Received: 14 March 2019 | Revised: 23 January 2021 | Accepted: 28 January 2021

DOI: 10.1111/biom.13442

BIOMETRIC PRACTICE

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

The Tukey trend test: Multiplicity adjustment using multiple marginal models

Frank Schaarschmidt[1] | **Christian Ritz**[2] | **Ludwig A. Hothorn**[3,*]

[1] Department of Biostatistics, Institute of Cell Biology and Biophysics, Leibniz University Hannover, Hannover, Germany

[2] Department of Nutrition, Exercise and Sports, University of Copenhagen, Frederiksberg C, Denmark

[3] Institute of Biostatistics, Faculty of Natural Sciences, Leibniz University Hannover, Hannover, Germany

Correspondence
Frank Schaarschmidt, Department of Biostatistics, Institute of Cell Biology and Biophysics, Leibniz University Hannover, D-30419 Hannover, Germany.
Email: schaarschmidt@cell.uni-hannover.de

*Retired.

Abstract

In dose–response analysis, it is a challenge to choose appropriate linear or curvilinear shapes when considering multiple, differently scaled endpoints. It has been proposed to fit several marginal regression models that try sets of different transformations of the dose levels as explanatory variables for each endpoint. However, the multiple testing problem underlying this approach, involving correlated parameter estimates for the dose effect between and within endpoints, could only be adjusted heuristically. An asymptotic correction for multiple testing can be derived from the score functions of the marginal regression models. Based on a multivariate t-distribution, the correction provides a one-step adjustment of p-values that accounts for the correlation between estimates from different marginal models. The advantages of the proposed methodology are demonstrated through three example datasets, involving generalized linear models with differently scaled endpoints, differing covariates, and a mixed effect model and through simulation results. The methodology is implemented in an R package.

KEYWORDS

adjustment of p-values, dose–response, multiple endpoints, multivariate normal, toxicology

1 | INTRODUCTION

In applications of dose–response analysis in early phases of research, the shape or functional form of the dose–response relationship is usually unknown a priori. Often, experimental designs with only few dose levels and one control group are used in these situations. This is especially common in toxicological studies or early-phase clinical trials for proof of concept (Ting, 2006; Hothorn, 2016). Nonzero dose levels are often chosen using logarithmic dose escalation to cover a wide range, reflecting lack of prior knowledge about onset of effects. It may, however, be difficult to prespecify a dose–response model that can accommodate all possible trends, which such flexible designs allow.

A naive, yet common data-driven solution would be to try out different models and choose the best fitting one for subsequent inference. For instance, models corresponding to different data transformations as well as subsequent pairwise comparisons of dose levels may be considered. Unfortunately, this leads to inflation of the type I error rate. Another solution would be to define a certain model beforehand and then use it for hypothesis testing. No inflation of type I error would be the consequence, but at the cost of discarding adaptive strategies for choosing a model that would lead to more efficient inference. Such problems are further aggravated in toxicology because usually many endpoints are involved: endpoints may differ in scale as different toxicological effects of a single compound may be recorded in terms of binary data, count data, or metric

measurements taken from the same experimental units. The onset of dose effects may differ between endpoints, and, consequently, the shape of dose–response relationships may differ between endpoints. Also, relevant covariates may differ between endpoints (Hothorn, 2016).

In a 1985 *Biometrics* paper, Tukey, Ciminera, and Heyse proposed a solution in the form of a series of trend tests for a given single endpoint: Different transformations of the dose levels were applied, and for each transformation a linear regression model was fitted (Shirley, 1996). In other words, multiple marginal models were fitted to the same observations. This approach is sensitive to a wide range of dose–response shapes, and it can be expected to have high power in case of studies with small sample size because it uses the sparse parameterization of a linear regression model. We refer to this approach as the Tukey trend test. Based on power assessment for various dose–response shapes, it has been recommended over other procedures (Capizzi *et al.*, 1992; Aras *et al.*, 2011). As different marginal models were fitted to the same data in related parameterizations, the resulting estimates and test statistics were highly correlated, rendering *p*-value adjustment challenging as Bonferroni adjustment would be too conservative whereas no adjustment would lead to inflation of the type I error rate (Capizzi *et al.*, 1992). For application to multiple endpoints, Tukey *et al.* (1985) proposed to use a correction factor for the *p*-values that was derived in a heuristic manner as the actual joint distribution of parameter estimates from different marginal models was intractable.

Recently, it has become possible to approximate joint distributions of estimates from multiple marginal models. Indeed using such approximations is not a new idea: For a number of special cases in survival analysis and analysis of longitudinal data, results were obtained by Wei and collaborators in a series of papers (Wei and Lachin, 1984; Wei and Johnson, 1985; Wei *et al.*, 1989), their initial developments have been further generalized (Lin, 2005; So and Sham, 2011; Pipper *et al.*, 2012). These authors paved the way for a unified framework to obtain the asymptotic joint distributions of parameter estimates from multiple marginal models. This development provides a means to carry out simultaneous inference for differently scaled multiple endpoints, modeled by different classes of linear models, with different distributional assumptions, and possibly also different covariate adjustments, but without having to specify a joint statistical model. Pipper *et al.* (2012) provided a general implementation of the approach in the statistical environment **R** (R Core Team, 2017), which in the meanwhile has become part of the extension package *multcomp* (Hothorn *et al.*, 2008).

