Testing marginal homogeneity against stochastic order in multivariate ordinal data

B. Klingenberg[1,*], **A. Solari**[2], **L. Salmaso**[3], and **F. Pesarin**[2],

[1]Department of Mathematics and Statistics, Williams College

Williamstown, MA 01267, U.S.A.

[2] Department of Statistics, University of Padova,

Via Cesare Battisti, 241/243, 35121 Padova, Italy

[3] Department of Management and Engineering, University of Padova,

Stradella San Nicola, 3, 36100 Vicenza, Italy

[*] *email:*bklingen@williams.edu

SUMMARY: Many assessment instruments used in the evaluation of toxicity, safety, pain or disease progression consider multiple ordinal endpoints to fully capture the presence and severity of treatment effects. Contingency tables underlying these correlated responses are often sparse and imbalanced, rendering asymptotic results unreliable or model fitting prohibitively complex without overly simplistic assumptions on the marginal and joint distribution. Instead of a modeling approach, we look at stochastic order and marginal inhomogeneity as an expression or manifestation of a treatment effect under much weaker assumptions. Often, endpoints are grouped together into physiological domains or by the body function they describe. We derive tests based on these subgroups which might supplement or replace the individual endpoint analysis because they are more powerful. The permutation or bootstrap distribution is used throughout to obtain global, subgroup and individual significance levels as they naturally incorporate the correlation among endpoints. We provide a theorem that establishes a connection between marginal homogeneity and the stronger exchangeability assumption under the permutation approach. Multiplicity adjustments for the individual

endpoints are obtained via step-down procedures, while subgroup significance levels are adjusted via the full closed testing procedure. The proposed methodology is illustrated using a collection of 25 correlated ordinal endpoints, grouped into 6 domains, to evaluate toxicity of a chemical compound.

KEY WORDS: Adverse events; Correlated ordinal observations; Drug safety; Multiple endpoints; Stochastic order

1. Introduction

Assessing the risks and benefits of a treatment is more comprehensive and sensitive when several variables are considered simultaneously rather than ignoring some or analyzing them separately. For instance, the large number and variety of possible manifestations of a dose effect in a subject's clinical response usually necessitates that several endpoints are observed jointly as not to miss out on any crucial effects or interactions. One such collection of endpoints designed to evaluate neurophysiological effects in animals after exposure to a toxin is a biological screening assay composed of roughly 25 endpoints, termed the Functional Observational Battery (FOB; Moser, 1989; see also the United States Environmental Protection Agency's guideline 40CFR 798.6050). The FOB tries to group endpoints by a common domain, each domain describing a possibly distinct neurological function. Table 1 shows the structure of the FOB and displays data from a study designed to measure the presence and severity of neurotoxic effects in animals after exposure to perchlorethylene (PERC), a chemical used in the dry cleaning industry (and suspected of leading to an unusual concentration of leukemia cases in the city of Woburn, MA, see Lagakos, Wessen and Zelen, 1986). In the study, a total of 40 animals were randomly assigned to either placebo or 4 exposure levels of PERC, with 8 animals per dose group, and the FOB was administered at several time intervals. Each animal was evaluated on 25 endpoints, classified into six domains, and the data were converted to a scale from 1 to 4, where a score of 1 indicated absence of the corresponding adverse effect and a score of 4 denoted the most severe reaction (Moser et al., 1995). Table 1 summarizes the result of the FOB at the time peak effect, 4 hours into exposure, for the placebo and the 1.5 gram per kg body weight exposure level.

[Table 1 about here.]

Many similar comprehensive batteries of tests exist in other fields, such as the presence and severity of several adverse events (which are also usually grouped by body function according

to some dictionary) in drug safety studies or various assessment scales for medical diagnosis of pain or diseases such as Alzheimer's or Parkinson's. The statistical analysis of these data structures is challenging, firstly because underlying contingency tables for the multivariate categorical responses are very sparse and imbalanced and secondly because associations of various degrees among the endpoints may mask or enhance effects if not properly taken into account. One simple approach treats the observed scores (and the resulting mean score) as Gaussian, opening up the tool-box for normal-based theory and methods. This might be appropriate if both, the number of categories and the sample size are large, which is not the case for the types of multivariate data considered here.

Han et al. (2004) used data from the FOB to illustrate methods for testing a dose-response relationship with multivariate *binary* responses, addressing some of the challenges via an exact, conditional analysis. They transformed the ordinal responses into binary ones (absence/presence of adverse effect), thereby losing information on severity, and assumed equal correlation among endpoints within a domain coupled with independence of endpoints across different domains. The correlated binary endpoints within a domain were modeled using the distribution of Molenberghs and Ryan (1999), conditioning out two-way interactions and setting higher ones equal to 0. However, associations to various degrees exist both within and across domains, partly because the grouping is often rather subjective and neurotoxic effects defy a simple categorization (Baird et al., 1997). Finally, they assumed a common trend (across the 5 exposure levels) in the marginal probabilities for *all* endpoints within a domain, and used the significance of this trend to assess potential domain effects.

In this article we only focus on the two-sample case (the multiple sample case is subject of a follow-up paper), but treat the data as multivariate ordinal with general correlation structure that we don't model explicitly. We analyze all endpoints simultaneously and not one domain at a time. Moreover, we focus on stochastic order and marginal inhomogeneity among the

response vectors as an expression or manifestation of a dose effect rather than explicitly modeling marginal probabilities under overly simplistic assumptions on both, the marginal (such as a common shift) and joint (such as equal correlation within and independence across domains) distributions. For the univariate case, see Cohen and Sackrowitz (2000) for a treatment of the stochastic order hypothesis. In Section 2 we specify and discuss hypotheses of interest and develop test statistics based on assigning scores to the ordinal outcomes. All inference is based on the permutation distribution (Section 3) that naturally handles the associations in the multiple endpoints and provides exact significance levels. We present a theorem showing that the permutation approach (which assumes exchangeability of entire profiles) is valid for testing marginal homogeneity under a prior assumption of stochastic order. This assumption is usually implicitly made in the multivariate normal case, where one considers a shift in the marginal means but assumes identical covariance matrices, which leads to stochastic order (Müller, 2001). In particular, in the literature on multiple endpoints, O'Brien (1984) and Pocock, Geller and Tsiatis (1987) consider a (common) shift in the marginal means to describe a treatment effect, but assume identical distributions under the null (when the shift is zero). Similarly, Westfall and Young (1989) and Troendle (2005), both using resampling procedures for analyzing multivariate binary endpoints assume identical joint distributions under the null, although the hypothesis of interest focuses solely on the margins. In Section 4, we analyze the FOB data, investigating toxicity at the endpoint and domain level with the methods developed in Section 3. We present an alternative bootstrap approach in case the prior assumption on stochastic ordering is implausible in Section 5, together with simulation results. Finally, Section 6 discusses the validity of both approaches.

2. Two-sample multivariate ordinal data

In this article we discuss the case of comparing two treatments (doses) based on observing for each subject k ordinal variables with possibly different number of categories. Let $\boldsymbol{Y}_i =$

