i(1-\tilde{p}_i)/(n_i + 4/g)$, with g the number of non-zero contrast coefficients. Alternatively, p_i in Equation (3) is replaced by $\tilde{p}_i = (Y_i + 1)/(n_i + 2)$, and $\hat{V}(p_i)$ by $\tilde{p}_i(1-\tilde{p}_i)/(n_i + 2)$. Both intervals are the Agresti and Caffo interval for the difference of two binomials if the contrast has only two non-zero coefficients.

Price and Bonett (2004) investigated the performance of their method in a simulation study involving different types of single contrasts, and concluded that, compared to the Wald limit, it exhibited better coverage probability for the improved limits.

These simulation studies, however, considered only the coverage probability of two-sided $(1-\alpha)$ confidence intervals. The coverage probability of one-sided limits was investigated only by Cai (2005) in the case of a single binomial proportion. The direction of interest is usually known in trend tests, and only the upper or lower limit is necessary for a decision regarding the trend. Therefore, we are mainly interested in the performance of one-sided confidence limits. However, when the assumption of monotonicity is in question, two-sided confidence intervals may be more appropriate. See for example the recent discussion of Shirley and Peddada (Shirley, 2007).

3.2 Approximate simultaneous confidence limits for multiple linear combinations of proportions

Approximate simultaneous confidence limits for M linear combinations L_m can be constructed using Equation (4):

$$\sum_{i=1}^{I} c_{im}\tilde{p}_i - q_{M,R,1-\alpha}\sqrt{\sum_{i=1}^{I} c_{im}^2 \tilde{V}(p_i)} \tag{4}$$

Here, $q_{M,R,1-\alpha}$ is the equicoordinate $(1-\alpha)$ quantile of a M-variate normal distribution with correlation matrix $\mathbf{R}$, with its CDF denoted $\Phi_M(q;\mathbf{0},\mathbf{R})$. The quantile $q_{M,R,1-\alpha}$ is chosen such that

$$\Phi_M(q = q_{M,R,1-\alpha};\mathbf{0},\mathbf{R}) = P(Z_m \leq q,\ \forall m = 1,...,M) = 1-\alpha,$$

where $\mathbf{Z}$ is an M-variate normal random vector with elements Z_m, $m = 1,...,M$, expectation $\mathbf{0}$ and correlation matrix $\mathbf{R}$. Due to the specific choice of the quantile $q_{M,R,1-\alpha}$, the probability that at least one of the M values of $\mathbf{L}$ is excluded by the confidence limits is α if $n \to \infty$. Upper confidence limits can be calculated accordingly. By using quantiles of the multivariate normal distribution, the number of estimated linear combinations as well as the correlation between them is taken into account.

When both increasing or decreasing trends are of interest, two-sided confidence intervals can be calculated:

DOI: 10.2202/1557-4679.1180

$$\sum_{i=1}^{I} c_{im}\tilde{p}_i \pm q'_{M,R,1-\alpha}\sqrt{\sum_{i=1}^{I} c_{im}^2 \tilde{V}(p_i)} \tag{5}$$

Then, a quantile $q'_{M,R,1-\alpha}$ has to be chosen such that

$$\Phi_{M,two-sided}\left(q' = q'_{M,R,1-\alpha}; \mathbf{0}, \mathbf{R}\right) = P\left(|Z_m| \leq q' \ \forall m = 1,...,M\right) = 1-\alpha.$$

Numerically, probabilities $\Phi_M(q; \mathbf{0}, \mathbf{R})$, $\Phi_{M,two-sided}(q'; \mathbf{0}, \mathbf{R})$ as well as quantiles $q_{M,R,1-\alpha}$ and $q'_{M,R,1-\alpha}$ can be calculated using the package `mvtnorm` (Hothorn et al., 2001) in R. Following a similar approach as Piegorsch (1991) and Schaarschmidt et al. (2008a, 2009), the correlation matrix $\mathbf{R}$ is computed depending on the known constants c_{im} and n_i and the estimates $\tilde{\pi}_i$ and $\tilde{V}(p_i)$ in Table 3. For computational details we refer to Schaarschmidt et al. (2008a).

Regarding the approaches of Agresti and Caffo (2000) and Price and Bonett (2004), we investigated five different methods of adjustment. Table 3 summarizes choices for $\tilde{p}_i$ and $\tilde{V}(p_i)$ in Equation (4), where g is the number of contrast coefficients with $c_{im} \neq 0$ in the mth contrast. When referring to these confidence limits below we use the notations in Table 3.

