Order-Restricted Inference for Multivariate Binary Data With Application to Toxicology

Ori Davidov and Shyamal Peddada

In many applications, researchers collect multivariate binary response data under two or more naturally ordered experimental conditions. In such situations, one is often interested in using all binary outcomes simultaneously to detect an ordering among the experimental conditions. To make such comparisons, we develop a general methodology for testing for the multivariate stochastic order between \( K \geq 2 \) multivariate binary distributions. Our proposed test uses order-restricted estimators, which, according to our simulation study, are more efficient than the unrestricted estimators in terms of their mean squared error. We compared the power of the proposed test with that of several alternative tests, including procedures that combine individual univariate tests and a Bonferroni-based test. We also compared the proposed test with an unrestricted Hotelling \( T^2 \)-type test. Our simulations suggest that the proposed method competes well with these alternatives. The gain in power is often substantial. The proposed methodology is illustrated by applying it to a two-year rodent cancer bioassay data obtained from the U.S. National Toxicology Program. Supplemental materials are available online.

KEY WORDS: Dose–response study; Multivariate binary data; Order-restricted statistical inference; Stochastic order relation.

1. INTRODUCTION

In this article we develop a methodology for analyzing multivariate binary (MVB) response data when the explanatory variable is an ordered categorical variable, such as the dose of a drug or toxin, the level of a treatment, or the stage of a cancer.

The random vector \( \mathbf{X} = (X_1, \ldots, X_m) \) is said to follow a MVB distribution if its components \( X_i, i = 1, \ldots, m \), are binary random variables (RVs). Various probability models for MVB data have been proposed (e.g., Joe 1997). MVB response data arise in diverse applications, including epidemiology (Ananth and Kantor 2004), dentistry (Burnside, Pine, and Williamson 2007), genetics and animal breeding (Hoschele et al. 1986), psychology and economics (Maydeu-Olivares, Hernandez, and McDonald 2006), clinical trials (Pocock, Geller, and Tsiatis 1987), and quality of life studies (Ribaudo and Thompson 2002). Our investigation is motivated by applications in toxicology.

Several statistical methods are available for analyzing unordered MVB response data, depending on the scientific question of interest. One popular methodology is to model the MVB outcomes using a latent multivariate normal random vector and a set of fixed thresholds that induce binary RVs (Hoschele et al. 1986; Maydeu-Olivares, Hernandez, and McDonald 2006). This approach is appealing because it is easy to interpret. Moreover, the data can be analyzed using well-studied hierarchical models. However, a problem with such a methodology is the arbitrary choice of various thresholds and the suitability of the latent variable model in the given application. Another family of methods is based on the generalized estimating equation (GEE) methodology of Liang and Zeger (1986), which fits marginal models to the data. One concern with this approach is the potential loss of power (e.g., Fitzmaurice, Laird, and Zahner 1996). Several modifications to the GEE approach have been proposed in the literature by considering multivariate logistic regression where, in addition to the marginal probabilities, marginal odds ratios are modeled (e.g., Fitzmaurice and Laird 1993; Glonke and McCullach 1995). Another reason for not considering GEE in this article is that in our application, dealing with animal studies, we cannot be certain of the cause of death. The use of logistic regression models for modeling tumor incidence in terms of the lifetime of an animal implicitly assumes that the tumor is nonlethal, which might not be appropriate. Finally, resampling methods provide another strategy for testing for equality among \( K \geq 2 \) MVBs (Westfall and Young 1989; Troendle 2005). One nice feature of resampling methods is that they naturally account for dependence without explicitly modeling it. Many of these methods and others are described in greater detail by Song (2007).

Industries and various agencies, including the Environmental Protection Agency, the Organization for Economic Cooperation and Development, the National Toxicology Program (NTP), routinely conduct dose–response studies in rodents to investigate the toxicity and/or carcinogenicity of various chemicals (Keenan et al. 2009). Toxicologists have long believed that some tumors may be associated or coexist. Therefore, while making decisions regarding the effect of a chemical on a particular tissue, the NTP also may consider information regarding an associated tumor. For example, tumors of the pituitary gland and the mammary gland are believed to be associated through the prolactin pathway (e.g., McComb et al. 1984; National Toxicology Program 2005). It follows that a dose-related trend in both tumors may provide stronger evidence of carcinogenic activity. Often the NTP draws conclusions informally by examining the two tumors individually, because formal methods for...
combining information from different tumors sites do not exist. Similar issues arise in clinical trials and drug safety studies. For example, in a clinical trial one treatment may be preferred over another if it reduces the probability of observing a collection of prespecified adverse events.

Researchers are often interested in demonstrating that the joint probability of events is increasing with dose. The foregoing methods may not be directly applicable for making such comparisons. A simple approach to the problem, which is often used, is to conduct univariate analysis on each binary variable separately and then perform a Bonferroni correction on the resulting p-values. Because Bonferroni corrections do not require any information on the underlying dependence structure, the resulting methodology does not require any complicated modeling of the association among the response variables. A potential drawback of such a methodology is that it may lose power as the number of binary variables increases. Rather than testing for the equality of each marginal binomial probability separately, Agresti and Klingenberg (2005) and Klingenberg and Agresti (2006) introduced binary analogs to the classical Hotelling $T^2$ statistic. Although the methodology developed in those articles is ideal for comparing two or more experimental conditions in terms of all marginal probabilities simultaneously, they are not designed to test for an ordering among the experimental conditions. To the best of our knowledge, the only methodology in the literature that addresses the problem considered in this article is based on combining p-values from marginal tests in a permutation framework (e.g., Pesarin 2001; Finos, Salmaso, and Solarli 2007). It is important to emphasize that although the permutation framework honors the dependence among the binary RVs, the underlying test statistics ignores it. Moreover, the methodology to carry out such tests when the data are high-dimensional. The new methodology does not require any model of the dependence structure. Unlike the existing methodologies, the proposed formulation allows us not only to draw inferences on the joint probabilities, but also, and more importantly, to test for an ordering among the experimental conditions. We provide a precise formulation of the problem in Section 2. In Section 3 we develop estimation and testing procedures that explicitly account for multivariate stochastic order constraints and show through simulations that the proposed methodology improves the precision of the estimators and increases power for testing hypotheses. In Section 4 we modify the proposed methodology so that it can be applied to rodent cancer bioassay data, where tumor incidence data are frequently incomplete. We also provide results of a small simulation study that demonstrate that the proposed test maintains the type I error rates. In Section 5 we apply the methodology to a dataset obtained from the NTP, where female rats were exposed to four different doses of malachite green chloride, an antifungal agent used in the fish industry. We formally demonstrate that the joint incidence of mammary gland and pituitary gland tumors increases in a dose-related manner. The NTP noted this outcome informally. We provide a brief discussion and concluding remarks in Section 6, and relegate proofs to an Appendix.