$(Y_{i1}, \ldots, Y_{ik})^t$ be the multivariate response at dose $i = 1, 2$, where Y_{ih} is ordinal with $c_h \geq 2$ categories, $h = 1, \ldots, k$. Suppose we have a total of $n_1 + n_2$ subjects randomly assigned to the two doses, such that $\boldsymbol{Y}_{11}, \ldots, \boldsymbol{Y}_{1n_1}$ are n_1 i.i.d. observations from a distribution $\pi_1(j_1, \ldots, j_k)$ and, independently, $\boldsymbol{Y}_{21}, \ldots, \boldsymbol{Y}_{2n_2}$ are n_2 i.i.d. observations from a distribution $\pi_2(j_1, \ldots, j_k)$, where $\pi_i(j_1, \ldots, j_k)$ denotes the joint probabilities $\Pr(Y_{i1} = j_1, \ldots, Y_{ik} = j_k)$, $j_h \in \{1, \ldots, c_h\}$ at dose i. To investigate a possible dose effect, we initially set up the null hypothesis $H_0 : \boldsymbol{Y}_1 \stackrel{d}{=} \boldsymbol{Y}_2$, where '$\stackrel{d}{=}$' means 'equal in distribution', i.e. $\pi_1(j_1, \ldots, j_k) = \pi_2(j_1, \ldots, j_k)$ for all $(j_1, \ldots, j_k) \in \{1, \ldots, c_1\} \times \ldots \times \{1, \ldots, c_k\}$, against the one-sided alternative that the $\boldsymbol{Y}_2$ distribution is stochastically larger and not equal to the $\boldsymbol{Y}_1$ distribution, $H_1 : \boldsymbol{Y}_2 \stackrel{st}{\gneq} \boldsymbol{Y}_1$. The random vector $\boldsymbol{Y}_2$ is said to be stochastically larger than $\boldsymbol{Y}_1$, written $\boldsymbol{Y}_2 \stackrel{st}{\geq} \boldsymbol{Y}_1$, if $E[g(\boldsymbol{Y}_2)] \geq E[g(\boldsymbol{Y}_1)]$ for all functions $g : \mathbb{R}^k \to \mathbb{R}$ that are increasing in each argument and have finite expectations (Marshall and Olkin, 1979). Note that $\boldsymbol{Y}_2 \stackrel{st}{\geq} \boldsymbol{Y}_1$ implies both, order of the cdf's $\Pr(Y_{11} \leq j_1, \ldots, Y_{1k} \leq j_k) \geq \Pr(Y_{21} \leq j_1, \ldots, Y_{2k} \leq j_k)$, i.e., smaller or equal values are more likely to occur under the first dose, and order of the survival functions $\Pr(Y_{11} > j_1, \ldots, Y_{1k} > j_k) \leq \Pr(Y_{21} > j_1, \ldots, Y_{2k} > j_k)$, i.e., larger values are more likely to occur under the second dose. As is natural, we assume two independent multinomial distributions $(n_i, \{\pi_i(j_1, \ldots, j_k)\})$ for the counts in each of the two tables of size $c_1 \times \cdots \times c_k$ that cross-classify the n_i multivariate responses at dose i.

2.1 *Simultaneous Marginal Homogeneity*

The counts displayed in Table 1 refer to the $k = 25$ one-way marginal distributions $\{\pi_{ih}(j_h) = \Pr(Y_{ih} = j_h), j_h = 1, \ldots, 4\}$ at the two dose levels $i = 1, 2$. These one-way margins are usually the parameters of interest when it comes to establishing a dose effect. Hence, instead of testing the much narrower H_0, we consider the less restrictive null hypothesis of equality of the vectors of marginal multinomial parameters under the two dose levels. That is, for each

adverse event h and outcome category j_h, $\pi_{1h}(j_h) = \pi_{2h}(j_h)$, and we have the hypothesis

$$H_0^{\text{marg}} \;:\; \bigcap_{h=1}^{k} \{H_{0h}\} = \bigcap_{h=1}^{k} \left\{ Y_{1h} \stackrel{d}{=} Y_{2h} \right\}, \tag{1}$$

noting that $H_0 \Rightarrow H_0^{\text{marg}}$. We refer to (1) as *simultaneous marginal homogeneity* (SMH) of the two multivariate ordinal distributions. For multivariate binary data, tests for SMH were investigated by Klingenberg and Agresti (2006) and Agresti and Klingenberg (2005). There is an interesting relationship between H_0 and the SMH hypothesis:

Theorem: *Under the prior assumption* $\boldsymbol{Y}_2 \stackrel{st}{\geq} \boldsymbol{Y}_1$,

$$H_0 : \boldsymbol{Y}_1 \stackrel{d}{=} \boldsymbol{Y}_2 \;\; \Leftrightarrow \;\; H_0^{\text{marg}} : \bigcap_{h=1}^{k} \left\{ Y_{1h} \stackrel{d}{=} Y_{2h} \right\}$$

$$H_1 : \boldsymbol{Y}_2 \stackrel{st}{\gneq} \boldsymbol{Y}_1 \;\; \Leftrightarrow \;\; H_1^{\text{marg}} : \bigcup_{h=1}^{k} \left\{ Y_{2h} \stackrel{st}{\geq} Y_{1h} \right\}$$

It follows that under the prior assumption, testing SMH against the alternative that there is marginal inhomogeneity $\Pr(Y_{2h} \geq j_h) \geq \Pr(Y_{1h} \geq j_h)$ (with strict inequality for at least one h and j_h) is equivalent to testing equality of the two multivariate distributions against the one-sided alternative that the treatment 2 distribution is stochastically larger and not equal to the treatment 1 distribution. The Theorem (proof in the Web Appendix) is also important for the validity of a permutation test of SMH and the multiplicity adjustments introduced later. The effect of the prior assumption is to restrict the total parameter space $\Omega = \left\{ \pi_i(j_1, \ldots, j_k), i = 1, 2 : \pi_i(j_1, \ldots, j_k) \geq 0, \sum_{(j_1, \ldots, j_k)} \pi_i(j_1, \ldots, j_k) = 1 \right\}$ spanned by the two unconstrained multinomials to the subset $\Omega_r = \{\pi_i(j_1, \ldots, j_k), i = 1, 2 : \boldsymbol{Y}_2 \stackrel{st}{\geq} \boldsymbol{Y}_1\}$. Under Ω_r, rejection of H_0 (or, equivalently, H_0^{marg}) directly leads to H_1 (or, equivalently, H_1^{marg}, i.e., marginal inhomogeneity). This restriction is similar to methods in constrained inference in the univariate case, where one tries to construct more efficient tests under some order restrictions on the parameter space (such as stochastic order or a monotone dose-response) that seem plausible for the given context (for a recent overview, see Silvapulle and Sen, 2005). A motivation for restricting the parameter space in toxicity studies is based

on the following observation: It is not unrealistic to expect that an increase in exposure to the toxin results in a shift towards higher outcome categories for some endpoints, while others are unaffected. On the other hand, a shift towards lower categories is unrealistic, as all endpoints describe adverse effects. Hence, we can assume *a priori* that $\boldsymbol{Y}_2 \overset{st}{\geq} \boldsymbol{Y}_1$, which implies $Y_{2h} \overset{st}{\geq} Y_{1h}$ for all h. In Section 5 we present an alternative approach of testing SMH when this prior assumption is not plausible, but show that a permutation approach can still be appropriate.

2.2 *Simple tests for SMH*

SMH corresponds to the hypothesis $H_0 : \beta_h = 0, h = 1, \ldots, k$ in the marginal cumulative logit model (the choice of the cumulative link is secondary)

$$\log[\Pr(Y_{ih} \leq j_h)/\Pr(Y_{ih} > j_h)] = \alpha_{hj_h} + \beta_h I(i = 2), h = 1, \ldots, k,\ j_h = 1, \ldots, c_h - 1,$$

that assumes a common shift β_h (proportional odds) across all c_h categories of adverse event h in the second dose group. However, estimating the large number of parameters via ML is impossible for sparse and/or high dimensional data because of the many empty cells. Also, the likelihood refers to the multinomial probabilities $\pi_i(j_1, \ldots, j_k)$ while the model above refers to logits of their cumulative one-way margins, which makes ML-fitting very complicated. Hence, LR-based inference might be impossible to conduct in such situations, even for moderate values of k and c_h. To reduce the number of parameters, Han et al. (2004) condition on the marginal totals to eliminate the α_{hj_h}'s in the binary context. An additional complication with ordinal responses is that the proportional odds assumption is unrealistic to hold for all k endpoints.

The simplest approach to testing SMH with ordinal data is to form a test statistic based solely on the differences in marginal sample proportions and their variances and possibly covariances. This has the advantage of being amenable to computer-intensive resampling procedures, as asymptotic results will not be valid in small sample or sparse/imbalanced sit-

uations. Let $n_{ih}(j)$ denote the marginal count of subjects with outcome j of endpoint h under dose i, as displayed in Table 1, with corresponding sample proportion $\hat{\pi}_{ih}(j) = n_{ih}(j)/n_i$. Many different statistics can be formed from the vector of marginal sample proportions $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{i1}(1), \hat{\pi}_{i1}(2), \ldots, \hat{\pi}_{i1}(c_1), \hat{\pi}_{i2}(1), \ldots, \hat{\pi}_{ik}(c_k))^t$, the most basic one being the difference in marginal sample proportions, $\boldsymbol{d} = \hat{\boldsymbol{\pi}}_2 - \hat{\boldsymbol{\pi}}_1$. With standard results from the underlying multivariate distributions, $\mathrm{E}[\boldsymbol{d}] = \boldsymbol{\pi}_2 - \boldsymbol{\pi}_1$ and $\mathrm{Cov}(\boldsymbol{d})$ is constructed from the variances and covariances within and between endpoints given by