In this article, we develop appropriate asymptotically correct multiplicity adjustment for the Tukey trend test, replacing the heuristic correction proposed by Tukey *et al.* (1985). We extend the Tukey trend test to combine several regression-based tests with multiple contrasts. Furthermore, the Tukey trend test is extended to joint test procedures for multiple endpoints of generalized linear models and linear mixed models. A simulation study examines the proposed methodology in detail. The extended Tukey trend test is also applied to examples from toxicology. The methodology is implemented in the **R** package *tukeytrend*.

2 | MATERIALS AND METHODS

2.1 | The Tukey trend test

Consider a design with G dose levels, where the index $g = 1, \dots, G$ identifies increasing dose levels, $d_1 < d_2 < \cdots < d_G$ of a compound of interest. Usually there are several replicates for each dose level. Tukey *et al.* (1985) propose to test for a dose–response relation by fitting three marginal regression models with three different transformations of the original dose levels used as predictor. Denoted by $x_k^{(g)}$, the transformed value of dose level d_g under transformation $k = 1, \dots, K = 3$. We refer to these transformed values as dose scores.

The first dose score ($k = 1$) uses the untransformed dose levels: $x_1^{(g)} = d_g$. The second transformation ($k = 2$) is the index g of ordered dose levels d_g: $x_2^{(g)} = g$. Finally, the third transformation is the logarithm ($k = 3$), $x_3^{(g)} = \log d_g$ if $d_1 > 0$. For experiments including an untreated or vehicle control group, that is $d_1 = 0$, Tukey *et al.* (1985) proposed the following dose score for $g = 1$:

$$x_3^{(1)} = \log d_2 - \frac{d_2 - d_1}{d_3 - d_2}(\log d_3 - \log d_2) \qquad (1)$$

and to use the log-transformation for $g > 1$. We refer to these three dose scores as "ari" (arithmetic), "ord" (ordinal), and "arilog," respectively.

Let the experimental units have running index $n = 1, \dots, N$ with N denoting the total number of units in the experiment. Initially, we assume that one measurement, y_n say, of one endpoint is taken for the nth experimental unit. Fitting three marginal simple linear regression models of y_n versus the transformed dose levels x_{n1}, x_{n2}, x_{n3}, respectively, $y_n = \beta_{0k} + \beta_{1k} x_{nk} + \epsilon_{nk}$, yields estimates of the three slopes, $\widehat{\beta}_{1k}$ with corresponding standard errors $\widehat{\sigma}_{1k}$ ($k = 1, \dots, 3$).

Evaluation of the null hypotheses $H_0 : \beta_{11} = 0 \cap \beta_{12} = 0 \cap \beta_{13} = 0$ versus $H_A : \beta_{11} \neq 0 \cup \beta_{12} \neq 0 \cup \beta_{13} \neq 0$ with test statistics $t_k = \widehat{\beta}_{1k}/\widehat{\sigma}_{1k}$, each following the $t_{df=N-2}$ distribution under H_0, leads to an union-intersection test, requiring a multiplicity adjustment of the three marginal

p-values, p_k ($k = 1, \dots, 3$). Obviously, the three parameter estimates and thus the resulting test statistics are highly correlated, and Tukey *et al.* (1985) proposed to use the minimal p-value for conclusions concerning an overall trend because a Bonferroni correction $3 \cdot \min_k(p_k)$ or $2 \cdot \min_k(p_k)$ was considered too conservative due to the high correlations. Also, Capizzi *et al.* (1992) showed that using no multiplicity adjustment violates the type I error and proposed an adjustment using the trivariate t-distribution.

Moreover, Tukey *et al.* (1985) discussed the necessity to jointly test for the presence of a dose–response relationship in at least one of J multiple endpoints ($j = 1, \dots, J$) where for each endpoint the three marginal regression models are fitted. For assessing an overall trend in a group of J endpoints, they propose to adjust the minimal p-value among all p-values, p_{jk} ($j = 1, \dots, J, k = 1, \dots, 3$), using the approximation $1 - \{1 - \min_{j,k}(p_{jk})^{\sqrt{J}}\}$. Tukey *et al.* (1985) proposed the above procedure as an alternative to (1) multiple comparisons of dose levels to control (Dunnett, 1955) and (2) the Williams trend test (Williams, 1971).

Combining the approaches of Pipper *et al.* (2012) and Hothorn *et al.* (2008), the following section will extend the basic idea of Tukey *et al.* (1985) to formulate a union-intersection test for the joint test of multiple, possibly differently scaled endpoints. The heuristic adjustment of p-values suggested by Tukey *et al.* (1985) is replaced by a single-step adjustment that accounts for the empirical correlation structure of the parameter estimates between different regression models for the same endpoints and between the regression models for different endpoints. The set of dose scores is extended to include dummy-coded categorical dose variables. Thus, joint tests for a dose–response relation, using the regression approach and multiple contrast tests corresponding to Dunnett or Williams-type contrasts, can be performed.

2.2 | Extensions to multiple endpoints and categorical dose variables

Denote with $\mathbf{Y}$ an $(N \times J)$ matrix of response variables, where $n = 1, \dots, N$ is the index of observational units, and $j = 1, \dots, J$ is the index of J response variable of interest. The J variables $\mathbf{y}_j$ may be differently scaled (e.g., continuous or counts), but need to be observed for the same observational units $n = 1, \dots, N$.