Table 3: Choices for $\tilde{p}_i$ and $\tilde{V}(p_i)$ in Equation (4)

Notation	$\tilde{p}_i$	$\tilde{V}(p_i)$
Wald	Y_i/n_i	$p_i(1-p_i)/n_i$
add-1	$(Y_i+0.5)/(n_i+1)$	$\tilde{p}_i(1-\tilde{p}_i)/(n_i+1)$
add-2	$(Y_i+1)/(n_i+2)$	$\tilde{p}_i(1-\tilde{p}_i)/(n_i+2)$
add-2/g	$\left(Y_i+\frac{1}{g}\right)/\left(n_i+\frac{2}{g}\right)$	$\tilde{p}_i(1-\tilde{p}_i)/\left(n_i+\frac{2}{g}\right)$
add-4/g	$\left(Y_i+\frac{2}{g}\right)/\left(n_i+\frac{4}{g}\right)$	$\tilde{p}_i(1-\tilde{p}_i)/\left(n_i+\frac{4}{g}\right)$

3.3 Corresponding multiple tests

In order to use adjusted p-values instead of simultaneous confidence limits to perform the test procedure defined above, M test statistics have to be calculated:

$$T_m = \frac{\sum_{i=1}^{I} c_{im}\tilde{p}_i}{\sqrt{\sum_{i=1}^{I} c_{im}^2 \tilde{V}(p_i)}} \tag{6}$$

with adjustments listed in Table 3. The global null hypothesis $H_0 : \bigcap_{m=1}^{M} L_m \leq 0$ can be rejected if $max(T_m) > q_{M,R,1-\alpha}$, i.e., if at least one of the test statistics exceeds the critical value of an M-variate normal distribution.

Corresponding to the above hypotheses, adjusted p-values for each of the M contrasts can be computed as: $1-\Phi_M\left(q=T_m;\mathbf{0},\mathbf{R}\right)$. In case of two-sided tests $1-\Phi_{M,two-sided}\left(q'=T_m;\mathbf{0},\mathbf{R}\right)$ has to be used instead.

3.4 Software

The availability of relevant and free statistical software is important in toxicology, because fewer statisticians engage in pre-clinical statistics compared to clinical statistics. The methods presented here are included in the R-package `MCPAN` (Schaarschmidt et al., 2008b), which can be downloaded via CRAN of R. Before use in regulatory toxicology, this software needs to be further validated in detail.

4 Simulation study

4.1 Parameter settings

We investigated the coverage probability of the simultaneous confidence limits in a simulation study that focused on small to moderate sample sizes and $n_i = 100$ for nearly asymptotic behavior. Results for the situations $I = 3,4,6,10$, with balanced sample sizes $n_i = 10,20,40,60,100$ are summarized in Tables 4 and 5. Since most designs in toxicology involve three dose groups and a negative control, the situation $I = 4$ is the most relevant. Although sample sizes smaller than $n_i = 20$ are hardly reasonable from the perspective of power, sample sizes as small as $n_i = 10$ can be found in practice, e.g. in Kandori et al. (2005). Also, markedly unbalanced sample sizes may occur, e.g., $n_i = 42,...,74$ in Bell et al. (2007). For this reason, we investigated a number of unbalanced situations (Tables 6 and 7).

To assess the methods' performance in the entire parameter space, 10,000 combinations $\{\pi_1,...,\pi_I\}$ were sampled from independent uniform distributions $[0,1]$. For each of these combinations and sample size settings, 10,000 random samples $\{y_1,...,y_I\}$ were drawn from binomial distributions $Bin(n_i,\pi_i)$. Limits were considered to cover the true value when all estimated confidence limits included the corresponding true linear combination L_m. Known values of π_i, instead sample estimates, were used to calculate the correlation matrix in this main part of the simulation study.