$$\begin{aligned} \mathrm{Var}[d_h(j)] &= \mathrm{Var}[\hat{\pi}_{2h}(j) - \hat{\pi}_{1h}(j)] = \sum_{i=1}^{2} \pi_{ih}(j)[1 - \pi_{ih}(j)]/n_i \\ \mathrm{Cov}[d_h(j), d_{h'}(j')] &= \sum_{i=1}^{2} \left[\delta_{hh'}\pi_{ihh'}(j, j') - \pi_{ih}(j)\pi_{ih'}(j')\right]/n_i, \end{aligned} \tag{2}$$

where $\delta_{hh'} = 0$ if $h = h'$, $\delta_{hh'} = 1$ if $h \neq h'$ and $\pi_{ihh'}(j, j') = \Pr(Y_{ih} = j, Y_{ih'} = j')$ is the two-way marginal distribution for endpoints h and h'. However, this construction doesn't take advantage of the ordinal nature and the data are usually too sparse or imbalanced to give a positive definite estimate of $\mathrm{Cov}(\boldsymbol{d})$. Alternatively, let $A = \mathrm{diag}(\boldsymbol{\nu}_h^t, h = 1, \ldots, k)$ be a block-diagonal matrix with scores $\boldsymbol{\nu}_h^t = (\nu_h(1), \nu_h(2), \ldots, \nu_h(c_h))$ as blocks, where $\nu_h(.)$ is some monotone increasing scoring function for the c_h categories of endpoint h. Then, $\boldsymbol{s} = A\hat{\boldsymbol{\pi}}_2 - A\hat{\boldsymbol{\pi}}_1 = A\boldsymbol{d}$ compares mean scores among the two treatments, with covariance matrix $\Sigma = A\mathrm{Cov}(\boldsymbol{d})A^t$. Again, due to sparseness and/or imbalance, Σ and in particular its off-diagonal elements involving the two-way marginals may be impossible to estimate without further simplifying assumptions that lead to a positive definite matrix. One can assume homogeneity across all possible pairs of endpoints for the two-way margins, or, as in Han et al. (2004), equal correlation among endpoints within a domain and no higher order associations, paired with independence across domains. If neither of these assumptions are plausible, one can always form a test statistic ignoring the covariances between outcomes for different adverse effects, assuming working independence among endpoints.

In any case, it is more efficient to estimate elements of Σ assuming SMH, where one can

pool data from the two dose groups to obtain score-type statistics such as $W_0 = \mathbf{1}^t \widehat{\Sigma}_0^{-1/2} \boldsymbol{s}/k$, the average of weighted (by the elements of $\widehat{\Sigma}_0^{-1/2}$) mean score differences. Here, $\widehat{\Sigma}_0 = A\widehat{\mathrm{Cov}}_0(\boldsymbol{d})A^t$ is the estimated covariance matrix of $\boldsymbol{s}$ under the SMH hypothesis, where for each $\pi_{ih}(j)$ appearing in $\mathrm{Cov}(\boldsymbol{d})$ the pooled estimator $\hat{\pi}_{+h}(j) = [n_{1h}(j) + n_{2h}(j)]/(n_1 + n_2)$ is plugged in. Even after pooling the data, estimating off-diagonal elements of Σ can be problematic due to imbalance and sparseness, with potentially many combinations (j, j') sparsely or never observed for a given endpoint pair (h, h'). By assuming equal correlation in estimating these two-way margins, i.e., $\pi_{+hh'}(j, j')$ are identical for all $k(k-1)$ pairs (h, h'), constructing $\widehat{\Sigma}_0$ and hence W_0 may be possible for data not too sparse. Alternatively, as mentioned at the end of the last paragraph, one can consider $W_0' = \mathbf{1}^t \Delta_0^{-1/2} \boldsymbol{s}/k$ which, with $\Delta_0 = \mathrm{diag}(\widehat{\Sigma}_0)$, only incorporates as weights the variances and covariances among the c_h categories at a given endpoint, but ignores correlation among different endpoints. W_0' is then the average of standardized mean score differences formed at each endpoint. When some adverse effects are considered more important than others, a weighting scheme can be incorporated by simply replacing the column vector of $\mathbf{1}$ in the definition of W_0 or W_0' by a normalized vector of weights.

3. Permutation Testing of SMH

The significance of all statistics mentioned in the previous section should be evaluated via a resampling procedure such as permuting profiles among the two treatments. Note that by our Theorem, the null hypothesis of SMH together with the prior assumption $\boldsymbol{Y}_2 \overset{st}{\geq} \boldsymbol{Y}_1$ implies identical distributions across treatments, i.e., exchangeability. The permutation approach has the advantage that it automatically incorporates the correlation among endpoints by resampling entire profiles. Hence, even though statistics such as W_0 with the equal correlation assumption or W_0' with working independence simplify or ignore the correlation, their significance is evaluated taking it into account.

3.1 *Multiplicity adjusted P-values for endpoints*

Statistics such as W_0' only give a global evaluation of the dose effect but do not indicate which adverse effects (or domains) are responsible for the shift in the marginal distribution. Let $z_h, h = 1, \ldots, k$ denote the k components of $\Delta_0^{-1/2}\boldsymbol{s}$, the standardized mean score differences, which we use as endpoint-specific test statistics. Note that $z_h = [\boldsymbol{\nu}_h^t \widehat{\mathrm{Cov}}_0(\boldsymbol{d}_h)\boldsymbol{\nu}_h]^{-1/2} s_h$ and $W_0' = \frac{1}{k}\sum_{h=1}^{k} z_h$, where $\boldsymbol{d}_h$ is the vector of differences in marginal sample proportions between the two exposure levels for the c_h categories of endpoint h, $\widehat{\mathrm{Cov}}_0(\boldsymbol{d}_h) = (1/n_1 + 1/n_2)\left[\mathrm{diag}(\hat{\boldsymbol{\pi}}_{+h}) - \hat{\boldsymbol{\pi}}_{+h}\hat{\boldsymbol{\pi}}_{+h}^t\right]$ with $\hat{\boldsymbol{\pi}}_{+h} = (\hat{\pi}_{+h}(1), \ldots, \hat{\pi}_{+h}(c_h))^t$ the pooled sample proportions, and $s_h = \boldsymbol{\nu}_h^t \boldsymbol{d}_h$. We obtain multiplicity adjusted endpoint P-values following the closed testing principle of Marcus, Peritz and Gabriel (1976) that controls the familywise error rate (FWER) in the strong sense (i.e., the probability of rejecting one or more true null hypotheses H_{0h} under *any* constellation of true and false hypotheses is less than or equal to α). Let H_{0K}^{marg} denote the intersection hypothesis $\bigcap_{h\in K}\{H_{0h}\}$ for a subset $K \subseteq \{1, \ldots, k\}$. Under a full closed testing procedure there are $2^k - 1$ such intersection hypotheses to test, e.g., more than two million for the FOB data. Hence, we consider the shortcut provided by the step-down approach of Westfall and Young (1993), which is based on the maximum test statistic $\max_h z_h$ and only needs to consider k intersection hypothesis. By our Theorem, the permutation test for each of these intersection hypothesis is a valid (because exact) α-level test, so the FWER control is guaranteed by the closed testing principle. However, we will see in Section 4.1 that the support of the permutation distribution of $\max_h z_h$ is rather discrete (e.g., only 49 different realized values for the FOB data), making it automatically conservative. This effect is compounded if one considers a test statistic such as $\max_h s_h$, without standardizing by the estimated variance under SMH (e.g., a support of only 18 different values for the FOB data). Note that standardizing is imperative if endpoints are measured on different scales. Because of discreteness, we actually use the mid P-value

approach throughout in finding endpoint (and domain) adjusted P-values. The mid P-value (Lancaster, 1961; Hirji, 2006) is defined as 1/2 times the probability (estimated as the percentage out of all permutations) of observing a test statistic exactly equal to the one observed, plus the probability of observing any larger one. Although strictly controlling the FWER cannot be guaranteed with the mid P-value approach, simulations in Section 5 show that the distribution of the mid-P value is close to uniform under SMH, even for very sparse data.

3.2 *Domain Analysis*

The focus in the FOB analysis is not only on the individual adverse effects, but also which of the domains show increased toxicity, i.e., testing

$$H_0 : \bigcap_{\text{dom}} \{H_0^{\text{dom}}\} \qquad \text{vs.} \qquad H_1 : \bigcup_{\text{dom}} \{H_1^{\text{dom}}\},$$

where the individual domain hypotheses $H_0^{\text{dom}} : \bigcap_{h \in \text{dom}} \{H_{0h}\}$ and $H_1^{\text{dom}} : \bigcup_{h \in \text{dom}} \{H_{1h}\}$ are the intersection hypothesis of SMH and its complement over all endpoints within a domain (i.e., the intersection hypothesis H_{0K}^{marg} corresponding to the subset K of endpoints $\{1, \ldots, k\}$ that make up the domain). By grouping endpoints into domains (or body functions for drug safety data), the analysis at the domain level may increase efficiency for identifying the nature of toxicity effects. This is because some of the endpoints may measure similar effects and thus are redundant. Then, the use of step-down multiplicity corrections at the endpoint level may be too conservative and have low statistical power, leading to false negatives, an undesirable feature in toxicology.