2. FORMULATION

Let $X_i$ and $X_j$ be $m$-dimensional RVs. The RV $X_i$ is said to be smaller than $X_j$ in multivariate stochastic order, denoted by $X_i \preceq_{st} X_j$, if $P(X_i \leq X_j) = 1$ for all upper sets $U \in \mathbb{R}^m$. A set $U \in \mathbb{R}^m$ is called an upper set if $u \in U$ implies that $v \in U$ whenever $u_i \leq v_i$, $i = 1, \ldots, m$. It follows that if $X_i$ and $X_j$ are MVBS, then $X_i \preceq_{st} X_j$ if and only if for all upper sets $U \in \mathbb{R}^m$

$$\sum_{t \in U} p_i(t) \leq \sum_{t \in U} p_j(t),$$

(2.1)

where $p_i(t) = P(X_{i1} = t_1, \ldots, X_{im} = t_m)$ for $t_i \in \{0, 1\}$, $i = 1, \ldots, m$, and $p_j(t)$ is defined similarly. Furthermore, if there exists $U \in \mathbb{R}^m$ for which the inequality in (2.1) is strict, then $X_i$ is said to be strictly smaller than $X_j$ in the multivariate stochastic order. This relationship is denoted by $X_i \prec_{st} X_j$. Thus the event $\{X_i \preceq_{st} X_j\}$ is equivalent to $\{X_i =_{st} X_j\} \cup \{X_i \prec_{st} X_j\}$. This notion of multivariate stochastic order generalizes the univariate one and simply indicates that one random vector is more likely than another to take on larger values. In toxicology, this statement implies that there is equal or greater chance of observing tumors in the high-dose group (group $j$) than in the low-dose group (group $i$).

Equation (2.1) implies that two MVBS are ordered provided that a set of linear inequalities on their probabilities hold. It is easily verified that an upper set $U$ generates a nontrivial inequality if the probabilities $\sum_{t \in U} p_i(t)$ and $\sum_{t \in U} p_j(t)$ are in $(0, 1)$. For example, if $m = 1$, then $X_i$ and $X_j$ are binary RVs, and $X_i \preceq_{st} X_j$ if and only if $p_i(1) \leq p_j(1)$. If $m = 2$ (i.e., $X_i$ and $X_j$ are bivariate binary RVs), then it can be verified that $X_i \preceq_{st} X_j$ if and only if the following inequalities hold:

$$p_i(1, 1) \leq p_j(1, 1),$$

$$p_i(1, 0) + p_i(1, 1) \leq p_j(1, 0) + p_j(1, 1),$$

$$p_i(0, 1) + p_i(1, 1) \leq p_j(0, 1) + p_j(1, 1),$$

$$p_i(0, 0) + p_i(1, 0) + p_i(1, 1) \leq p_j(0, 0) + p_j(1, 0) + p_j(1, 1).$$

(2.2)

In principle, one could enumerate all upper sets for any given $m$. Unfortunately, and as explained in Remark 1, the number of upper sets grows at a superexponential rate.

Remark 1. Let $N_m$ be the number of upper sets for which the inequality (2.1) is nontrivial. Let $S_m = \{(t_1, \ldots, t_m) : t_i \in \{0, 1\}\}$ denote the support of an $m$-dimensional MVB and let $\mathcal{U}(S_m)$ be the family of upper sets on $S_m$. Note that the sets $\emptyset$ and $S_m$ are the only members of $\mathcal{U}(S_m)$ that guarantee a trivial inequality in (2.1) (i.e., $0 \leq 0$ and $1 \leq 1$). It is well known that there is a one-to-one correspondence between the elements of $\mathcal{U}(S_m)$ and the antichains in $S_m$. The antichains are those subsets of $S_m$ in which no two elements are ordered (e.g., Davey and Priestley 2002). Furthermore, the number of antichains is given by the Dedekind numbers, $D_m$, $m \geq 0$. It follows that $N_m = D_{m+1} - 2$, where $m \geq 1$. The first six elements of this sequence are $N_1 = 1,$
We have $N_2 = 4$, $N_3 = 18$, $N_4 = 166$, $N_5 = 7579$, and $N_6 = 7,828,352$. The values of $N_7$ and $N_8$ are also known, but $N_9$ is not (lower bounds and approximations are available for all $m$). For more details, see the On-Line Encyclopedia of Integer Sequences at http://www.research.att.com/~njas/sequences/A000372.

In many applications, $m$ is large. For instance, in neurotoxicology a researcher may be interested in studying the effect of a toxin on the nervous system of an animal by measuring as many as 20–30 binary response variables, known as functional observational battery; examples include response to touch, tail pinch, and so forth (e.g., Moser 2000; Han et al. 2004). In such cases the enumeration of all upper sets is practically impossible. Thus, to proceed we must first reduce the dimension of the problem. This can be done with the help of Theorem 1.

**Theorem 1.** Suppose that $X_i \preceq_{st} X_j$ then $X_i \prec_{st} X_j$ if and only if $X_{il} \prec_{st} X_{jl}$ for some $1 \leq l \leq m$.