In an additional simulation study, we explored the coverage probabilities of limits with a correlation matrix estimated from the samples. This was done for a small subset of scenarios in order to show that the above simulations appropriately characterize the proposed methods. We considered a balanced sample size of 40 with 4 groups. Analogously to the main study, combinations $\{\pi_1,...,\pi_I\}$ were

DOI: 10.2202/1557-4679.1180

Table 4: Lower 0.95 confidence limits in balanced designs: Proportion of situations with coverage probability between 0.94 to 0.96

I	n	Wald	add-1	add-2	add-2/g	add-4/g
3	10	0.208	0.506	0.234	0.481	0.315
3	20	0.359	0.700	0.359	0.676	0.463
3	40	0.502	0.832	0.492	0.827	0.617
3	60	0.600	0.884	0.576	0.880	0.702
3	100	0.722	0.931	0.688	0.933	0.797
4	10	0.202	0.487	0.269	0.385	0.433
4	20	0.320	0.674	0.394	0.540	0.598
4	40	0.436	0.814	0.541	0.717	0.748
4	60	0.519	0.875	0.632	0.813	0.813
4	100	0.641	0.925	0.740	0.900	0.882
6	10	0.198	0.440	0.324	0.317	0.418
6	20	0.295	0.584	0.461	0.422	0.594
6	40	0.394	0.744	0.620	0.547	0.772
6	60	0.458	0.827	0.705	0.641	0.841
6	100	0.559	0.903	0.797	0.772	0.908
10	10	0.180	0.364	0.379	0.254	0.316
10	20	0.263	0.467	0.535	0.326	0.432
10	40	0.339	0.622	0.693	0.435	0.578
10	60	0.397	0.681	0.769	0.512	0.681
10	100	0.492	0.836	0.848	0.617	0.802

drawn from a uniform distribution, and random samples were drawn for each combination. The coverage probability of each limit and computation method of the correlation was calculated, resulting in only a negligible difference in the second decimal position of the coverage probability value.

4.2 Criteria

Due to the discreteness of the binomial distribution, the coverage probability of confidence limits for contrasts of proportions oscillates depending on π_i and n_i. Confidence limit methods that exhibit coverage probabilities greater than or equal to the nominal level for all parameter settings will, on average, necessarily be conservative. For this reason, Agresti and Coull (1998) and Brown and Li (2005), as well as others, recommend confidence limit methods if their coverage probability is close to the nominal level, but not necessarily for all cases equal to or greater than

Table 5: Lower 0.95 confidence limits in balanced designs: Mean coverage probability for 10,000 randomly chosen settings

I	n	Wald	add-1	add-2	add-2/g	add-4/g
3	10	0.913	0.951	0.959	0.948	0.957
3	20	0.933	0.950	0.955	0.949	0.954
3	40	0.942	0.950	0.953	0.949	0.952
3	60	0.945	0.950	0.952	0.949	0.952
3	100	0.947	0.950	0.951	0.950	0.951
4	10	0.908	0.951	0.959	0.944	0.954
4	20	0.930	0.950	0.956	0.946	0.952
4	40	0.941	0.950	0.953	0.948	0.951
4	60	0.944	0.950	0.952	0.948	0.951
4	100	0.946	0.950	0.951	0.949	0.950
6	10	0.902	0.951	0.960	0.938	0.950
6	20	0.927	0.950	0.956	0.943	0.950
6	40	0.939	0.950	0.954	0.946	0.950
6	60	0.943	0.950	0.953	0.950	0.947
6	100	0.946	0.950	0.952	0.948	0.950
10	10	0.894	0.950	0.961	0.928	0.941
10	20	0.923	0.949	0.957	0.938	0.944
10	40	0.936	0.949	0.954	0.943	0.947
10	60	0.941	0.949	0.953	0.945	0.948
10	100	0.945	0.949	0.952	0.947	0.949

the nominal level. Accordingly, we present the proportion of 10,000 parameter settings $\{\pi_i, ..., \pi_I\}$ for which coverage probability was between 0.94 to 0.96. We consider this to be the main criterion for recommending a method (Table 4). Additionally, the mean coverage probabilities over all 10,000 settings are given in Table 5. These values provide additional information whether the considered confidence limit is, on average, liberal or conservative.

4.3 Results

The results for nominal 0.95 lower confidence limits in situations with balanced sample sizes are summarized in Tables 4 and 5. In the case of $3, 4$, or 6 groups, the add-1 limit achieved the highest proportion of coverage probabilities between 0.94 and 0.96, and a mean coverage probability closest to the nominal confidence level 0.95. However, for small sample sizes, such as $n_i = 10$ or 20, and larger number

DOI: 10.2202/1557-4679.1180

of groups, this method can be liberal for large proportions of settings. In these situations, our results suggested that the add-2 method constitutes a better choice if conservative performance is acceptable. The add-2/g and add-4/g methods tended to be liberal for large number of groups. As expected, the Wald limit was more liberal than all the other methods under all simulated situations.