By the closed testing principle, the adjusted P-value p_h^{adj} for endpoint h is the maximum over P-values formed by considering all possible intersection hypotheses that include endpoint h. One of these intersections must consist of all the other endpoints in h's domain, which leads to the following fact (proof in the Web Appendix): The adjusted P-value for a domain is always less than or equal to the minimum of the adjusted P-values of endpoints within the

domain, i.e, $p^{\text{adj}}_{\text{Dom}} \leq \min_{h\in\text{Dom}} p^{\text{adj}}_h$. This property results in more power to detect toxicity, albeit only at the domain level. The inequality shows that if a domain test is insignificant, no endpoint within the domain will achieve significance. Conversely, a single significant endpoint within a domain implies a significant domain effect. Finally, if the inequality is strict, domain significance can occur without any single endpoint being significant.

Let $\max_{h\in\text{Dom}_\text{m}} z_h$ be the test statistic for the m-th domain, $m = 1, \ldots, M$. Then, the step-down adjusted endpoint P-values from the previous section provide an upper bound for the domain tests. However, taking the maximum focuses only on the strongest toxicity effect, which may not be desirable when one wants to capture effects that accumulate over endpoints within a domain. Then, an appropriate test statistic is W_0' calculated over each domain m, i.e., $\frac{1}{|\text{Dom}_\text{m}|}\sum_{h\in\text{Dom}_\text{m}} z_h$, where $|.|$ denotes the number of endpoints in the domain. The decision on which type of domain test to use is not straightforward. As a guideline, if one expects endpoints within a domain to be correlated and with similar but possibly moderate effect, a test statistic such as W_0' is appropriate. On the other hand, multiplicity adjustments at the endpoint level are only possible via the distribution based on the maximum. Further, if one wants to ensure that non-significant endpoints are not unduely influencing domain results, taking $\max_{h\in\text{Dom}_\text{m}} z_h$ as domain test ensures this. Since grouping of endpoints into domains (or body functions) is often controversial or not well defined, this robustness feature may be a desirable property.

4. Analysis of the FOB data

We first present some simplistic results based on assigning equally spaced scores $\nu_h(j) = j$ to the categories and treating them as Gaussian. The standard way to deal with multiple endpoints in this context was presented by O'Brien (1984), who used as statistic the standardized sum of the individual t-statistics over all endpoints. Logan and Tamhane (2004) approximate

the distribution of this so called OLS-statistic by a t-distribution with $(n_1+n_2-2)(1+1/k^2)/2$ degrees of freedom. Table 1 shows O'Brien's OLS statistic and simple Bonferroni-Holm (Holm, 1979) adjusted p-values based on this approximation for testing an exposure effect in each of the 6 domains. Note that we can't compute the OLS statistic for the global test because there is insufficient information on the covariance matrix (only 16 observations for estimating the dependence among 25 endpoints) and hence we have to make the simplifying assumption of independence between domains. Then, the OLS test $T_{\text{glob}} = 4.2$ is significant (P= 0.002). Alternatively, the smallest adjusted domain P-value (0.034) can be used as the global test. The Table also shows raw and Bonferroni-Holm adjusted P-values for individual endpoints, based on forming, for each endpoint, a regular t statistic comparing the mean scores between the two exposures. The analysis reveals a significant shift in severity scores for endpoints in the Neuromuscular domain, containing the endpoint "Gait" that is the only one marginally significant at a 5% level.

With regard to a more appropriate permutation analysis treating the data as ordinal, there are $16!/(8!8!) = 12870$ distinct permutations of the 16 observed animal profiles into the two treatment groups, but many of them lead to identical values of a test statistic such as W_0 or W_0'. For the 4 endpoints "Salivation", "Tonic", "Palpebral" and "Piloerection" the exact same severity categories were observed for all 8 animals under both doses, and hence $s_h = 0$ with $\widehat{\text{Cov}}_0(\boldsymbol{d}_h) = \mathbf{0}$. We define $z_h \equiv 0$ for such cases. Alternatively, these 4 endpoints could have been excluded from the analysis as they hold no information about marginal inhomogeneity (results under both approaches are almost identical). We choose $W_0' = \mathbf{1}^t \Delta_0^{-1/2} \boldsymbol{s}/k$ as the test statistic, as with only 8 animals per dose group and $k = 25$ it is impossible to obtain a reliable (positive definite) estimate of the full covariance matrix Σ. Note that Δ_0 only needs to be computed once, since it is based on the pooled data and therefore invariant under permutations of profiles. The first panel in Figure 1 shows the

exact (i.e., using complete enumeration) permutation distribution of W_0'. The 95 percentile of this distribution equals 0.55, and hence the observed W_0' of 1.06 is significant (permutation P-value 0.0002; only 2 of the 12870 permutations yielded a larger W_0'), indicating a shift in marginal toxicity accumulated over the endpoints. For large sample sizes (which is not the case here) and $\pi_{ih}(j_h)$ bounded away from 0 and 1, the standardized mean score differences z_h follow an asymptotic $N(0,1)$ distribution. Under independence among endpoints, their average (which is W_0') is then asymptotically $N(0, 1/\sqrt{k})$. This distribution is superimposed in the Figure and one clearly sees that it's tails are too light, owing to small sample size and dependence among endpoints.

[Figure 1 about here.]

4.1 *Endpoint and Domain analysis*

After establishing a global exposure effect, naturally we are interested which endpoints (or domains) contribute to the significant difference. The raw (i.e., unadjusted) permutation mid P-value for endpoint h is simply the proportion of permutations that yield a z_h larger than the one observed, plus 1/2 times the proportion of permutations that yield a z_h equal to the one observed. Table 2 displays these using all possible permutations, while the second panel in Figure 1 shows the exact permutation distribution of $\max_h z_h$ (and its asymptotic distribution under independence), marking the observed z_h's as crosses on the x-axis. This is the starting point for finding the multiplicity adjusted P-values for all endpoints in a step-down procedure, where the permutation distribution of $\max_h z_h$ is used to find the adjusted P-value for the endpoint with the largest z_h statistic (e.g., "Gait"). Successive steps delete the endpoint corresponding to $\max_h z_h$ and find the permutation distribution of the maximum over the remaining endpoints, yielding the adjusted P-value for the next largest z_h, "Hindlimb" in our example. Table 2 lists all resulting step-down adjusted P-values for the $k = 25$ adverse effect in the FOB data. We observe two points: Firstly, even when

using $\max_h z_h$ as the global test statistic (instead of W_0' which incorporates *all* standardized differences) we still obtain a significant result indicating a shift in marginal toxicity. Secondly, we can identify the couple of endpoints that are largely responsible for this shift by inspecting the individual multiplicity adjusted P-values.

The Neuromuscular domain is one that includes significant adjusted P-values, hence it is guaranteed to show a significant (at the domain level) increase in toxicity at the $1.5g/kg$ exposure when compared to control. However, the domain adjusted P-values displayed in Table 2 are based on $\frac{1}{|\mathrm{Dom_m}|}\sum_{h\in\mathrm{Dom_m}} z_h$, using the full closed testing procedure for the multiplicity adjustments, which now only comprise testing $2^6 - 1 = 63$ intersection hypotheses, a more manageable number than the more than two million that would have been necessary at the endpoint level. We view this domain statistic more appropriate when the focus is on accumulated toxicity over many endpoints. It provides one more significant domain (Sensorimotor, P = 0.033) compared to the domain test with $\max_{h\in\mathrm{Dom_m}} z_h$, which yielded P = 0.079 for that domain. Apparently, the evidence against SMH provided jointly by the endpoints "Approach" and "Touch" is sufficient for an overall domain effect.

[Table 2 about here.]