Theorem 1 shows that if $X_i$ is equal to or smaller stochastically than $X_j$, then a strict ordering between $X_i$ and $X_j$ holds provided that at least one of the marginal distributions is strictly ordered. Note that in many dose–response studies, such as those conducted by toxicologists at the NTP, the investigators believe a priori that the chemical under study is harmful and that its effect worsens as the dose increases. Without such prior belief agencies such as the NTP would not spend millions of dollars to conduct 2-year bioassays. Thus the assumption that $X_i \preceq_{st} X_j$ whenever $i \leq j$ is reasonable in our application.

Although the focus of this article is binary outcomes, Theorem 1 is valid for discrete and continuous variables and combinations thereof. Consequently, it can be used to develop methods for a much broader collection of applications than those described here. For example, in early-phase clinical trials one expects the number and severity of adverse effects to have a monotone trend in dose. In such studies researchers collect continuous outcomes (e.g., blood chemistry), ordinal outcomes (e.g., quality of life measures), and binary outcomes (e.g., existence of specified adverse effects).

**Remark 2.** The relationship between the ordering of multivariate distributions and their marginals was noted by Finos, Salmaso, and Solari (2007) and Finos et al. (2008), who quoted Baccelli and Makowski (1989). Giancristofaro and Bonnini (2008) showed that if the marginal distributions are supported on $\{0, \ldots, s\}$ then they can be ordered by ordering their first $s – 1$ moments. Klingenberg et al. (2009) formulated a theorem similar to Theorem 1 for the case of multivariate discrete distributions with finite support. Their proof establishes the orthant orders. More specifically, they showed that, under the conditions of Theorem 1, $P(X_i \leq x) \geq P(X_j \leq x)$ and $P(X_i > x) \leq P(X_j > x)$ for all $x$, with a strict inequality for some $x$. The first inequality is equivalent to the lower orthant order, denoted by $X_i \preceq_{lo} X_j$, and the second is equivalent to the upper orthant order, denoted by $X_i \preceq_{uo} X_j$. It can be shown (see Supplemental Materials) that if $X_i$ and $X_j$ are MVBS, then $X_i \preceq_{lo} X_j$ and $X_i \preceq_{uo} X_j$ are jointly equivalent to $X_i \preceq_{st} X_j$ if and only if $m = 2$. However, Theorem 1 establishes a strict multivariate stochastic order that is stronger than the orthant orders; moreover, it does so for arbitrary RVs. Further comparison of the multivariate stochastic order and the orthant orders has been provided by Shaked and Shanthikumar (2007).

**Remark 3.** Note that in general, the univariate stochastic ordering of marginal distributions does not imply the multivariate stochastic ordering of the joint distribution. Consider, for example, two bivariate binary distributions. The Bahadur (1961) representation of their probability function is

$$p_4(t_1, t_2) = \frac{2}{\sqrt{p_{11}(1 – p_{11})} \sqrt{p_{12}(1 – p_{12})}} \left(1 + \rho_1 \frac{t_1 – p_{11}}{\sqrt{p_{11}(1 – p_{11})}} \frac{t_2 – p_{12}}{\sqrt{p_{12}(1 – p_{12})}} \right),$$

where $i = 1, 2$ and $(t_1, t_2) \in [0, 1]^2$. In our notation, $p_{11} = p_1(1, 1) + p_1(1, 0)$ is the probability of observing the first outcome in the $i$th group. All other probabilities are defined similarly. The quantity $\rho_1$ is the Pearson correlation coefficient in the $i$th group. Suppose that $(p_{11}, p_{12}, \rho_1) = (0.1, 0.1, 0.4)$ and $(p_{21}, p_{22}, \rho_2) = (0.15, 0.15, 0.4)$. Clearly $p_{11} < p_{21}$ for $j = 1, 2$, so the marginal distributions are ordered. However, $p_{11} = 0.0460 > 0.0225 = \rho_2(1, 1)$, so the joint distributions are not ordered by the multivariate stochastic order [see the first inequality in (2.2)]. A simple calculation shows that $p_{11} < p_{21}$ if $\rho_2 > 0.185$, in which case they will be ordered.

Theorem 1 is particularly useful in the context of MVBS because $X_{il} \preceq_{st} X_{jl}$ means that the marginal probabilities (i.e., the probabilities of observing specific tumors) are ordered. Thus

$$P_{il} = \sum_{t \in T_i} p_i(t) \leq \sum_{t \in T_j} p_j(t) = P_{jl} \tag{2.3}$$

for all $1 \leq l \leq m$, where $T_i = \{t : t_1 = 1\}$. Therefore, only $m$ inequalities, rather than $N_m$, need be considered (see Remark 1). For example, in the bivariate case we need to consider only the second and third inequalities in (2.2). For large $m$, the reduction in the number of constraints to be considered is extraordinary, and thus ordered inference on MVBS of any dimension is practical. We exploit this reduction in dimensionality in this article. It is particularly useful in the context of testing problems in which the hypothesis

$$H_0 : X_i =_{st} X_j \quad \text{versus} \quad H_1 : X_i \prec_{st} X_j \tag{2.4}$$

can be reformulated using Theorem 1 as

$$H_{0i} : \bigcup_{l=1}^{m} [X_{il} =_{st} X_{jl}] \quad \text{versus} \quad H_{1i} : \bigcup_{l=1}^{m} [X_{il} \prec_{st} X_{jl}] \tag{2.5}.$$  

The null in (2.5) was first referred to as simultaneous marginal homogeneity by Agresti and Klingenberg (2005).