For each endpoint $\mathbf{y}_j$, a number of marginal models K_j may be fit, with index $k = 1, \dots, K_j$. These models may differ with respect to the variables and the number of parameters that are used to model the dose–response relationship. They may differ with respect to their design matrices $\mathbf{X}_{jk}$ because some models may contain covariates that other models do not use. Denote the number of parameters in the jkth model with D_{jk}. The models fitted for differently scaled endpoints may belong to different model classes such as the linear model in Equation (2) or generalized linear models with different distributional assumptions and different link functions $g_{jk}()$ and linear predictors $\boldsymbol{\eta}_{j'}$ as in Equation (3):

$$\mathbf{y}_j = \mathbf{X}_{jk}\boldsymbol{\beta}_{jk} + \boldsymbol{\epsilon}_{jk}, \tag{2}$$

$$\boldsymbol{\eta}_{j'} = g_{j'k}(\mathbf{X}_{j'k}\boldsymbol{\beta}_{j'k}). \tag{3}$$

We now decompose the parameter vector $\boldsymbol{\beta}_{jk}$ of the jkth model into a subvector $\boldsymbol{\gamma}_{jk}$, corresponding to effects of covariates, sample stratifications, blocks, etc. and into a subvector $\boldsymbol{\delta}_{jk}$, containing the parameters of interest for inference with respect to the dose–response relation; the length of this subvector is B_{jk}.

$$\boldsymbol{\beta}_{jk}^T = \left(\boldsymbol{\gamma}_{jk}^T, \boldsymbol{\delta}_{jk}^T\right). \tag{4}$$

For some endpoint and/or models, $\boldsymbol{\delta}_{jk}$ may contain only one slope parameter, that is, $B_{jk} = 1$, whereas for other models the dose effect may be represented by dummy-coded differences to the reference level leading to $B_{jk} > 1$ parameters in $\boldsymbol{\delta}_{jk}$. The above framework is also applicable for linear and generalized linear mixed models by utilizing that estimation of fixed-effects parameters may be based on generalized least squares (Ritz *et al.*, 2017).

2.2.1 | Simultaneous inference for the joint parameter vector

Now, interest is in joint inference for all $M = \sum_{j=1}^{J}\sum_{k=1}^{K_j} B_{jk}$ parameters representing potential dose effects for the set of models defined above: Let

$$\boldsymbol{\theta}^T = \left(\boldsymbol{\delta}_{11}^T, \boldsymbol{\delta}_{12}^T, \dots, \boldsymbol{\delta}_{1K_1}^T, \boldsymbol{\delta}_{21}^T, \dots, \boldsymbol{\delta}_{JK_J}^T\right) \tag{5}$$

define the vector of stacked parameter vectors $\boldsymbol{\delta}_{jk}^T$ with inner order by k and outer order by j. For notational simplicity, we identify its elements by $m = 1, \dots, M$ in the following. Our interest is in testing hypotheses of the form

$$H_0 : \bigcap_{m=1}^{M} \theta_m = 0 \text{ versus } H_A : \bigcup_{m=1}^{M} \theta_m \neq 0, \tag{6}$$

or

$$H_0 : \bigcap_{m=1}^{M} \theta_m \leq 0 \text{ versus } H_A : \bigcup_{m=1}^{M} \theta_m > 0, \tag{7}$$

that is, performing an intersection union test, aiming at a one-step procedure that controls the familywise error rate (FWER) for the M individual hypotheses.

Marginally fitting the jkth model yields vectors with the parameter estimates $\widehat{\boldsymbol{\delta}}_{jk}$, the corresponding standard error estimates $\widehat{\boldsymbol{\sigma}}_{jk}$, and $(N \times B_{jk})$ matrices containing the standardized score contributions evaluated at the maximum likelihood (ML) estimates of the model parameters (Pipper *et al.*, 2012; Jensen and Ritz, 2015), denoted by $\widehat{\boldsymbol{\Psi}}_{jk}$. To define a joint test for the M parameters of interest, we stack in the order defined above and obtain: $\widehat{\boldsymbol{\theta}}^T = (\widehat{\boldsymbol{\delta}}_{11}^T, \widehat{\boldsymbol{\delta}}_{12}^T, \dots, \widehat{\boldsymbol{\delta}}_{JK_J}^T)$, $\widehat{\boldsymbol{\sigma}}^T = (\widehat{\boldsymbol{\sigma}}_{11}^T, \widehat{\boldsymbol{\sigma}}_{12}^T, \dots, \widehat{\boldsymbol{\sigma}}_{JK_J}^T)$, and $\widehat{\boldsymbol{\Psi}}^T = (\widehat{\boldsymbol{\Psi}}_{11}^T, \widehat{\boldsymbol{\Psi}}_{12}^T, \dots, \widehat{\boldsymbol{\Psi}}_{JK_J}^T)$. The test statistic corresponding to the mth element of $\boldsymbol{\theta}$, then is denoted

$$t_m = \frac{\widehat{\theta}_m}{\widehat{\sigma}_m}. \tag{8}$$

Note that the M estimators $\widehat{\theta}_m$ may be correlated, and thus test statistics in $\boldsymbol{t} = (t_1, \dots, t_M)$ will also be correlated. In the following, results of Pipper *et al.* (2012) will be used to account for the correlation among test statistics: For large sample sizes, the covariance among the estimators $\widehat{\boldsymbol{\theta}}$ obtained from multiple marginal models can be estimated from the row vectors $\widehat{\boldsymbol{\psi}}_n$ of $\widehat{\boldsymbol{\Psi}}$, $\widehat{\boldsymbol{V}} = 1/N \sum_{n=1}^{N} \widehat{\boldsymbol{\psi}}_n^T \widehat{\boldsymbol{\psi}}_n$. Standardizing $\widehat{\boldsymbol{V}}$ by its diagonal elements yields an asymptotic estimator of the correlation between the M test statistics of interest, $\widehat{\boldsymbol{R}}$.