4.2 *Score-free test statistics*

Typically used scoring systems $\boldsymbol{\nu}_h$ are based on equally spaced scores, used above, or midranks. However, if the chosen scores do not adequately reflect the numerical scale that underlies the ordered classification, the resulting test will not be sensitive. With $z_h(\boldsymbol{\nu}_h)$ as test statistic, it is straightforward to show that for any linear transformation of scores that preserves the monotonicity $z_h(a\mathbf{1} + b\boldsymbol{\nu}_h) = z_h(\boldsymbol{\nu}_h)$ with $a \in \mathbb{R}, b \in \mathbb{R}^+$. Hence, in the following we can focus on standardized scores $0 = \nu_h(1) \leq \ldots \leq \nu_h(c_h) = 1$ obtained by transforming the original scores via $a = -\nu_h(1)/\left[\nu_h(c_h) - \nu_h(1)\right]$ and $b = 1/\left[\nu_h(c_h) - \nu_h(1)\right]$ to the $[0, 1]$ interval. Consider the statistic $z_h^{\max} = \max_{\boldsymbol{\nu}_h}\{z_h(\boldsymbol{\nu}_h)\}$, where the data-driven scores $\boldsymbol{\nu}_h^{\max}$

that maximize $z_h(\boldsymbol{\nu}_h)$ are found by considering the following two cases (Kimeldorf, Sampson and Whitaker, 1992): If $\sum_{l=1}^{j_h} \hat{\pi}_{2h}(l) \geq \sum_{l=1}^{j_h} \hat{\pi}_{1h}(l)$ for all $j_h = 1, \ldots, c_h - 1$, the scores $\boldsymbol{\nu}_h^{\max}$ are one of the $(c_h - 1)$ monotone extreme points $0 = \nu_h^{\max}(1) = \ldots = \nu_h^{\max}(j_h) < \nu_h(j_h + 1) = \ldots = \nu_h(c_h) = 1$, $j_h = 1, \ldots, c_h - 1$. Otherwise, the scores $\boldsymbol{\nu}_h^{\max}$ are given by $v_h^{\max}(j) = (a_j - a_1)/(a_{c_h} - a_1)$, where $a_1, \ldots, a_{c_h}$ are determined via weighted isotonic regression from

$$\min_{a_1 \leq \ldots \leq a_{c_h}} \sum_{l=1}^{c_h} \left(\frac{n_{2h}(l)}{n_{+h}(l)} - a_l \right)^2 n_{+h}(l).$$

To compute $a_1, \ldots, a_{c_h}$, one uses the well known pool adjacent violators algorithm. If $n_{+h}(l) = 0$ for some category l, one simply collapses over this category, resulting in the corresponding shorter vector of scores. Maximum scores $\boldsymbol{\nu}_h^{\max}$ seem especially appropriate in the toxicity and safety context in the sense that they maximize the contrast (if we were to use normalized scores) of standardized mean score differences for the c_h categories of endpoint h. The second part of Table 2 shows individual and domain raw and adjusted P-values with scores $\boldsymbol{\nu}_h^{\max}$, using the same multiplicity adjusting procedures as outlined in Section 3. Results are comparable to equally spaced scores, although the order of significance of endpoints in the Neuromuscular domain is different and we gain one more significant endpoint ("Forelimb") at a 5% level.

All results of this Section (raw and adjusted endpoint and domain P-values in Table 2) can be reproduced using R-code displayed in the Web Appendix, with general R-functions available at `www.williams.edu/~bklingen`. Statistics that avoid assigning scores altogether, such as the likelihood ratio test computed under order restriction or the direct chi-squared test (Cohen, Madigan and Sackrowitz, 2003) resulted in slightly less significant results. We also tried using the minimum P-value (which may be more appropriate if endpoints are mixtures of categorical and continuous variables) instead of the maximum statistic for the

multiplicity adjustments at the endpoint level and alternative combinations of test statistics for the domain tests, all of which yielded similar conclusions.

5. A bootstrap approach to SMH

The validity of significance levels for testing the SMH hypothesis via the permutation approach rests on our theorem that established an equivalence between SMH and identical joint distributions (IJD, $\boldsymbol{Y}_1 \overset{d}{=} \boldsymbol{Y}_2$) under the prior condition $\boldsymbol{Y}_2 \overset{st}{\geq} \boldsymbol{Y}_1$, i.e, that joint toxicity profiles are stochastically ordered with more severe responses equally or even more likely under exposure. This assumption seems plausible in a toxicity study where all endpoints describe adverse effects, but in many clinical trials such as safety studies measuring side-effects of a drug, it can be the case that several marginal sample proportions are larger under the control (placebo) than under some treatment. If these differences are large enough to rule out sampling variability, this is in direct violation of the prior assumption on stochastic ordering. Then, it may not be appropriate to use the permutation approach to construct the null distribution (of any test statistic) under SMH, since exchangeability (i.e. IJD) is a stronger assumption.

We consider using the nonparametric bootstrap to test SMH when the prior assumption does not seem plausible. There are two options on how to resample to obtain a distribution reflecting SMH: We can draw bootstrap samples from centered scores within each group and form a test statistic (such as Welch's t statistic), or we can center an appropriate statistic (such as z_h) obtained from a bootstrap sample of the original responses within each group. Here, we use the latter approach, which is preferred by Pollard and van der Laan (2004) because it has asymptotic FWER control for the endpoint analysis. Also, the first approach will not be sensible when using data-driven scores. Let $\boldsymbol{Y}^*_{11}, \ldots, \boldsymbol{Y}^*_{1n_1}$ and $\boldsymbol{Y}^*_{21}, \ldots, \boldsymbol{Y}^*_{2n_2}$ denote bootstrap samples within each group. Under SMH, $\mathrm{E}_{\mathrm{SMH}}[\boldsymbol{d}_h] = \boldsymbol{0}$ and we expect the mean score difference s^*_h or its standardized version z^*_h computed from

the bootstrap sample to (asymptotically) equal 0. To reflect this, we center s_h^* or z_h^* by subtracting the observed s_h^{obs} or z_h^{obs}, respectively, the estimates of their means under the true data generating mechanism (which does not necessarily obey SMH). That is, for each resample we compute $t_h^{1*} = [\boldsymbol{\nu}_h^t \widehat{\text{Cov}}_0(\boldsymbol{d}_h^*)\boldsymbol{\nu}_h]^{-1/2}(s_h^* - s_h^{\text{obs}})$ or $t_h^{2*} = z_h^* - z_h^{\text{obs}}$ with $z_h^* = [\boldsymbol{\nu}_h^t \widehat{\text{Cov}}_0(\boldsymbol{d}_h^*)\boldsymbol{\nu}_h]^{-1/2} s_h^*$. Note that under SMH but not IJD, we can still use pooling to estimate the common marginal probabilities appearing in $\text{Cov}_0(\boldsymbol{d}_h^*)$. We mention both possibilities of first centering s_h^* and then standardizing it or centering z_h^* directly because they show rather different small sample behavior. This is due to the fact that with e.g. only 8 observations per group and many of them making the same response as in the FOB study, the bootstrap estimate of $\text{Var}(s_h)$ can be very small, which in our case resulted in many large values of t_h^{1*}. Consequently, raw and adjusted P-values for the FOB study are rather large when using t_h^{1*}, with no endpoint reaching significance even at a 20% significance level. By contrast, step-down adjusted P-values using the bootstrap distribution with t_h^{2*} for endpoints and domains (using $\frac{1}{|\text{Dom}_\text{m}|}\sum_{h\in\text{Dom}_\text{m}} t_h^{2*}$ for the m-th domain) are displayed in Table 2. Results are comparable to the permutation analysis, however, adjusted P-values are somewhat smaller for the significant endpoints and some of the domains. This is partly due to the fact that the bootstrap distribution of the maximum is far less discrete (699 distinct points in 10,000 resamples for the FOB data), but more importantly, that it does not control the FWER for small samples as simulations show below. Since the bootstrap test for endpoint h is only *asymptotically* level α, these P-values may be inappropriate for the FOB data with only 8 observations per group. Both centering methods give more congruent results (and again comparable to the permutation approach) when based on maximum scores $\boldsymbol{\nu}_h^{\max}$ and, as simulations show below, for large samples. The adjusted P-values using t_h^{2*} with maximum scores are also displayed in Table 2. Note that with data-driven scoring, it is necessary to recalculate s_h^{obs} or z_h^{obs} for each bootstrap iteration, as the scoring changes from one to the

next. R-functions for testing SMH with the two bootstrap approaches are also available from `www.williams.edu/~bklingen` and illustrated for the FOB data in the Web Appendix.