Table 6: Lower 0.95 confidence limits in unbalanced designs: Proportion of situations with coverage probability between 94% to 96%

n_1, n_2, n_3, n_4	Wald	add-1	add-2	add-2/g	add-4/g
64, 32, 32, 32	0.513	0.864	0.509	0.862	0.723
80, 40, 30, 10	0.280	0.503	0.467	0.474	0.664
10, 30, 40, 80	0.147	0.312	0.608	0.241	0.497
20, 30, 50, 60	0.228	0.494	0.645	0.367	0.736
60, 50, 30, 20	0.382	0.800	0.495	0.765	0.717

Table 7: Lower 0.95 confidence limits in unbalanced designs: Mean coverage probability for 10,000 randomly chosen settings.

n_1, n_2, n_3, n_4	Wald	add-1	add-2	add-2/g	add-4/g
64, 32, 32, 32	0.941	0.950	0.953	0.948	0.952
80, 40, 30, 10	0.910	0.947	0.955	0.945	0.954
10, 30, 40, 80	0.903	0.951	0.961	0.941	0.952
20, 30, 50, 60	0.930	0.950	0.956	0.945	0.951
60, 50, 30, 20	0.934	0.950	0.953	0.947	0.952

For the unbalanced four group designs considered in Tables 6 and 7, the add-2 and add-4/g limits approached the highest proportions of coverage probability between 0.94 and 0.96. The add-2 limit is conservative in situations with extremely small sample sizes. Especially when the control group sample size was small, the performance of all limits became weak.

5 Approximative power calculation

Although the main focus of this article is confidence limit estimation, users may be interested in power calculation for the global test on trend. Bretz and Hothorn (2002) derive an approximative calculation for the power of multiple contrast tests for binary data. Their method is based on a Wald-type test statistic, using the maximum likelihood estimators for variance estimation, and a pooled variance estimator

under the null hypothesis. This method can be easily adapted to the test statistic presented in Section 3.3, assuming that we wish to detect an increasing trend using Williams contrasts with the global null hypothesis $H_0 : \bigcap_{m=1}^{M} L_m \leq 0$ and the alternative $H_1 : \bigcup_{m=1}^{M} L_m > 0$. Under the alternative, we assume true proportions π_i and sample sizes n_i. Then, for large n_i, a single test statistic follows a normal distribution with expectation

$$E(T_m) = \frac{\sum_{i=1}^{I} c_{im}\tilde{\pi}_i^*}{\sqrt{\sum_{i=1}^{I} c_{im}^2 \tilde{\pi}_i^* \left(1-\tilde{\pi}_i^*\right)/\tilde{n}_i^*}} \tag{7}$$

and variance $V(T_m) = 1$, where $\tilde{\pi}_i^* = (n_i\pi_i + 0.5)/(n_i + 1)$ and $n_i^* = n_i + 1$, e.g. for the add-1 adjustment. The M test statistics jointly follow an M-variate normal distribution with $\mathbf{e}$ being the vector of expectations with elements $E(T_m)$, and correlation matrix $\mathbf{R}$ as defined in Schaarschmidt et al. (2008a). The power to reject the global null hypothesis is the probability that at least one T_m exceeds the equi-coordinate critical value $q_{M,R,1-\alpha}$. It can, therefore, be calculated using: $1 - \Phi_M\left(\mathbf{q}_{\mathbf{M},\mathbf{R},\mathbf{1}-\alpha}; \mathbf{e}, \mathbf{R}\right)$ or, equivalently, using a central multivariate normal distribution after subtracting the vector of expected values from the quantiles: $1 - \Phi_M\left(\mathbf{q}_{\mathbf{M},\mathbf{R},\mathbf{1}-\alpha} - \mathbf{e}; \mathbf{0}, \mathbf{R}\right)$. A simulation study involving a variety of settings revealed that only a negligible difference in the second decimal position exists between the values of approximate and simulated power. In Table 8, the approximate power calculation is compared to simulated power (10,000 replications) for tests on increasing trend with nominal level $\alpha = 0.05$, using the add-1 adjustment. The expected values π_i for two different dose response shapes, which could be underlying the data in Table 2, were assumed, and power is calculated for balanced samples sizes $n_i = 10, 20, 40, 60$.