For an overall decision with respect to the hypotheses in Equation (6), we may thus choose a critical value $z_{1-\alpha,2-sided,M,\widehat{\boldsymbol{R}}}$ as the equi-coordinate, two-sided quantile of an M-variate normal distribution with correlation $\widehat{\boldsymbol{R}}$. Let z_q, with index $q = 1, \dots, M$, denote the elements of an M-variate normal random vector $\boldsymbol{Z}_{M,\boldsymbol{0},\widehat{\boldsymbol{R}}}$ with expectation $\boldsymbol{0}$, and covariance $\widehat{\boldsymbol{R}}$. Then, the critical value $z_{1-\alpha,2-sided,M,\widehat{\boldsymbol{R}}}$ is chosen such that $P\{\max_{q=1,\dots,M}(|z_q|) \leq z_{1-\alpha,2-sided,M,\widehat{\boldsymbol{R}}}\} = 1-\alpha$. One can then reject the null hypotheses in Equation (6) if at least one test statistic exceeds the critical value, that is $|t_m| > z_{1-\alpha,2-sided,M,\widehat{\boldsymbol{R}}}, \in m = 1, \dots, M$. For one-sided hypotheses as in Equation (7), a one-sided equi-coordinate quantile, $z_{1-\alpha,1-sided,M,\widehat{\boldsymbol{R}}}$, is chosen to ensure that $P\{\max_{q=1,\dots,M}(z_q) \leq z_{1-\alpha,1-sided,M,\widehat{\boldsymbol{R}}}\} = 1-\alpha$, and the null can be rejected if at least one t_m exceeds $z_{1-\alpha,1-sided,M,\widehat{\boldsymbol{R}}}$.

Asymptotically, adjusted p-values, p_m, for each of the M hypotheses of interest can be computed based on the probabilities of the M-variate normal distribution with correlation $\widehat{\boldsymbol{R}}$. For two-sided hypotheses as in Equation (6), $p_m = P(\max_{q=1,\dots,M}(|z_q|) > |t_m|)$, and for one-sided hypotheses with positive trends in the alternatives, $p_m = P(\max_{q=1,\dots,M}(z_q) > t_m)$. For the computational details of multivariate normal quantiles and probabilities, we refer to Genz and Bretz (2009) and Genz *et al.* (2017).

2.2.2 | Heuristic small sample adjustment

In general, linear models with limited sample size N, univariate tests of δ_{jk} based on the standard normal distribution would lead to inflated type I errors while using the t-distribution with degree of freedom $\nu_{jk} = N - D_{jk}$ would be exact. In generalized linear models, improved small sample performance can be expected when using the t-distribution instead of standard normal distribution in marginal tests, at least when dispersion parameters are estimated from the data. In linear mixed models, tests for δ_{jk} would be based on the t-distribution with estimated denominator degrees of freedom (Kenward and Roger, 1997). It is thus natural to use p-values calculated from the multivariate t-distribution instead of the multivariate normal distribution above. Several marginal models may have the same residual degree of freedom, $\nu_{jk} = N - D_{jk}$. However, different marginal models may involve different number of covariates in γ_{jk}, may use different number of parameters to model the trend (one regression slope or several differences to control), or may differ in the estimated denominator degrees of freedom ν_{jk} due to differently sized variance components in linear mixed models. Then, the marginal parameter vectors $\boldsymbol{\delta_{jk}}$ would imply different degrees of freedom for the corresponding t-distributions. So far, computational methods for multivariate t-distributions with different marginal degrees of freedom are not available (Genz and Bretz, 2009). We thus propose to use the arithmetic mean of (estimated) degrees of freedom across models, $\overline{\nu}$, in a multivariate t-distribution $t_{1-\alpha,2-sided,\overline{\nu},M,\widehat{\boldsymbol{R}}}$ instead of the multivariate normal distribution above.

2.3 | Software

The methods to combine multiple marginal models and estimate the joint covariance matrix and corresponding correlation matrix are implemented in the function `mmm` in the R package `multcomp` (Hothorn *et al.*, 2008). For computing multivariate normal and t probabilities or critical values, these functions rely on the R package `mvtnorm` (Genz and Bretz, 2009; Genz *et al.*, 2017). The R package `tukeytrend` is available from the CRAN repository and provides wrapper functions for computing the dose scores, automatic updating of given model object of classes `lm`, `glm` (the package `stats`; R Core Team, 2017) as well as `merMod` (the package `lme4`; Bates *et al.*, 2015) and `lme` (the package `nlme`; Pinheiro *et al.*, 2017) with different dose scores. It also contains functions that simplify the combination of multiple marginal models with different dose scores and endpoints before doing the main computations using the

functionality in `multcomp`. Further, a vignette is part of the package `tukeytrend`, containing numerous additional examples for multiple primary endpoints, several glms, and mixed effect models.

2.4 | Remarks

2.4.1 | Redundant dose scores

In designs with dose levels being equidistantly spaced on log scale, for example, dose levels being powers of 2, the ordinal and log-transformed model have predictors and thus slope estimates that are identical up to a rescaling factor. In this case, the estimated correlation between the two marginal dose parameter estimates is 1. The methods of computing multivariate normal or t probabilities (Genz and Bretz, 2009) can deal with this case: Jointly considering such two identical models does not lead to any multiplicity correction.