5.1 *A simulation study comparing FWER under permutation and bootstrap resampling*

To evaluate and compare the behavior of the proposed tests, we simulated 4000 datasets from a multivariate ordinal distribution with 9 endpoints (25 would be too computationally demanding), each with 4 categories. To simulate under the assumption of SMH without IJD we generated, for each group, a random vector of 4^9 multinomial probabilities $\{\pi(j_1, \ldots, j_9)\}$ and then used Iterative Proportional Fitting (Deming and Stephan, 1940) to modify the two vectors such that they agree in their 9 margins. These margins were set equal to the pooled proportions $\hat{\pi}_{+h}(j), j = 1, \ldots, 4$ for the $h = 1, \ldots, 9$ endpoints of the Neuromuscular and Sensimotor domains of our FOB study. To generate samples under IJD, we just used one of those vectors to simulate samples for both groups. Table 3 shows the actual type I error rate with regard to erroneously (at a nominal 5% level) declaring marginal inhomogeneity with both ways of centering under the bootstrap approach and under the permutation analysis of Section 3, for test statistics W_0' and $\max_h z_h$. The approximate nature of the bootstrap test for both W_0' and $\max_h z_h$ with either way of centering is obvious for small samples, where tests are either too conservative or too liberal. The permutation test has actual size closer to the nominal one, even under SMH without IJD (i.e., $\boldsymbol{Y}_2 \overset{st}{\geq} \boldsymbol{Y}_1$ does not hold). For larger sample sizes, all procedures have actual size close to the nominal one for both scenarios. Figure 2 shows QQ plots, comparing the distribution of mid P-values from the bootstrap and permutation approaches to the uniform distribution, when simulating under SMH without IJD and IJD. For sample sizes larger than 25, the QQ plots fall close to the diagonal, attesting to the near uniform distribution of mid P-values obtained from the bootstrap or permutation analysis. In Figure 2, we zoom in onto the most interesting part up to the 15th percentile. Then, one clearly sees the liberal behavior of the bootstrap test

(we only display the one based on centering z_h), which vanishes as sample size increases. The permutation test has excellent performance throughout, under SMH without IJD and under IJD.

[Table 3 about here.]

[Figure 2 about here.]

6. Discussion

The behavior of a permutation test of SMH when IJD does not hold was studied by Huang et al. (2006). They showed that under multivariate normality with balanced sample sizes, the permutation distribution of a vector of test statistics based on differences of sample means leads to valid tests and thus to a valid multiple testing procedure that controls the FWER. Hence, under asymptotic normality for the vector of mean score differences $\boldsymbol{s} = A\boldsymbol{d} = A(\hat{\boldsymbol{\pi}}_2 - \hat{\boldsymbol{\pi}}_1)$ or the standardized version $\boldsymbol{z} = \Delta_0^{-1/2}\boldsymbol{s}$, we obtain asymptotically valid tests with the permutation approach, even if the prior condition does not hold. However, even for finite but balanced samples, the exact distribution of $\boldsymbol{d}$ under SMH without IJD (i.e., the prior assumption does not hold) has the same mean and correlation matrix as the distribution of $\boldsymbol{d}$ under IJD (i.e., when the prior assumption holds). Trivially, $\mathrm{E}_{\mathrm{SMH}}[\boldsymbol{d}] = \mathrm{E}_{\mathrm{IJD}}[\boldsymbol{d}] = \boldsymbol{0}$, but with $n_1 = n_2 = n$ we also have from (2) that $\mathrm{Var}_{\mathrm{SMH}}[d_h(j)] = \mathrm{Var}_{\mathrm{IJD}}[d_h(j)] = \pi_{+h}(j)[1-\pi_{+h}(j)]/n$ and $\mathrm{Cov}_{\mathrm{SMH}}[d_h(j), d_{h'}(j')] = \mathrm{Cov}_{\mathrm{IJD}}[d_h(j), d_{h'}(j')] = 2\left[\delta_{hh'}\pi_{+hh'}(j,j') - \pi_{+h}(j)\pi_{+h'}(j')\right]/n$, where $\pi_{+hh'}(j,j') = [\pi_{1hh'}(j,j') + \pi_{2hh'}(j,j')]/2$.

Summing up, if the prior condition does not hold but $n_1 = n_2$, both the permutation and bootstrap procedures are asymptotically valid (which can also be seen from our simulations and Figure 2), whereas only the bootstrap approach is asymptotically valid when $n_1 \neq n_2$. However, the permutation approach still showed good performance for the unbalanced cases we considered in our simulations. For small and balanced samples, strictly speaking neither

procedure is valid (when $k \geq 3$), although the permutation distribution differs from the true distribution only in cumulants of order three and higher (Huang et al., 2006), which may have negligible effect as evidenced by our simulations. When the prior condition holds, the permutation approach is always exact, and no asymptotic tests are needed for the SMH hypothesis.

Finally, we conducted a limited power study where we evaluated the power of detecting marginal inhomogeneity. For this, we generated random vectors for each group that agreed in all but two of their marginal distributions, but differ in their higher order probabilities. These two margins were set equal to the observed marginal sample proportions of the endpoints "Gait" and "Approach" under the control and $1.5g/kg$ exposure from the FOB data. Power values for rejecting the SMH hypothesis (globally and at the individual level) are displayed in Table 3. The bootstrap and the permutation approach have similar power when the actual type I error is controlled by both procedures. Table 3 also displays the actual FWER of rejecting any true null hypothesis (for the scenario in Table 3, 7 null hypotheses are correct and 2, the ones corresponding to endpoints "Gait" and "Approach" are incorrect), using the endpoint specific adjusted P-values constructed via the step-down procedure. As the Table shows, the FWER is well controlled under all scenarios considered.

Supplementary Materials

All proofs and R code for the analysis of the FOB data can be found under the Paper Information link at the Biometrics website `http://www.biometrics.tibs.org`.

Acknowledgements

We would like to thank all reviewers for a stimulating discussion of our manuscript that improved its quality and Dr. Virginia Moser for providing the FOB data.

References

Agresti, A. and Klingenberg, B. (2005). Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Applied Statistics* **54**, 691–816.

Baird, S.J., Catalano, P.J., Ryan, L.M. and Evans, J.S. (1997). Evaluation of effect profiles: Functional Observational Battery outcomes. *Fundamental and Applied Toxicology* **40**, 37–51.

Cohen, A., Madigan, D. and Sackrowitz, H. (2003). Effective directed tests for models with ordered categorical data. *Australian & New Zealand Journal of Statistics* **45**, 285–300.

Cohen, A. and Sackrowitz, H. (2000). Testing whether treatment is "better" than control with ordered categorical data: definitions and complete class theorems. *Statistics & Decisions* **18**, 1–25.

Deming, W.E. and Stephan, E.F. (1940). On a least squares adjustment of a sampled frequency table when the expected margins are known. *Annals of Mathematical Statistics* **11**, 427–444.

Han, K.E., Catalano, P.J., Senchaudhuri, P. and Mehta, C. (2004). Exact analysis of dose response for multiple correlated binary outcomes. *Biometrics* **87**, 241–247.

Hirji, K.F. (2006). *Exact Analysis of Discrete Data.* London: Chapman & Hall/CRC.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

Huang, Y., Xu, H., Calian, V. and Hsu, J.C. (2006). To permute or not to permute. *Bioinformatics* **22**, 2244–2248.

Kimeldorf, G., Sampson, A.R. and Whitaker, L.R. (1992) Min and Max scoring for two-sample ordinal data. *Journal of the American Statistical Association* **60**, 216–224.

Klingenberg, B. and Agresti, A. (2006) Multivariate extensions of McNemar's test. *Biometrics* **62**, 921–928.

Lagakos, S.W., Wessen, B.J. and Zelen, M. (1986) An analysis of contaminated well water and health effects in Woburn, Massachusetts. *Journal of the American Statistical Association* **81**, 583–596.

Lancaster, H.O. (1961) The combination of probabilities: an application to orthonormal functions. *Australian Journal of Statistics* **3**, 20–33.

Logan, B.R., and Tamhane, A.C. (2004). On O'Brien's OLS and GLS tests for multiple endpoints. In: Benjamini, Y., Bretz, F., and Sarkar, S. (Eds), *Recent Developments in Multiple Comparison Procedures*, IMS Lecture Notes, Monograph Series **47**, 76–88.

Marcus, R., Peritz, E. and Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

Marshall, A. and Olkin, I., (1979). *Inequalities: Theory of Majorization and Its Applications.* New York: Academic Press.

Molenberghs, G. and Ryan, L.M. (1999). An exponential family model for clustered multivariate binary data. *Environmetrics* **10**, 279–300.

Moser, V.C. (1989). Screening approaches to a neurotoxicity: A functional observational battery. *Journal of the American College of Toxicology* **8**, 85–93.