The foregoing formulation can be extended from 2 to $K > 2$ groups as follows. Let $P = (P_1, \ldots, P_K)^T$, where $P_i = (p_{1i}, \ldots, p_{mi})^T$ are the marginal probabilities for the $i$th group as defined by (2.3). With this notation and using (2.5), we formulate our hypothesis as

$$H_0 : P \in P_0 \quad \text{versus} \quad H_1 : P \in P_1 \setminus P_0, \tag{2.6}$$

where $P_0 = \{P : P_1 = \cdots = P_K\}$, $P_1 = \{P : P_1 \leq \cdots \leq P_K\}$ and the inequalities defining $P_1$ are interpreted component-wise. Note that $P_0$ is a linear space and $P_1$ is a convex polyhedral cone, often called the simple order cone.
Remark 4. It is important to note that the null hypothesis in (2.6) is valid only if the marginal distributions are equal, that is, if \( X_j = X_{j_1} \cdots X_{j_r} \) for \( j = 1, \ldots, m \). Thus (2.6) can be tested without assuming that the multivariate distributions of \( X_1, \ldots, X_K \) are all equal. If we are willing to assume a priori (as we do here) that \( X_j \leq X_{j_1} \cdots X_{j_r} \), then rejecting the null in (2.6), we may conclude that the \( K \) MVBs are (strictly) ordered by the multivariate stochastic order. Otherwise, rejecting the null in (2.6) leads to the conclusion that some univariate distributions are strictly ordered by the univariate stochastic order.

For completeness, we note that Sampson and Whitaker (1989) and Lucas and Wright (1991) studied stochastically ordered distributions. They focused on bivariate discrete distributions, that is, on \( I \times J \) cross-classified data. Sampson and Whitaker (1989) developed min–max formulas for estimation, whereas Lucas and Wright (1991) extended Robertson and Wright’s (1981) projection formulas. Testing procedures with chi-bar limits were developed. These authors formulated their procedure not in terms of the parameters \( P_1, \ldots, P_K \) as in (2.6), but rather in terms of \( p_1, \ldots, p_K \), where \( p_i \) is the \( (2^m - 1) \)-dimensional probability mass function of \( X_i \). It is not surprising, in light of Remark 1, that these authors noted the difficulty of extending their methodology to high-dimensional data.

Following Pesaran (2001), in a series of articles Finos, Salmaso, and Solarli (2007) and Finos et al. (2008) developed union-intersection-type tests that do not explicitly incorporate any ordering. Their tests are based on combining \( p \)-values using prespecified \( p \)-value combination functions from pairwise one-sided tests and are implemented using permutation methods. Klingenberg et al. (2009) proposed a two-sample test for multivariate ordinal data. Their test is equivalent to the testing of homogeneity in a cumulative logit model and is implemented by assigning scores to the ordinal outcomes, with \( p \)-values computed by permutation methods. The foregoing body of work has been summarized by Basso et al. (2009).

In contrast to these approaches, our proposed methodology explicitly incorporates (1) the multivariate stochastic order constraints (i.e., that \( P_i \in P_0 \) under the null and that \( P_i \in P_1 \setminus P_0 \) under the alternative) and (2) the dependence among all binary RVs. Thus our methodology addresses the ordering in (2.6) in a very natural way. In addition, our methodology is completely model-free and does not require the specification of any extraneous quantities, such as scores or \( p \)-value combination functions. It is also important to note that unlike the aforementioned methods, our methodology can be applied when the data are not exchangeable and resampling methods fail. Animal studies subject to censoring by survival are an important example of such a situation (e.g., Peddada, Dinse, and Kissling 2007).

3. INFERENCES FOR ORDERED MULTIVARIATE BINARY DATA

3.1 The Proposed Methodology

Recall that \( P_i = (P_{i1}, \ldots, P_{im})^T \) is the vector of marginal probabilities for the \( i \)-th group. For \( i = 1, \ldots, K \) and \( j = 1, \ldots, m \), we may estimate \( P_{ij} \) by its maximum likelihood estimator (MLE), \( \hat{P}_{ij} = n_{ij}^{-1} \sum_{t_j \in T_j} N_i(t) \), where \( T_j \) is as defined in (2.3), \( N_i(t) \) is the random number of subjects in the \( i \)-th group with outcome \( t \), and \( n_i = \sum_t N_i(t) \) is the size of the \( i \)-th group.

It follows that \( \text{Var}(\hat{P}_i) = n_i^{-1} \Sigma_i \) where the variance matrix \( \Sigma_i \) is a function of \( \Pi_i \), the probability mass function of \( X_i \).

We also note that under the null in (2.6), \( P_i \) is estimated by \( \hat{P} = (\hat{P}_1, \ldots, \hat{P}_K) \), where

\[
\hat{P}_1 = \cdots = \hat{P}_K = \frac{\sum_{i=1}^K n_i \hat{P}_i}{\sum_{i=1}^K n_i}.
\]

We propose estimating \( P_i \) under the alternative in (2.6) by

\[
\tilde{P} = \Pi(\hat{P}/P_0) = \arg\min \{ (\hat{P} - P)^T \Sigma^{-1} (\hat{P} - P) : P_i \in P_1 \}.
\]

Thus \( \tilde{P} \) is the projection of \( \hat{P} \) onto the convex cone \( P_1 \) with respect to the matrix \( \Sigma \), defined as

\[
\Sigma = \text{Var}(\sqrt{n}(\hat{P} - P)) = \text{BlockDiag}(n/n_1, \ldots, n/n_K)^{-1},
\]

where \( n_i \) are the usual plug-in estimators; that is, the \( (j, j) \) element of \( \Sigma_i \) is given by \( n_{ij}^{-1} \sum_{a=1} n_{ij} (X_{ij}(a) - N_{ij}/n_i) (X_{ij}(a) - N_{ij}/n_i) \), where \( X_{ij}(a) = 1 \) if the \( a \)-th individual in the \( j \)-th group experienced outcome \( j \) and \( X_{ij}(a) = 0 \) otherwise, and \( N_{ij} \) is the number of individuals with the \( j \)-th outcome in the \( i \)-th group.

Note that the projection (3.1) accounts for the dependence among the components of \( \hat{P} \). Projection estimators of the type (3.1) have been proposed for a wide variety of constrained inference problems (see Silvapulle and Sen 2005 for a recent summary). The value of \( \tilde{P} \) can be readily obtained by quadratic programming.

Proposition 1. Let \( P \in P_1 \). If \( \min(n_1, \ldots, n_K) \to \infty \), then \( \tilde{P} \) is consistent for \( P \).