2.4.2 | Additional shift parameter for zero dose level in arilog dose score

In case an untreated control $d_1 = 0$ is log-transformed using the extrapolation in Equation (1), the resulting transformed value may still deviate substantially from an otherwise linear relation between dose and response. In such cases, it is possible to add further transformations k (see, e.g., Ekwaru and Veugelers, 2018). For example, these may differ from the "arilog" transformation only by a factor ($f_k > 0$) that modifies the amount of shifting of the first dose score relative to second- and third-ordered dose levels:

$$x_k^{(1)} = \log d_2 - f_k \frac{d_2 - d_1}{d_3 - d_2}(\log d_3 - \log d_2) \text{ for } g = 1. \quad (9)$$

The above framework allows to globally test on the presence of a trend while correcting for the multiple testing problem of trying different amounts of shifting of the first dose level. The vignette of the package `tukeytrend` contains examples illustrating these modified versions of the "arilog" dose scores.

2.5 | Simulation study

To assess the small sample performance of the multiplicity adjustment based on the empirical correlation matrix $\widehat{\boldsymbol{R}}$, we ran several simulation studies under the global null hypothesis of no trend. For single endpoints, we investigated: (1) A linear model with homoscedastic Gaussian residuals, with 4, 8, or 12 dose levels and balanced sample size ranging from 2 to 100 per dose level, considering different sets of hypotheses ($K = 3$ regression slopes with added Dunnett contrasts or multiple contrast tests, and multiple tests of different dose shifts); (2) A generalized linear model for overdispersed binomial data; (3) A simple linear mixed effect model with random intercepts to account for imbalanced subsampling. For the latter two simulations, $G = 4$ dose levels and different sets of hypotheses were simulated.

We used copulas, as implemented in the R package `copula` (Hofert *et al.*, 2018) to simulate multivariate normal, binomial, and Poisson data, to be fitted by linear models, and generalized linear models with logit and log link functions, respectively. To cover differently scaled endpoints, correlated multivariate datasets with equal numbers of endpoints from these three distributions were simulated. The number of endpoints was varied from 2 to 20, the sample size per dose level ranged from 5 to 100, and the correlation parameter was set 0, 0.5, and 0.9. Detailed descriptions of the simulation scenarios and results are available as Supporting Information.

Results indicate the control of the FWER when sample sizes per dose level are 20 or larger. Clear inflation of the FWER is observed for situations, where the number of tested hypotheses M is close to or higher than the number of observations per marginal model. All simulation results involving generalized linear models for count data (single overdispersed binomial endpoint, multivariate data including binomial and Poisson endpoints) indicate that the proposed method is conservative for low sample sizes, even when including high numbers of endpoints.

3 | EXAMPLES

3.1 | Tukey *et al.* (1985) revisited: Serum albumin in dogs

Tukey *et al.* (1985) present an example where $N = 36$ dogs in $G = 5$ treatment groups (with dose levels $d_1, \ldots, d_G = 0, 0.25, 1, 4, 16$). The only endpoint is the serum albumin concentration. The control group $d_1 = 0$ consisted of 12 dogs, while the treated groups consisted of six dogs. The dataset (Web Figure A9) is simulated to closely reproduce the summary statistics published by Tukey *et al.* (1985, Table 1 therein). Table 1 shows estimated slopes, test statistics, unadjusted p-values, and adjusted p-values obtained from our approach, using the three-variate t-distribution with $df = 34$. The adjustment depends on the empirical correlation of the three test statistics, which are 0.772, 0.843, and 0.986 for the pairs (t_1, t_2), (t_1, t_3), and (t_2, t_3), respectively. The smallest unadjusted p-value follows from the arithmetic-logarithmic dose scores,

TABLE 1 Estimated slopes ($\widehat{\beta}_{1k}$), standard errors ($\widehat{\sigma}_{1k}$) (s.e.), test statistics (t), as well as raw and adjusted p-values for the serum albumin data for arithmetic, ordinal, and arithmetic–logarithmic dose scores

Dose score	Slope	s.e.	t	p(raw)	p(adj)
ari	0.013	0.006	2.198	0.0348	0.0579
ord	0.067	0.022	3.096	0.0039	0.0068
arilog	0.058	0.018	3.156	0.0033	0.0058

$p_3 = 0.0033$, and the corresponding adjusted p-value is increased by slightly less than factor 2 as conjectured by Tukey *et al.* (1985).

3.2 | Multiple, differently scaled endpoints with different covariates: Litter weight and size

A total of $N = 74$ rat dams were randomized to four groups with dose levels $(d_1, \ldots, d_4 = 0, 5, 50, 500)$, and sample size 20, 19, 18, and 17, respectively. For each dam, $J = 2$ endpoints were recorded: litter weight ($j = 1$), possibly depending on the gestational time as a covariate, and the number of pups per litter($j = 2$) that may be modeled as overdispersed Poisson (see Figure 1).

We applied our extension of the Tukey trend test to each of three basic models, leading to a total of nine marginal models: Litter weight ($\boldsymbol{y}_1$) was described by linear models without any covariates, assuming independent Gaussian errors in each model,

$$\boldsymbol{y}_1 = \boldsymbol{X}_{1k}\boldsymbol{\beta}_{1k} + \boldsymbol{\epsilon}_{1k}, \tag{10}$$

where $\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \boldsymbol{X}_{13}$ are the $(N \times 2)$ design matrices corresponding to arithmetic, ordinal, and arithmetic–logarithmic dose scores, respectively (intercept and dose scores columns). In the parameter vectors, $\boldsymbol{\beta}_{1k}^T = (\beta_{1k0}, \beta_{1k1})$, the β_{1k1}, $k = 1, \ldots, 3$ are the regression slopes that are of interest for testing the dose–response relation.