Table 8: Approximate and simulated power of tests for increasing trend ($\alpha = 0.05$) using add-1 adjustment

n_i	π_1	π_2	π_3	π_4	Approximate power	Simulated power
10	0.30	0.30	0.30	0.50	0.1763	0.1802
10	0.05	0.05	0.20	0.50	0.7224	0.7650
20	0.30	0.30	0.30	0.50	0.2863	0.3300
20	0.05	0.05	0.20	0.50	0.9592	0.9596
40	0.30	0.30	0.30	0.50	0.4868	0.4713
40	0.05	0.05	0.20	0.50	0.9996	0.9994
60	0.30	0.30	0.30	0.50	0.6484	0.6345
60	0.05	0.05	0.20	0.50	1.0000	1.0000

DOI: 10.2202/1557-4679.1180

Table 8 shows marked differences between the calculated and the simulated power for the smallest group wise sample size $n_i = 10$. For such small sample sizes, the approximate power calculation may show even larger deviations from the true power than the deviations displayed in Table 8, depending on the chosen parameter combination. However, taking into account the uncertainty in assuming $\pi_1, ..., \pi_I$, the method of approximate power calculation is a helpful tool in experimental design involving moderate sample sizes.

6 Evaluation of the examples

Applying the add-1 method to the chronic toxicity data presented in Table 1 led to the approximate simultaneous 0.95 confidence limits listed in Table 9. Since all three linear combinations were significantly larger than 0, we concluded that there is a significant increase in the mortality rate with increasing dosage. Pooling the 50 mg/kg and 100 mg/kg dose groups led to the most pronounced change in mortality rate compared to the untreated control group.

Table 9: Simultaneous 0.95 lower add-1 confidence limits for Williams-type contrasts of the proportions presented in Table 1.

Comparison	Contrast coefficients				Lower limit	Estimate
C1: Control vs. high	-1	0	0	1	0.069	0.300
C2: Control vs. medium & high	-1	0	0.5	0.5	0.078	0.250
C3: Control vs. all doses	-1	$0.\bar{3}$	$0.\bar{3}$	$0.\bar{3}$	0.014	0.150

Next, we evaluated the histopathological findings of the 28-day toxicity study in rats (Tab. 2). We applied the add-1 adjustment and calculated multiplicity-adjusted p-values. For the three Williams contrasts C1: $\pi_{250mg/kg} - \pi_{0mg/kg}$; C2: $(\pi_{250mg/kg} + \pi_{50mg/kg})/2 - \pi_{0mg/kg}$; and C3: $(\pi_{250mg} + \pi_{50mg} + \pi_{10mg})/3 - \pi_{0mg/kg}$, we obtain p-values $1.7e-07$, $7.4e-06$ and $2.0e-04$, respectively. The minimal p-value smaller than 0.05 reveals the presence of a dose-related trend, and because the smallest p-value was obtained for contrast C1, we concluded that an increase in the high 250 mg/kg dose dominated this trend. An alternative approach to evaluating this dataset consisted of employing a closed test to identify the minimum observed effect dose by performing Williams trend tests with stepwise omission of the highest remaining dose. In the second step, this approach already yielded a non-significant p-value of 0.052 for the trend test without the 250 mg/kg dose, i.e., 250 mg/kg was the minimum observed effect dose.

7 Discussion

This paper shows how trends among ordered binomial proportions can be detected without the strong assumption of linearity in settings with small or moderate sample size. Although the use of multiplicity-adjusted p-values is common, the use of simultaneous confidence limits allows further interpretation with respect to the effect size and shape of the dose-response relationship.

A simulation study showed that the computationally simple add-1 and add-2 adjustments perform better than the commonly used Wald limit. The methods proposed here fill the gap between methods that are appropriate for designs with large sample sizes (e.g., 100 per group) when central limit theorem holds for the Wald limit, and designs involving small sample sizes of 20 or less, which are not appropriate in trials with a binomial response due to insufficient power. A comparison of the different adjustments revealed that the add-1 adjustment provides a simple and acceptable solution for one-sided confidence limits. Advantages of the small sample adjustments investigated here are their simplicity and computational availability. Additionally, due to the simplicity of the methods, an approximate power calculation can be derived using previous results of Bretz and Hothorn (2002).