Note that although \( \tilde{P} \) is unbiased, the constrained estimator \( \tilde{P} \) is biased. This is typical of constrained estimators. The bias is negligible, however.

We propose testing (2.6) using the statistic \( D_n \), given by

\[
D_n = n \left( \| \hat{P} - \Pi(X_0 \hat{P})/P_0 \|_2^2 - \| \tilde{P} - \Pi(X_0 \tilde{P})/P_1 \|_2^2 \right),
\]

where \( \| x \|_2 \equiv (x^T V^{-1} x)^{1/2} \) and \( X_0 \) estimates the scaled variance of \( P \) under the null. The matrix \( \Sigma_0 \) may be estimated by \( \tilde{X}_0 = \text{BlockDiag}(n/n_1, \ldots, n/n_K)^{-1} \Sigma(n/n_1, \ldots, n/n_K) \), where \( \Sigma = \text{Var}(\hat{P}_i) = n_i^{-1} \sum_{i=1}^K n_i \hat{P}_i \). This estimator is appropriate if we further assume that the data are exchangeable under the null. If only marginal homogeneity is assumed, then we use (3.2) instead of \( \tilde{X}_0 \). Furthermore, observe that (3.3) simplifies to

\[
2 (\hat{P} - \tilde{P})^T \Sigma^{-1} (\hat{P} - \tilde{P}) - (\hat{P} - \tilde{P})^T \Sigma^{-1} (\hat{P} - \tilde{P})
\]

and has the form of a likelihood ratio statistic for an asymptotically normal random vector (Silvapulle and Sen 2005). In fact, \( D_n \) may be viewed as test based on a likelihood derived from a normal approximation to the joint distribution of the marginal means. We now provide the limiting distribution of the proposed test statistic under the null hypothesis.

Theorem 2. If \( P \in P_0 \) and \( n_i \to \infty \) with \( n_i/n \to \lambda_i > 0 \), then we have

\[
\lim_{n \to \infty} n \left( \frac{d_n - \chi^2_k}{\chi^2_k} \right) = \sum_{k=1}^{m(K-1)} w_k(X_0, \Sigma) \frac{d_n - \chi^2_k}{\chi^2_k} \geq x
\]

for some weights \( w_k(X_0, \Sigma) \), \( k = 1, \ldots, m(K-1) \), discussed later.
Thus, asymptotically, (3.3) follows a chi-bar distribution. The weights \( w_k(\Sigma_0, R) \), \( k = 1, \ldots, m(K - 1) \), are nonnegative and sum to unity. Note that the weights depend on the unknown but estimable variance of the unconstrained estimator under the null and the restriction matrix \( R \) defined in the Appendix. In general, the weights can not be computed in closed form whenever \( m(K - 1) \geq 5 \); however, the weights can be estimated by Monte Carlo methods. [For a thorough discussion on calculation of the weights and the simulation of chi-bar distributions see Silvapulle and Sen (2005, section 3.5) or Davidov, Fokianos, and Iliopoulos (2010).] To circumvent these problems, we test (2.4) using standard bootstrap methods, which we have found to be more powerful than permutation-based methods. We also have found that the (approximate) parametric bootstrap works well. The latter methodology compares the methods. We also have found that the (approximate) parametric bootstrapping guarantees their invertibility and is a widely used strategy in a variety of contexts (see Montgomery, Peck, and Vining 2001 or Tusher, Tibshirani, and Chu 2001 for examples). Alternatively, one advantage of the (approximate) parametric bootstrap is that it can be applied in situations where the data are not exchangeable and standard resampling methods do not hold.

We describe one such situation in the next section.

When dealing with binary variables with small “success” probabilities, it is inevitable to observe a small number of or even no “successes” in the sample, which could potentially lead to singular or near-singular variance matrices. To deal with this problem, following Teoh et al. (2008), we applied the classical ridge regression-based methodology by adding a small positive constant \( \delta \) to the diagonal elements of \( \Sigma \) and/or \( \Sigma_0 \). This modification guarantees their invertibility and is a widely used strategy in a variety of contexts (see Montgomery, Peck, and Vining 2001 or Tusher, Tibshirani, and Chu 2001 for examples). Although there is a considerable body of literature on the choice of \( \delta \), there is no consensus. Fortunately, in our specific situation, the exact value of the ridge parameter is not critical, for two reasons. First, our simulation studies suggest that the impact on the estimators is minimal for a reasonably wide range of values experimented in our simulation study. Second, and more interestingly, it can be shown that when the ridge parameter goes to 0, we obtain the restricted estimator, and as it goes to infinity, we actually recover the pooled adjacent violator estimator.

Finally, we remark that in the special case of \( m = 1 \), the multivariate methodology developed in this article generalizes the well-known procedures for the univariate case. The constrained estimator (3.1) reduces to the isotonic regression problem \( \hat{P} = \arg \min_i \{ \sum_{k=1}^{K} \hat{w}_i (\hat{p}_i - p_i)^2 : p_1 \leq \cdots \leq p_K \} \), where \( \hat{w}_i = \hat{w}_i / \hat{\delta}_i \). It is well known (e.g., Silvapulle and Sen 2005, p. 53) that the MLE in this case is \( \hat{P} = \arg \min_i \{ \sum_{k=1}^{K} n_i (\hat{p}_i - p_i)^2 : p_1 \leq \cdots \leq p_K \} \). Thus \( \hat{P} \) may be different from the MLE in some cases. Nevertheless, it can be shown that \( P^* = \hat{P} + \delta p_i(1) \). Moreover, if we recompute the weights, \( w_k \), when samples are merged in the pool adjacent violator algorithm, then \( P^* = \hat{P} \).