Moser, V.C., Cheek, B.M. and MacPhail, R.C. (1995). A multidisciplinary approach to toxicological screening. III. Neurobehavioral toxicity. *Journal of Toxicology and Environmental Health* **45**, 173–210.

Müller, A. (2001). Stochastic ordering of multivariate normal distributions. *Annals of the Institute of Statistical Mathematics* **53**, 567–575.

O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087.

Pocock, S. J., Geller, N. L. and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487–498.

Pollard, K.S. and van der Laan, M.J. (2004). Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference* **125**, 85–100.

Silvapulle, J.S. and Sen, P.K. (2005). *Constrained Statistical Inference. Inequality, Order, and Shape Restictions.* Hoboken, New Jersey: John Wiley & Sons.

Troendle, J.F. (2005). Multiple comparisons between two groups on multiple Bernoulli outcomes while accounting for covariates. *Statistics in Medicine* **24**, 3581–3591.

Westfall, P.H. and Young, S.S. (1989). P value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association* **84**, 780–786.

Westfall, P.H. and Young, S.S. (1993). *Resampling-based multiple testing.* New York: John Wiley & Sons.

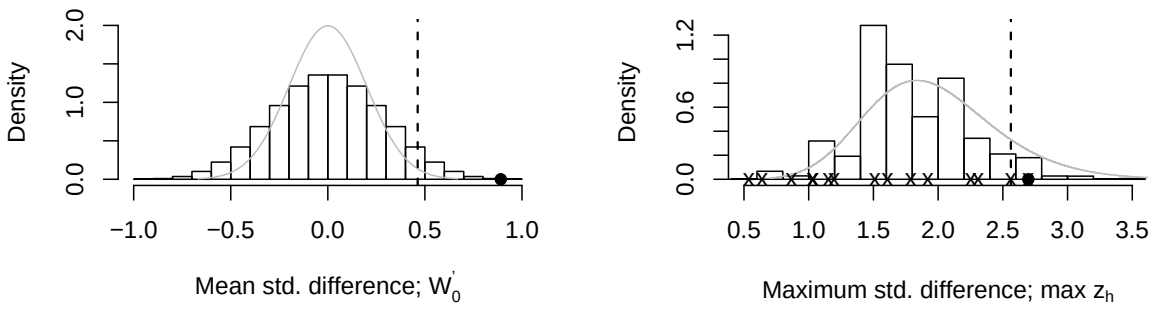

Figure 1. Permutation distributions of global test statistic W_0' (first panel) and $\max_h z_h$ (second panel) for comparing the $1.5g/kg$ exposure to control. Dashed lines indicate 95th percentiles and full circles observed values of test statistics. Crosses indicate observed values for individual endpoints. Grey curves show asymptotic distributions assuming independence among endpoints.

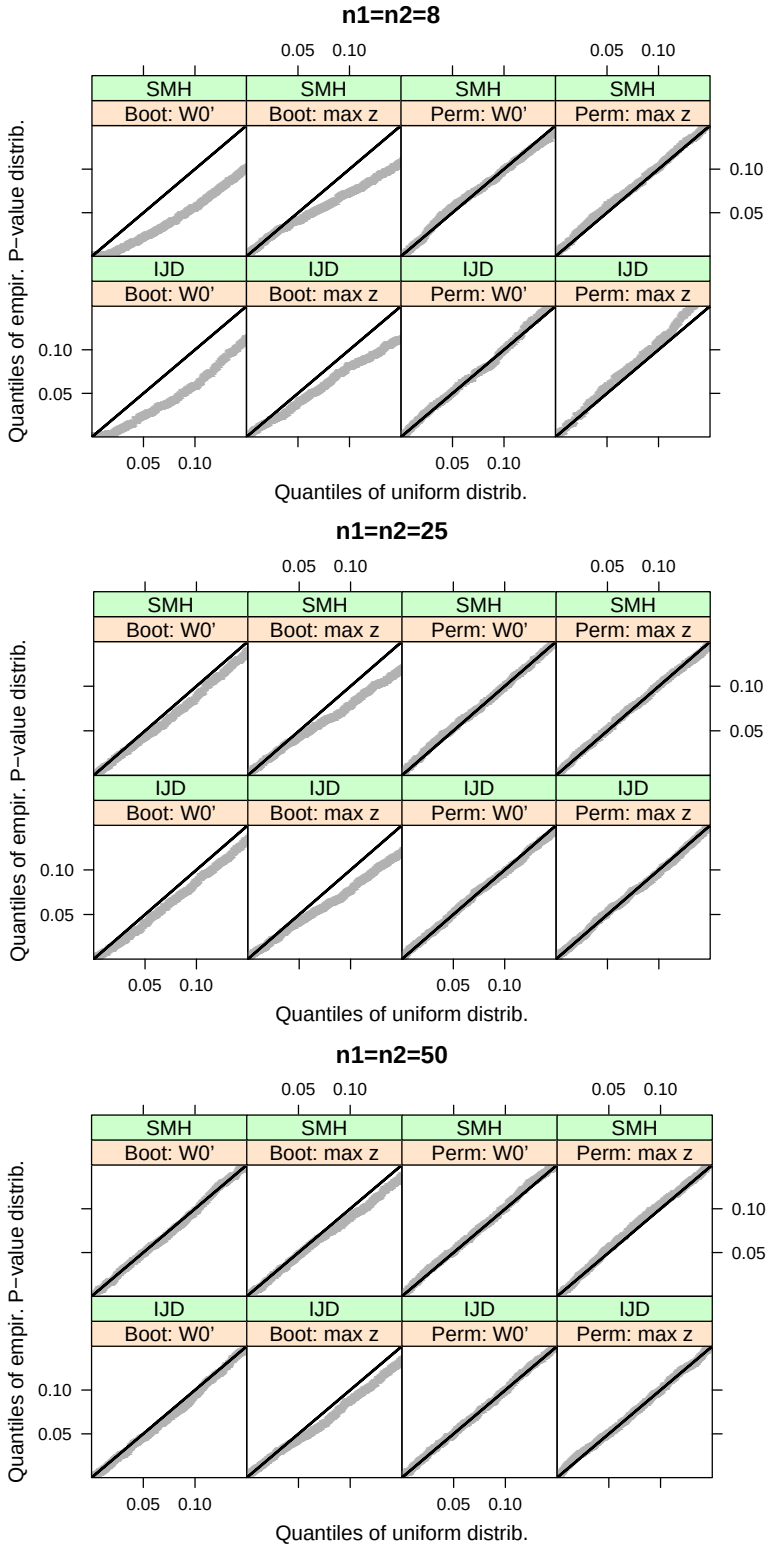

Figure 2. QQ-plots of empirical P-value (from 4000 simulations) against uniform distribution under the two scenarios SMH (w/o IJD) and IJD. P-values for the SMH hypothesis with both, the test based on W_0' and $\max_h z_h$ are derived from bootstrapping ("Boot: W0' " and "Boot: max z") or from permuting ("Perm: W0' " and "Perm: max z") the data. We only display results for the bootstrap test with centering z_h.

Table 1

FOB study: Marginal counts of severity-scores for adverse effects observed at two exposure levels. The last three columns give a simplistic analysis treating the scores as normal and list the corresponding t statistic with associated raw and Bonferroni-Holm adjusted P-values. The quoted domain test statistic T for each domain is based on O'Brien's (1984) OLS test applied to all endpoints within a domain, with p-values adjusted via Bonferroni-Holm.