In toxicology, downturn effects at high doses may sometimes occur. The Williams test is relatively robust against slight non-monotonicity, because of its pooling properties, but may yield misleading results in the presence of significant downturns. When monotonicity is in doubt, modified Williams-type contrasts for downturn alternatives are available (Bretz and Hothorn, 2003). Combining their approach with the methods discussed in this article is straightforward, and involves replacing the Williams-type contrast matrix with the modified matrix proposed by Bretz and Hothorn (2003). In cases when even a relaxed monotonicity assumption appears inadequate, Dunnett-type comparisons for proportions may be applied instead (Holford, 1989, Piegorsch, 1991, Schaarschmidt et al., 2009).

When the underlying trend is, in fact, linear, the Williams test exhibits lower power compared to alternative methods. However, in the case of Gaussian variables, this loss in power is modest (Williams, 1971, Bretz, 2006), except when the sample size allocated in the control group is unusually small (Bretz, 2006).

Exact procedures are often recommended for binomial proportions and designs with small or moderate sample sizes. Such methods are available and discussed extensively for the one-parameter problem, or when only simple null- hypotheses of no effect are considered. However, when two or more parameters are considered simultaneously, and confidence intervals are of interest in addition to p-values, the use of exact methods becomes either controversial, even in relatively simple problems (Roehmel, 2005), or simply impossible due to the computational burden of inverting exact binomial tests. Furthermore, requiring an exact method

DOI: 10.2202/1557-4679.1180

means that the size of the test is smaller than or equal to α. Due to the discreteness of the binomial variable, and hence of the test statistics, exact tests have size smaller than α for almost all parameter settings, especially if sample sizes are small and proportions are close to 0 or 1 (e.g. Brown et al., 2001). As a consequence, the test (and the corresponding confidence interval) is more conservative than nominally required. However, applying a conservative test in cases when the alternative hypothesis describes the hazardousness of the compound is counterintuitive under the precautionary principle.

In this paper, we discussed situations, in which simple binomial proportions are assumed to characterize the toxicologically interesting effects. However, due to competing risks, the crude incidence rates may result in biased estimates of the effects of interest. As an example, consider an evaluation of tumor incidences in long-term carcinogenicity studies confounded by mortality without cause-of-death information. For such data, related procedures based on mortality-adjusted poly-k estimates are available (Schaarschmidt et al., 2008a). Related procedures for the evaluation of graded histopathological findings and differential blood counts may serve as a subject for future research.

The approach discussed above concerns a one-way layout, and does not allow to include covariate information. However, testing trends for Williams-type contrasts while correcting for covariates is straightforward in the generalized linear model with the binomial family, logit link, and a subsequent application of the methods described by Hothorn et al. (2008) for multiple contrasts. In this approach, the effect would be described in terms of odds ratios rather than differences of proportions. However, its small sample properties have not been examined so far; the necessity of applying adjustments and their available options for small sample sizes may provide additional topics for further research.

References

Agresti A, Caffo B. (2000) Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician* 54(4):280–288.

Agresti A, Coull A. (1998) Approximate is better than ”exact” for interval estimation of binomial proportions. *American Statistician* 52(2):119–126.

Armitage P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics* 11(3):375–386.

Bell DR, Clode S, Fan MQ, Fernandes A, Foster PMD, Jiang T, Loizou G, MacNicoll A, Miller BG, Rose M, Tran L, White S. (2007) Toxicity of 2,3,7,8-Tetrachlorodibenzo-p-dioxin in the Developing Male Wistar(Han)

Rat. II: Chronic Dosing Causes Developmental Delay. *Toxicological Sciences* 99(1):224–233.

Bretz F. (2006) An extension of the Williams trend test to general unbalanced linear models. *Computational Statistics & Data Analysis* 50(7):1735–1748.

Bretz F, Hothorn LA. (2002) Detecting dose-response using contrasts: asymptotic power and sample size determination for binomial data. *Statistics in Medicine* 21(22):3325–3335.

Bretz F, Hothorn LA. (2003) Statistical analysis of monotone or non-monotone dose-response data from in vitro toxicological assays. *ATLA-Alternatives to Laboratory Animals* 31:81–96.

Brown LD, Cai TT and DasGupta A. (2001) Interval estimation for a binomial proportion. *Statistical Science* 16:101–133.