Additionally, litter weight was described by three Gaussian linear models including the covariates gestational time and number of pups per litter as third and fourth columns in the $(N \times 4)$ design matrices $\boldsymbol{X}_{1k}, k = 4, 5, 6$. In the corresponding parameter vectors $\boldsymbol{\beta}_{1k}^T = (\beta_{1k0}, \beta_{1k1}, \beta_{1k2}, \beta_{1k3})$, β_{1k1} model the dose–response relation with three different dose scores ($k = 4, 5, 6$), while adjusting for the two covariates via β_{1k2} and β_{1k3}.

Finally, three generalized linear models with log-link were used to model the number of pups per litter, $\boldsymbol{y}_2$. The mean and variance of $y_{n2}, n = 1, \ldots, N$ were assumed to depend on μ_{nk} and on the scale parameter ϕ that accounts for the potential extra-Poisson variation, such that

$$\log(\boldsymbol{\mu}_k) = \boldsymbol{\eta}_k, \tag{11}$$

$$\boldsymbol{\eta}_k = \boldsymbol{X}_{2k}\boldsymbol{\beta}_{2k}. \tag{12}$$

The design matrices $\boldsymbol{X}_{2k}$ have the same structure as $\boldsymbol{X}_{1k}$, $k = 1, \ldots, 3$, such that β_{2k1}, $k = 1, \ldots 3$, are the regression slopes of interest.

From Table 2, it is obvious that one cannot conclude for the presence of a dose–response relation with respect to litter weight or litter size. To illustrate the effect of multiplicity correction by using the multivariate t-distribution, Web Figure A10 shows the estimated correlation matrix, $\widehat{\boldsymbol{R}}$: Correlations between test statistics referring to the same endpoint show correlations between 0.81 and 0.96, correlations between parameters referring to different endpoints range between 0.1 and 0.46, and the corresponding multivariate t-quantile is 2.54. Ignoring the multiple testing problem would lead to a critical value of $t_{1-0.05/2, df=71} = 1.99$, while

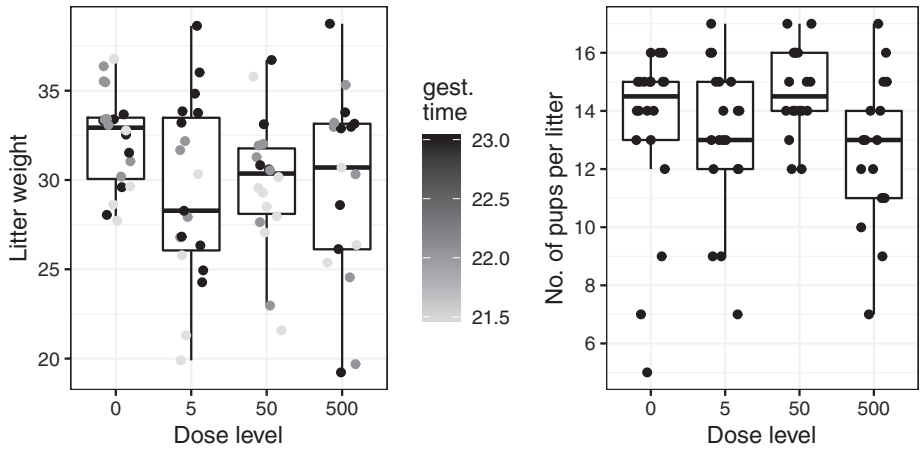

FIGURE 1 The left panel shows litter weights for each dam depending on the dose level, gray scale indicates different gestational time, and a potential covariate for litter weight. The right panel shows the number of pups per litter depending on the dose level

TABLE 2 Estimated slope parameters, standard errors (s.e.), test statistics (t), and adjusted p-values for nine marginal models analyzing litter weights without covariate linear model (LM) and with covariate adjustment (LM+cov) and litter size generalized linear model (GLM)

Model	Score	β_{jk1}	Estimate	s.e.	t	p(adj)
LM	"ari"	β_{111}	−0.002	0.003	−0.818	0.8273
LM	"ord"	β_{121}	−0.773	0.454	−1.703	0.2808
LM	"arilog"	β_{131}	−0.297	0.263	−1.128	0.6264
LM+cov	"ari"	β_{141}	−0.002	0.002	−0.777	0.8507
LM+cov	"ord"	β_{151}	−0.729	0.424	−1.717	0.2741
LM+cov	"arilog"	β_{161}	−0.256	0.247	−1.036	0.6887
GLM	"ari"	β_{211}	−0.000	0.000	−1.477	0.4010
GLM	"ord"	β_{221}	−0.006	0.020	−0.317	0.9962
GLM	"arilog"	β_{231}	−0.005	0.012	−0.449	0.9790

a Bonferroni correction would lead to $t_{1-0.05/(2\cdot9),df=71} = 2.86$.

3.3 | Linear mixed effect model with multiple laboratories as simple random factor

In an immunotoxicity study, the liver weights were assessed for $N = 316$ female rats in $G = 4$ dose levels 0, 3, 30, and 100 (Hothorn, 2003). The study comprised eight laboratories, each with initially 10 female rats per dose level. Drop out of single rats lead to a slightly unbalanced design (Web Figure A11). To account for the variation between the eight labs ($h = 1, \ldots, H$) and possibly different lab-specific deviation from overall trend, we fitted linear mixed models (using restricted maximum likelihood (REML)). Arithmetic, ordinal, and logarithmic dose scores were included as fixed effects in $\boldsymbol{X}_{1k}\boldsymbol{\beta}_{1k}$. The random effects included in $\boldsymbol{Z}\boldsymbol{u}_{1k}$ comprise overall deviations of laboratories and the dose-level–specific deviation within each laboratory, such that the residual variance is the variance of the 7–10 liver weights within each dose level g and lab h, $\boldsymbol{\epsilon}_{1k}$:

$$\boldsymbol{y}_1 = \boldsymbol{X}_{1k}\boldsymbol{\beta}_{1k} + \boldsymbol{Z}\boldsymbol{u}_{1k} + \boldsymbol{\epsilon}_{1k}, \tag{13}$$

where $\boldsymbol{y}_1$ contains the log-transformed liver weights. Results of applying our extension of the Tukey trend test are shown in Table 3. Corresponding denominator degrees of freedom according to Kenward–Roger are 23.15, 23.12, 23.15, estimated correlations between the three test statistics were 0.89, 0.87, and 0.95, and two-sided 0.95 quantile of the multivariate t-distribution is 2.315.