Domain	Endpoint	Exposure									P-value	
		0 g/kg				1.5 g/kg				t	raw	adj.
		1	2	3	4	1	2	3	4			
Autonomic	Lacrimation	8	0	0	0	5	0	3	0	2.05	0.030	0.626
	Salivation	8	0	0	0	8	0	0	0	0.00	0.500	1.000
$T = 0.93$	Pupil	7	0	1	0	5	0	3	0	1.13	0.139	1.000
$p_{\text{adj}} = 0.404$	Defecation	7	0	0	1	7	1	0	0	-0.63	0.731	1.000
	Urination	4	3	1	0	6	1	1	0	-0.67	0.744	1.000
Sensorimotor	Approach	8	0	0	0	4	0	3	1	2.55	0.011	0.253
	Click	7	0	1	0	7	0	1	0	0.00	0.500	1.000
$T = 2.08$	Tail pinch	6	0	2	0	5	0	3	0	0.51	0.309	1.000
$p_{\text{adj}} = 0.185$	Touch	8	0	0	0	6	0	0	2	1.53	0.074	1.000
CNS excitability	Handling	6	2	0	0	4	4	0	0	1.00	0.167	1.000
	Clonic	4	4	0	0	5	3	0	0	-0.47	0.679	1.000
$T = 1.25$	Arousal	4	3	1	0	3	0	3	2	1.64	0.061	1.000
$p_{\text{adj}} = 0.404$	Removal	0	8	0	0	0	7	1	0	1.00	0.167	1.000
	Tonic	8	0	0	0	8	0	0	0	0.00	0.500	1.000
CNS activity	Posture	8	0	0	0	7	0	1	0	1.00	0.167	1.000
$T = 1.28$	Rearing	5	2	1	0	4	2	2	0	0.61	0.277	1.000
$p_{\text{adj}} = 0.404$	Palpebral	8	0	0	0	8	0	0	0	0.00	0.500	1.000
Neuromuscular	Gait	8	0	0	0	3	5	0	0	3.42	0.002	0.052
	Foot splay	6	1	1	0	6	1	1	0	0.00	0.500	1.000
$T = 3.36$	Forelimb	5	2	1	0	2	1	0	5	2.65	0.010	0.221
$p_{\text{adj}} = 0.034$	Hindlimb	5	3	0	0	0	6	1	1	3.12	0.004	0.090
	Righting	8	0	0	0	5	2	1	0	1.87	0.041	0.824
Physiological	Piloerection	8	0	0	0	8	0	0	0	0.00	0.500	1.000
$T = 1.38$	Weight	6	1	1	0	4	2	0	2	1.17	0.130	1.000
$p_{\text{adj}} = 0.404$	Temperature	6	1	1	0	4	3	0	1	0.83	0.210	1.000

Table 2

Raw and multiplicity adjusted P-values for endpoints and domains, comparing the $1.5g/kg$ exposure to control, with test statistic $z_h(\boldsymbol{\nu}_h)$ based on equally spaced and maximum scores.

Domain	scores $\boldsymbol{\nu}_h = (1,2,3,4)$				maximum scores $\boldsymbol{\nu}_h^{\max}$			
Endpoint	z_h	raw	adj.[a]	adj.[b]	z_h	raw	adj.[a]	adj.[b]
Autonomic[c]		**0.214**	**0.214**	**0.216**		**0.135**	**0.135**	**0.184**
Lacrimation	1.92	0.050	0.331	0.236	1.92	0.050	0.434	0.219
Salivation	0.00	0.500	0.959	0.966	0.00	0.500	0.960	0.930
Pupil	1.15	0.162	0.844	0.726	1.15	0.162	0.875	0.664
Defecation	-0.67	0.633	0.959	0.966	0.00	0.633	0.960	0.930
Urination	-0.71	0.738	0.959	0.966	0.00	0.630	0.960	0.930
Sensorimotor[c]		**0.033**	**0.033**	**0.002**		**0.035**	**0.035**	**0.095**
Approach	2.25	0.019	0.119	0.067	2.31	0.019	0.138	0.064
Click	0.00	0.500	0.959	0.966	0.00	0.500	0.960	0.930
Tail pinch	0.54	0.321	0.927	0.897	0.54	0.321	0.933	0.885
Touch	1.51	0.117	0.566	0.543	1.51	0.117	0.731	0.494
CNS excitability[c]		**0.130**	**0.130**	**0.076**		**0.135**	**0.135**	**0.190**
Handling	1.03	0.182	0.883	0.788	1.03	0.182	0.921	0.819
Clonic	-0.50	0.671	0.956	0.966	-0.50	0.671	0.960	0.930
Arousal	1.61	0.064	0.456	0.463	2.14	0.047	0.235	0.095
Removal	1.03	0.250	0.883	0.788	1.03	0.250	0.921	0.769
Tonic	0.00	0.500	0.959	0.966	0.00	0.500	0.960	0.930
CNS activity[c]		**0.107**	**0.107**	**0.037**		**0.113**	**0.113**	**0.348**
Posture	1.03	0.250	0.833	0.788	1.03	0.250	0.921	0.769
Rearing	0.64	0.280	0.927	0.897	0.67	0.294	0.933	0.836
Palpepral	0.00	0.500	0.959	0.966	0.00	0.500	0.960	0.930
Neuromuscular[c]		**0.001**	**0.001**	**0.000**		**0.001**	**0.002**	**0.015**
Gait	2.70	0.006	0.025	0.013	2.70	0.006	0.041	0.013
Foot splay	0.00	0.500	0.959	0.966	0.00	0.597	0.960	0.930
Forelimb	2.31	0.012	0.096	0.061	2.70	0.009	0.041	0.013
Hindlimb	2.56	0.003	0.044	0.036	2.83	0.003	0.028	0.010
Righting	1.79	0.050	0.352	0.280	1.92	0.050	0.434	0.219
Physiological[c]		**0.089**	**0.089**	**0.045**		**0.119**	**0.119**	**0.191**
Piloerection	0.00	0.500	0.959	0.966	0.00	0.500	0.960	0.930
Weight	1.20	0.133	0.834	0.705	1.55	0.152	0.543	0.383
Temperature	0.87	0.224	0.883	0.843	1.26	0.285	0.875	0.658

[a] Multiplicity adjustments based on the step-down procedure with the maximum test statistic under permutation resampling

[b] Multiplicity adjustments based on the step-down procedure with the maximum test statistic under bootstrap resampling

[c] Domain P-values are based on $\sum_{h \in \text{Dom}} z_h / |\text{Dom}|$, with multiplicity adjustments via full closed testing

Table 3

Part A: *Actual Type I error rate (in %) with bootstrap (we give results for both types of centering, separated by a "/") and permutation resampling under SMH (w/o IJD) and IJD with 9 endpoints. The column labeled "W_0'" refers to the proportion out of 4000 generated data sets for which the global test of SMH using W_0' yielded a P-value less than 5%, while the column labeled "$\max z_h$" refers to the proportion out of the same 4000 generated data sets for which at least one of the 9 individual step-down adjusted P-values was less than 5%. Simulation margin of error: 0.7%.* ***Part B:*** *Power and FWER (in %) for establishing marginal inhomogeneity when 2 (Gait and Approach) out of the 9 endpoints show marginal inhomogeneity, using the global test statistic W_0' or the step-down adjusted P-values based on $\max z_h$ for these two endpoints. The column labeled "FWER" shows the proportion out of 4000 generated data sets for which one or more of the 7 true hypotheses were wrongly rejected (i.e., their step-down adjusted P-value < 0.05).*

Part A: Type I error rate

	SMH				IJD			
	Bootstrap		Permutation		Bootstrap		Permutation	
(n_1, n_2)	W_0'	$\max z_h$	W_0'	$\max z_h$	W_0'	$\max z_h$	W_0'	$\max z_h$
(8,8)	7.9/9.1	3.2/6.5	4.4	4.4	8.3/8.7	3.7/6.4	5.0	4.3
(25,25)	5.9/5.9	4.1/6.0	4.7	4.9	6.0/6.1	4.5/6.1	4.9	4.8
(50,50)	5.2/5.1	4.9/5.3	4.9	4.5	5.5/5.6	5.3/6.0	5.0	5.0
(100,100)	5.1/5.2	4.9/5.2	4.9	4.7	5.0/5.0	5.1/5.2	4.8	4.8
(8,25)	7.0/7.2	4.0/5.3	4.9	5.4	8.1/8.3	2.9/4.1	5.4	4.5
(25,50)	5.8/6.0	4.6/5.6	4.8	5.1	6.6/6.5	4.0/5.0	5.1	4.8
(25,100)	6.6/6.4	4.3/5.0	4.9	5.5	6.4/6.1	3.9/4.5	4.7	5.0
(100,25)	4.7/4.6	5.0/5.7	4.2	5.2	5.7/5.3	5.0/5.5	5.3	5.3

Part B: Power and FWER

	Bootstrap				Permutation			
(n_1, n_2)	W_0'	Gait	Appr.	FWER	W_0'	Gait	Appr.	FWER
(8,8)	50/52	48/61	31/42	3.9/6.2	42	54	36	4.0
(25,25)	79/80	89/89	94/96	4.4/5.6	79	91	95	4.4
(50,50)	96/96	99/99	100/100	4.9/5.9	96	99	100	4.9
(100,100)	100/100	100/100	100/100	5.4/5.6	100	100	100	5.2
(8,25)	56/56	64/75	42/55	2.6/3.3	48	79	67	4.6
(25,50)	87/88	95/95	99/99	4.6/5.4	87	96	99	4.9
(25,100)	92/92	99/98	100/100	4.1/4.8	91	99	100	5.1
(100,25)	98/98	99/98	100/100	5.2/5.3	98	99	100	5.4