3.2 A Simulation Study

We conducted a simulation study to evaluate the performance of our proposed methodology. This study was designed as follows. We considered four test groups: a control and three “dose” groups (low, medium, and high) and sample sizes of \( n = 50 \) subjects per group as in our application. MVB data were generated using the method of Emrich and Piedmonte (1991) (see also Qaqish 2003). Various patterns for the marginal probabilities and correlation coefficients and the dimension of the problem were investigated. A subset of these patterns reported in this article are summarized in Table 1. Here we report only on trivariate MVBs, that is, \( m = 3 \). Results for \( m = 5 \) are provided in our Supplemental Materials. We considered three different patterns of correlation structures among the three binary outcomes: (a) independent structure; (b) serial correlation structure, where the correlation coefficients between components 1 and 2 and between components 2 and 3 were both 0.5, whereas the correlation coefficient between components 1 and 3 was 0.25; and (c) intraclass correlation structure with a correlation coefficient of 0.5 between any pair of components. All simulations were based on 1000 simulation runs, and the null distribution of the test statistics were derived empirically using 1000 bootstrap samples.

Estimates of the total mean squared error (TMSE) for the unrestricted and the order-restricted estimators are provided in Table 2. Note that TMSE is defined as the sum of the MSEs of all components. The TMSE is computed using \( \hat{\Sigma}_0 \), the pooled variance estimator, under the null and using the unpooled estimator \( \hat{\Sigma} \) under the alternative. In most cases we observe that the proposed estimator achieves a substantial reduction in the TMSE compared with the unconstrained estimator. In some cases, a reduction of almost 40% was noted.

We compared the estimated type I error rate and power of the proposed test statistic (\( D_n \)) with six competitors. The first

<table>
<thead>
<tr>
<th>Pattern of probability</th>
<th>Group 1 ( P_1 )</th>
<th>Group 2 ( P_2 )</th>
<th>Group 3 ( P_3 )</th>
<th>Group 4 ( P_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern 1</td>
<td>(0.05, 0.05, 0.05)</td>
<td>(0.05, 0.05, 0.05)</td>
<td>(0.05, 0.05, 0.05)</td>
<td>(0.05, 0.05, 0.05)</td>
</tr>
<tr>
<td>Pattern 2</td>
<td>(0.1, 0.1, 0.1)</td>
<td>(0.1, 0.1, 0.1)</td>
<td>(0.1, 0.1, 0.1)</td>
<td>(0.1, 0.1, 0.1)</td>
</tr>
<tr>
<td>Pattern 3</td>
<td>(0.2, 0.2, 0.2)</td>
<td>(0.2, 0.2, 0.2)</td>
<td>(0.2, 0.2, 0.2)</td>
<td>(0.2, 0.2, 0.2)</td>
</tr>
<tr>
<td>Pattern 4</td>
<td>(0.1, 0.1, 0.1)</td>
<td>(0.1, 0.1, 0.15)</td>
<td>(0.1, 0.1, 0.1)</td>
<td>(0.1, 0.1, 0.25)</td>
</tr>
<tr>
<td>Pattern 5</td>
<td>(0.1, 0.1, 0.1)</td>
<td>(0.1, 0.1, 0.15)</td>
<td>(0.1, 0.2, 0.15)</td>
<td>(0.1, 0.2, 0.25)</td>
</tr>
<tr>
<td>Pattern 6</td>
<td>(0.1, 0.1, 0.1)</td>
<td>(0.1, 0.2, 0.15)</td>
<td>(0.1, 0.2, 0.15)</td>
<td>(0.1, 0.2, 0.25)</td>
</tr>
<tr>
<td>Pattern 7</td>
<td>(0.1, 0.1, 0.1)</td>
<td>(0.1, 0.2, 0.15)</td>
<td>(0.1, 0.2, 0.2)</td>
<td>(0.1, 0.2, 0.25)</td>
</tr>
</tbody>
</table>
of these is the unconstrained Hotelling $T^2$-type procedure ($U_n$) based on the statistic $n(\hat{P} - \bar{P})^T \Sigma^{-1} (\hat{P} - \bar{P})$, which is equivalent to the test of Agresti and Klingenberg (2005). Note that $U_n$ is not designed for ordered alternatives. It also is important to note that Theorem 1 can be used to develop tests for (2.4) that combine $m$ marginal tests for order in various ways. These include union-intersection-type tests (UIT). See Sen (2007) for a review of UIT methodology for ordered alternatives and a Bonferroni test. In particular, in our simulation study we consider the max test ($M_n$), the sum test ($S_n$), a Bonferroni test ($B_n$), a Williams-type test ($W_n$), and the test of Finos, Salmaso, and Solari (2007) ($F_n$). All of these tests except $F_n$ are new.

The max test is based on the statistic $M_n = \max_{1 \leq j \leq m} D_{n}^{[j]}$, where $D_{n}^{[j]}$ is the statistic (3.3) based only on the $j$th variable, as explained at the end of the previous section. Note that $D_{n}^{[j]}$ is the usual univariate likelihood ratio test for order. The sum statistic is given by $S_n = \sum_{j=1}^{m} D_{n}^{[j]}$. The Bonferroni test ($B_n$) is based on the marginal $p$-values of the tests $D_{n}^{[j]}$, $j = 1, \ldots, m$, while adjusting for the number of tests. The Williams-type test (Williams 1971) is based on the statistic

$$W_n = \frac{\max_{1 \leq j \leq m} \sqrt{n}(P^{*}_{Kj} - P^{*}_{ij})}{\sqrt{\text{var}_{j}},}$$

where $P^{*}_{Kj} - P^{*}_{ij}$ is the difference between the constrained MLEs for the first and last groups and $\text{var}_{j}$ is the estimated variance. Because of their simplicity and good power, Williams-type tests are common in order-restriction inference. The $F_n$ test is somewhat more involved and is based on the combination of $p$-values. The test statistic is given by $F_n = \Psi(Q_1, \ldots, Q_m)$, where $Q_j$ is the $p$-value associated with the $j$th marginal test and $\Psi$ is some $p$-value combination function. In our simulations we used Fisher’s product method (for more details, see Finos, Salmaso, and Solari 2007).