TABLE 3 Estimated slope parameters, standard errors (s.e.), test statistic (t), and adjusted p-values from three mixed model fits with arithmetic, ordinal and arithmetic–logarithmic dose scores for the immunotoxicity study on liver weights

Dose score	Slope	s.e.	t	p-value
"ari"	0.0042	0.00035	12.07	<0.0001
"ord"	0.1548	0.01125	13.76	<0.0001
"arilog"	0.1127	0.00582	19.36	<0.0001

4 | DISCUSSION

This paper provides an asymptotically correct and theoretically justified multiplicity adjustment for the trend test, which was proposed more than three decades ago by Tukey *et al.* (1985). The approach accounts for correlations between parameter estimates and test statistics from different marginally fitted models by using the methodology proposed by Pipper *et al.* (2012). The method can be applied on generalized linear and linear mixed models, extending the work by Quan and Capizzi (1999), and it allows simultaneous inference for the presence of a trend in multiple differently scaled endpoints that may be subject to different covariate adjustments. The original framework of Tukey *et al.* (1985) has been extended to joint tests of several regression slopes for different dose scores and for arbitrary multiple contrasts of means derived from the same dose variable. The framework is made available through an R package, including a manual (vignette) with many additional worked examples.

Compared to other approaches on dose–response analysis under model uncertainty, we focus on aspects relevant to toxicology. Establishing a dose–response relationship was of primary interest, while dose estimation or prediction of response is not. We focused on a scenario where the number of dose levels, the sample size is limited, and the interest lies in multiple, differently scaled endpoints and nonmonotone trends are a plausible complication: Joint testing of regression models and contrasts between means enables using sparse parameterization of regression models when the trend is monotone, while being robust against downturn effects or nonmonotone trends using appropriate contrasts.

Additional, prespecified transformations of the dose variable may be included in addition to three transformations proposed by Tukey *et al.* (1985). One application could be in epidemiology where exposure may be measured as a quantitative variable, but analyzed as a categorical variable: The "optimal" number of exposure categories and the position of the break points between them will be unclear. The proposed framework allows for simultaneous test of regression slopes and several definitions of exposure

categories, which may differ in the number and position of the breakpoints.

Alternative approaches to test for a dose–response under model uncertainty have been proposed, either using multiple contrast tests as in the MCPmod approach (e.g., Bretz *et al.*, 2005; Bornkamp *et al.*, 2009; Pinheiro *et al.*, 2014; Gould, 2019) or likelihood ratio tests for multiple nonlinear dose–response models (Gutjahr and Bornkamp, 2017) or max-T tests over several transformation of the covariates, such as multiple Box–Cox models (Liquet and Riou, 2019). Notably, the MCPmod approach considered different nonlinear models (using a priori defined guesses for the model parameters); these models were represented by multiple contrasts of the means of the dose groups. This approach has been extended to models involving covariates and assuming heterogeneous variances. There exist also extensions to generalized linear models and hierarchical mixed effect models (Pinheiro *et al.*, 2014). Further extensions developed by Dette *et al.* (2015) and Gutjahr and Bornkamp (2017) avoided the a priori definition of the nonlinear model parameters by using a likelihood ratio test instead of the maximum test of contrasts between dose groups. However, in most cases, these approaches relied on the assumption of independent, Gaussian errors. Moreover, they focused on single endpoints and, to our knowledge, they have not been extended to handle multiple differently scaled endpoints. These assumptions restrict a general application in fields like toxicology or epidemiology, where binomial or Poisson-type count data of pathological symptoms, lesions, or malformations are important response variables.

The MCPmod approach may indeed be extended using the proposed general framework for the Tukey trend test. The important step is to define contrasts in the parameters in θ in a matrix $\boldsymbol{C}$ with M columns. These contrasts would then represent both different nonlinear models, depending on guessed parameters of these models, and the original dose scores (see Example 3.1 in the Supporting Information). Based on $\widehat{\theta}$ and $\widehat{\boldsymbol{V}}$ as previously defined, tests for these linear combinations can be performed using $\widetilde{\theta} = \boldsymbol{C}\widehat{\theta}$ and $\widetilde{\boldsymbol{V}} = \boldsymbol{C}\widehat{\boldsymbol{V}}\boldsymbol{C}^T$. The similarity between MCPmod and parameter-dependent transformed nonlinear models into linear models was demonstrated for continuous response (Thomas, 2017).

ACKNOWLEDGEMENTS

Open access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper (Examples 3.1, 3.3) are available in the Supporting Information of this article. The dataset for Example 3.2 (Hothorn *et al.*, 2020) is openly available from CRAN at https://CRAN.R-project.org/package=multcomp.

ORCID

Frank Schaarschmidt https://orcid.org/0000-0002-6599-3803
Christian Ritz https://orcid.org/0000-0002-5095-0624
Ludwig A. Hothorn https://orcid.org/0000-0002-5162-1486

REFERENCES

Aras, G., Xue, A. and Liu, T. (2011) Tukey's contrast test versus two-sample test in a dose-response clinical trial. *Statistics in Biopharmaceutical Research*, 3, 31–39.