Brown L, Li X. (2005) Confidence intervals for two sample binomial distribution. *Journal of Statistical Planning and Inference* 130(1-2):359–375.

Cai TT. (2005) One-sided confidence intervals in discrete distributions. *Journal of Statistical Planning and Inference* 131(1):63–88.

Carriere KC. (2001) How good is a normal approximation for rates and proportions of low incidence events? *Communications in Statistics - Simulation and Computation* 30(2):327–337.

Dunn OJ. (1964) Multiple Comparisons Using Rank Sums. *Technometrics* 6(3): 241–249.

Dunnett CW. (1955) A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 50(272):1096–1121.

Holford TR, Walter SD and Dunnett CW (1989). Simultaneous interval estimates of the odds ratio in studies with two or more comparisons. *Journal of Clinical Epidemiology* 42:427–434.

Hothorn LA. (1994) Multiple comparisons in long-term toxicity studies. *Environmental Health Perspectives* 102:33–38.

Hothorn T, Bretz F and Genz A. (2001) On multivariate t and Gauss probabilities in R. *R News* 1(2):27–29.

Hothorn T, Bretz F and Westfall P. (2008) Simultaneous Inference in General Parametric Models. *Biometrical Journal* 50:346–363.

Kandori H, Suzuki S, Asamoto M, Murasaki T, Mingxi T, Ogawa K, Shirai, T. (2005) Influence of atrazine administration and reduction of calorie intake on prostate carcinogenesis in probasin/SV40 T antigen transgenic rats. *Cancer Science* 96(4):221–226.

Leuraud K, Benichou J. (2006) A comparison of stratified and adjusted trend tests for binomial proportions. *Statistics in Medicine* 25(3):529–535.

DOI: 10.2202/1557-4679.1180

National Toxicology Program. (2009) *Description of NTP Study Types - Expanded overview.* http://ntp.niehs.nih.gov [5 May 2009].

OECD408 (1998) *Repeated Dose 90-Day Oral Toxicity Study in Rodents.* OECD Paris: adopted 21st September 1998.

Piegorsch WW. (1991) Multiple comparisons for analyzing dichotomous response. *Biometrics* 47(1):45–52.

Price RM, Bonett DG. (2004) An improved confidence interval for a linear function of binomial proportions. *Computational Statistics & Data Analysis* 45(3):449–456.

Roehmel J. (2005) Problems with existing procedures to calculate exact unconditional p-values for non-inferiority/superiority and confidence intervals for two binomials and how to resolve them. *Biometrical Journal* 47(1): 37–47.

Rucker G, Schwarzer G, Carpenter J, Olkin, I. (2009) Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine* 28(5):721–738.

Schaarschmidt F, Sill M, Hothorn LA. (2008a) Poly-k-trend tests for survival adjusted analysis of tumor rates formulated as approximate multiple contrast test. *Journal of Biopharmaceutical Statistics* 18(5):934–948.

Schaarschmidt F, Gerhard D, Sill M. (2008b) MCPAN: Multiple comparisons using normal approximation. R package version 1.1-7.

Schaarschmidt F, Biesheuvel EHE, Hothorn LA. (2009) Asymptotic simultaneous confidence intervals for many-to-one comparisons of binary proportions in randomized clinical trials. *Journal of Biopharmaceutical Statistics* 19(2):292–310.

Shirley E. (1977) Nonparametric equivalent of Williams test for contrasting increasing dose levels of a treatment. *Biometrics* 33(2): 386–389.

Shirley E. (2007) Correspondence: Tests for a simple tree order restriction with application to dose-response studies. *Applied Statistics* 56:493–497.

Tang ML, Ng HKT, Guo JH, Chan W, Chan BPS. (2006) Exact Cochran-Armitage trend tests: comparisons under different models. *Journal of Statistical Computation and Simulation* 76(10):847–859.

Williams DA. (1971) Test for differences between treatment means when several dose levels are compared with a zero control. *Biometrics* 27(1):103–117.

Williams DA. (1972) Comparison of several dose levels with a zero dose control. *Biometrics* 28(2):519–531.

Woo GH, Shibutani M, Ichiki T, Inoue K, Hirose M. (2007) A repeated 28-day oral dose toxicity study of nonylphenol in rats, based on the Enhanced OECD Test Guideline 407 for screening of endocrine-disrupting chemicals. *Archives of Toxicology* 81(2):77–88.