Results of type I error rates and powers of the foregoing test statistics are provided in Table 3. We see that all test procedures
attain the nominal error rate of 0.05, although $U_n$, the Hotelling type $T^2$ statistic, was slightly liberal. In most cases considered in our simulation study, the proposed test $D_n$ has higher power than the alternative procedures when the data are correlated. In particular, $D_n$ has higher power than $F_n$ (Finos, Salmaso, and Solari 2007). The gain in power in the settings that we examined is usually substantial; for example, in some cases $D_n$ has double the power of $F_n$. As would be expected, the test statistic $F_n$ and the union-intersection-type tests $M_n$, $S_n$, $W_n$ are often comparable, and at times slightly superior, to $D_n$ in the independent case. Even then the gain in power by the best test (which varies by pattern) relative to $D_n$ is only minimal. We note that the Bonferroni test $B_n$ never performed better than $D_n$.

To summarize, $D_n$ is comparable to the best test under independence and much more powerful when the binary outcomes are dependent. We note that dependent outcomes are common in a variety of applications, such as the one considered here. Thus we conclude that the overall performance of $D_n$ in practical settings is likely to be superior to any of the alternatives considered.

4. APPLICATION TO ANALYSIS OF RODENT CANCER BIOASSAY

4.1 Modification of the Proposed Methodology

Toxicologists routinely conduct dose–response studies to evaluate the carcinogenicity of a chemical by examining multiple organs for tumors. For instance, a typical NTP bioassay involves exposing male and female rats and mice to one of several doses of a chemical of interest. Fifty animals are randomly assigned to each of the four dose groups, one of which is a control, and followed for 2 years. At the end of the study, all animals are necropsied, and various organs are evaluated for tumors. The NTP performs a statistical test to detect any dose-related trend in tumor incidence for each tumor type. Based on the statistical significance of the trend test and other biological evidence, the NTP determines whether or not a chemical is a carcinogen. This weight of evidence approach for declaring the statistical significance of the trend test and other biological covariates, terms appearing in (4.2) may be estimated empirically; for example, $\text{Cov}(O_i, W_j) = n_i \sum_{a=1}^{t_i} o_j(a) - o_j(n_i) (w_j(a) - w_j/n_i)$, where $o_j(a)$ is a binary RV indicating whether the $a$th animal in the $i$th dose group developed the $j$th tumor type before its death or sacrifice. The other covariances are estimated similarly.

Following the developments in Section 3, we define

$$\mathbf{P}' = \frac{\sum_{i=1}^{K} n_i \mathbf{P}}{\sum_{i=1}^{K} n_i}$$

for $i = 1, \ldots, K$,

$$\mathbf{P}^* = \Pi \mathbf{F}' \left( \mathbf{F}'^{-1} (\mathbf{P}') : \mathbf{P} \in \mathcal{P}_1 \right)$$
to be the estimators of $\mathbf{P}$ under the null and alternative, respectively. Here $\mathbf{S}_i^o = \text{BlockDiag}((n/n_1) \mathbf{S}_1^o, \ldots, (n/n_K) \mathbf{S}_K^o)$ is the estimated variance of $\mathbf{P}_i^o$. Furthermore, let

$$D_n^o = n \left\{ \left| \mathbf{P}_0 - \sum_{k=1}^{K} \mathbf{P}_k^0 \right| / \mathbf{S}_0^o \right\}^{1/2} - \left( \left| \mathbf{P}_0 - \sum_{k=1}^{K} \mathbf{P}_k^0 \right| / \mathbf{S}_0^o \right)$$

be the test statistic for (2.6).

It can be shown (by repeating the analysis in Proposition 1 and Theorem 2) that $\mathbf{P}_0$ is consistent and that $D_n^o$ follows a chi-bar distribution with weights $w_k(\mathbf{S}_0^o, \mathbf{R})$, $k = 1, \ldots, m(K - 1)$. The test $D_n^o$ may be applied using the (approximate) parametric bootstrap, as explained earlier.

Note that if survival rates differ among groups, then $\mathbf{S}_i^o$ will be generally different than $\mathbf{S}_j^o$. Nevertheless, our numerical calculations show that it is often better to estimate $\mathbf{S}_i^o$ by BlockDiag($n/n_1 \mathbf{S}_1^o, \ldots, n/n_K \mathbf{S}_K^o$), where $\mathbf{S}_i^o = n^{-1} \sum_{i=1}^{K} n_i \mathbf{S}_i^o$ is the pooled variance estimator.

### 4.2 A Simulation Study

We conducted a simulation study to evaluate the performance of the proposed modification of the methodology. The design of our simulation study mimics that of the NTP’s 2-year cancer bioassy studies and is similar to that of previous studies (e.g., Peddada, Dinse, and Haseman, 2005). We consider four test groups: a control and three dose groups (low, medium, and high), with sample sizes of 50 subjects per group.

Our simulation study honors the prevailing latent Weibull structure and the Poly-3 adjustment (Bailer and Portier 1988) assumed in rodent cancer bioassy studies. Thus, corresponding to each subject, we generated an $m$ variate normal random vector with mean 0 and covariance matrix $\Sigma = (\sigma_{ij})$, where $\sigma_{ij} = 1$ if $i = j$ and $\sigma_{ij} = \rho$ for $i \neq j$. Let $Q_{\lambda_i, \kappa_i}(\cdot)$ denote the quantile function of a Weibull RV with survival function given by $\exp(-\lambda_i \kappa_i t^\gamma)$. Note that $\lambda_i$ is the scale parameter and $\kappa_i$ is the shape parameter. Then $T_i(a) = Q_{\lambda_i, \kappa_i}(\Phi(Z(a)))$, where $a = 1, \ldots, n_i$ denotes the latent time to onset of the $j$th tumor for the $i$th subject in the $i$th group. By the foregoing construction, the components of $T_i(a) = (T_{i1}(a), \ldots, T_{im}(a))$ are correlated Weibull RVs. For each animal, we generated an additional independent Weibull RV $\Delta_i(a)$ that represents the time to death. The scale and shape parameters for $\Delta_i$ are denoted by $\beta_i$ and $\gamma_i$, respectively. Let $S$ be the duration of the study, 2 years in the present case. Then the $i$th animal in the $i$th group develops the $j$th tumor if $T_{ij}(a) \leq \min(\Delta_i(a), S)$. For $i = 1, \ldots, K$, $j = 1, \ldots, m_i$, and $a = 1, \ldots, n_i$ define the binary RVs $X_{ij}(a) = 1(T_{ij}(a) \leq \min(\Delta_i(a), S))$. Thus for each animal, we generate a vector of binary RVs and a time of death $\min(\Delta_i(a), S)$. This is exactly the information analyzed by the NTP.