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48.

Bornkamp, B., Pinheiro, J.C. and Bretz, F. (2009) MCPMod: an R package for the design and analysis of dose-finding studies. *Journal of Statistical Software*, 29, 1–23.

Bretz, F., Pinheiro, J.C. and Bransom, M. (2005) Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, 61, 738–748.

Capizzi, T., Survill, T., Heyse, J. and Malani, H. (1992) An empirical and simulated comparison of some tests for detecting progressiveness of response with increasing doses of a compound. *Biometrical Journal*, 34, 275–289.

Dette, H., Titoff, S., Volgushev, S. and Bretz, F. (2015) Dose response signal detection under model uncertainty. *Biometrics*, 71, 996–1008.

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

Ekwaru, J.P. and Veugelers, P.J. (2018) The overlooked importance of constants added in log transformation of independent variables with zero values: a proposed approach for determining an optimal constant. *Statistics in Biopharmaceutical Research*, 10, 26–29.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics, Vol. 195.* Heidelberg, Germany: Springer-Verlag.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F. and Scheipl, F. (2017). mvtnorm: multivariate normal and *t* distributions. R package version 1.0-5.

Gould, A.L. (2019) BMA-Mod: a Bayesian model averaging strategy for determining dose-response relationships in the presence of model uncertainty. *Biometrical Journal*, 61, 1141–1159.

Gutjahr, G. and Bornkamp, B. (2017) Likelihood ratio tests for a dose-response effect using multiple nonlinear regression models. *Biometrics*, 73, 197–205.

Hofert, M., Kojadinovic, I., Maechler, M. and Yan, J. (2018). copula: multivariate dependence with copulas. R package version 0.999-19.1.

Hothorn, L.A. (2003) Statistics with interlaboratory in vitro toxicological studies. *ATLA-Alternative to Laboratory Animals*, 31, 43–63.

Hothorn, L.A. (2016). *Statistics in Toxicology Using R*. Boca Raton, FL: CRC Press.

Hothorn, T., Bretz, F. and Westfall, P. (2008) Simultaneous inference in general parametric models. *Biometrical Journal*, 50, 346–363.

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R.M., Schuetzenmeister, A. and Scheibe, S. (2020) Litter weights data set. CRAN; R package multcomp; version 1.4-15.

Jensen, S.M. and Ritz, C. (2015) Simultaneous inference for model averaging of derived parameters. *Risk Analysis*, 35, 68–76.

Kenward, M.G. and Roger, J.H. (1997) Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983–997.

Lin, D.Y. (2005) An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, 21, 781–787.

Liquet, B. and Riou, J. (2019) Cpmcglm: an R package for p-value adjustment when looking for an optimal transformation of a single explanatory variable in generalized linear models. *BMC Medical Research Methodology*, 19, 79.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team (2017). nlme: linear and nonlinear mixed effects models. R package version 3.1-131.

Pinheiro, J., Bornkamp, B., Glimm, E. and Bretz, F. (2014) Model-based dose finding under model uncertainty using general parametric models. *Statistics in Medicine*, 33, 1646–1661.

Pipper, C.B., Ritz, C. and Bisgaard, H. (2012) A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. *Journal of the Royal Statistical Society Series C – Applied Statistics*, 61, 315–326.

Quan, H. and Capizzi, T. (1999) Adjusted regression trend test for a multicenter clinical trial. *Biometrics*, 55, 460–462.

R Core Team (2017) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Ritz, C., Laursen, R.P. and Damsgaard, C.T. (2017) Simultaneous inference for multilevel linear mixed models with an application to a large-scale school meal study. *Journal of the Royal Statistical Society Series C — Applied Statistics*, 66, 295–311.

Shirley, E. (1996) A literature review of statistical methods for the analysis of general toxicology data. In: Morgan, B.J.T. (Ed.) *Statistics in Toxicology*. Oxford, UK: Oxford University Press.

So, H.-C. and Sham, P.C. (2011) Robust association tests under different genetic models, allowing for binary or quantitative traits and covariates. *Behavior Genetics*, 41, 768–775.

Thomas, N. (2017) Understanding mcp-mod dose finding as a method based on linear regression. *Statistics in Medicine*, 36, 4401–4413.

Ting, N. (2006). *Dose Finding in Drug Development*. New York: Springer-Verlag.

Tukey, J.W., Ciminera, J.L. and Heyse, J.F. (1985) Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics*, 41, 295–301.

Wei, L.J. and Johnson, W.E. (1985) Combining dependent tests with incomplete repeated measurements. *Biometrika*, 72, 359–364.

Wei, L.J. and Lachin, J.M. (1984) Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal American Statistical Association*, 79, 653–661.

Wei, L.J., Lin, D.Y. and Weissfeld, L. (1989) Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal American Statistical Association*, 84, 1065–1073.

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

SUPPORTING INFORMATION

Web Appendices and Figures referenced in Sections 2.5, 3.1, 3.2, 3.3, and 4 are available with this paper at the Biometrics website on Wiley Online Library. R Code to reproduce the examples may be found online in the Supporting Information Section. The code primarily relies on the R package `tukeytrend` (https://CRAN.R-project.org/package=tukeytrend), and further packages listed in the Supporting Information.

How to cite this article: Schaarschmidt F, Ritz C, Hothorn LA. The Tukey trend test: Multiplicity adjustment using multiple marginal models. *Biometrics*. 2022;78:789–797. https://doi.org/10.1111/biom.13442