We evaluated the performance of the test procedure $D_n^o$ for detecting an ordering among $m$ correlated tumors in a 2-year cancer bioassy in terms of type I error. All of our simulations are based on 1000 random realizations, and in each case the critical values for $D_n^o$ were generated using 1000 parametric bootstrap samples.

Although we considered a broad range of patterns, in this article we summarize only a small subset of the results that we obtained. Various patterns of animal survival and tumor rates are provided in Table 4. The patterns of survival rates considered in this simulation study ranged from no survival difference among the dose groups to somewhat large survival differences between the control and the high-dose groups. Similarly, we considered a wide range of tumor rate patterns, ranging from rare tumors (approximately 1% rate) to somewhat common tumors (up to 50%). We considered three correlated tumors with two patterns correlations coefficients, namely $\rho = 0.1$ and 0.9. Corresponding to each of these pattern types, we considered three different patterns of shape parameters for tumor incidence curves: 1.5, 3, and 6, as commonly done in this area (e.g., Peddada, Dinse, and Haseman 2005). Results of the simulations are summarized in Table 5. When the data conform to the Poly-3 model (i.e., the shape parameter governing the Weibull distribution for time to tumor is equal to 3), our proposed test maintains the nomi-

### Table 4. Patterns of trivariate tumor proportions under the null hypothesis and the patterns of probability of death before the end of the study in the 4 dose groups

<table>
<thead>
<tr>
<th>Case</th>
<th>Probability of tumors before the end of the study</th>
<th>Probability of death before the end of the study</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9 (0.01, 0.01, 0.01)</td>
<td>(10^-5, 10^-5, 10^-5, 10^-5)</td>
</tr>
<tr>
<td>2</td>
<td>0.9 (0.01, 0.01, 0.01)</td>
<td>(0.1, 0.1, 0.3, 0.4)</td>
</tr>
<tr>
<td>3</td>
<td>0.9 (0.01, 0.2, 0.5)</td>
<td>(10^-5, 10^-5, 10^-5, 10^-5)</td>
</tr>
<tr>
<td>4</td>
<td>0.9 (0.01, 0.3, 0.3)</td>
<td>(0.2, 0.2, 0.3, 0.4)</td>
</tr>
</tbody>
</table>

### Table 5. Type I error rates of the unconstrained test and the proposed test

<table>
<thead>
<tr>
<th>Case</th>
<th>Shape parameter of tumor incidence curve</th>
<th>Proposed test ($D_n^o$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.5</td>
<td>0.033</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>0.067</td>
</tr>
<tr>
<td>3</td>
<td>1.5</td>
<td>0.053</td>
</tr>
<tr>
<td>4</td>
<td>1.5</td>
<td>0.082</td>
</tr>
<tr>
<td>5</td>
<td>1.5</td>
<td>0.012</td>
</tr>
<tr>
<td>6</td>
<td>1.5</td>
<td>0.075</td>
</tr>
<tr>
<td>7</td>
<td>1.5</td>
<td>0.054</td>
</tr>
<tr>
<td>8</td>
<td>1.5</td>
<td>0.061</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.033</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.050</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.053</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0.056</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.012</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.053</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>0.054</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0.041</td>
</tr>
</tbody>
</table>

8 Journal of the American Statistical Association, ??? 2011
nal level. The nominal level is also maintained when the data-generating model is misspecified as the Poly-6 model; however, it can be liberal if the data are generated by (the misspecified) Poly-1.5 model and the survival differences among groups are large.

5. ILLUSTRATION: AN ANALYSIS OF THE NTP’S MALACHITE GREEN CHLORIDE DATA

As part of their toxicity and carcinogenicity evaluation of chemicals that humans are exposed to, the NTP recently conducted a 2-year cancer bioassay on malachite green chloride, an antifungal agent used in the fish industry (National Toxicology Program 2005). The bioassay included three dose groups of the chemical (100, 300, and 600 ppm) and a control group (0 ppm), with 48 female rats randomly assigned to each group. The animals were followed for 2 years. The incidences of mammary gland adenomas and pituitary gland adenocarcinomas in female rats are reported in Table 6. Neither of these tumor types displayed a significant dose-related trend according to the NTP’s Poly-3 trend test (p-values of 0.113 and 0.162, respectively). However, an increased incidence of mammary gland carcinoma in the highest-dose group was one of the reasons why the NTP concluded that there was equivocal evidence of carcinogenic activity in female rats. It is interesting to note that the NTP emphasized the finding of a significant increase in the incidence rates of pituitary gland adenocarcinoma in female rats from the control group to the 100-ppm group (p = 0.014). They also noted that the incidences in the higher-dose groups (i.e., 300 and 600 ppm) were well above the historical control range. Note that in the case of mammary gland adenoma, the observed rate in the 300-ppm dose group violates the hypothesized monotone order, whereas in the case of pituitary gland adenocarcinoma, the violation may be in the 100-ppm dose group.

It is well documented in the literature that the pituitary gland tumors may be associated with mammary gland tumors via the prolactin pathway (e.g., McComb et al. 1984; National Toxicology Program 2005). In view of this known correlation between the two tumors, we applied the proposed methodology to these data. The order restricted estimates are provided in Table 6. By doing so, we provide a framework for comparing treatment conditions that are ordered as in dose–response studies, time-course experiments, tumor stages in oncology, and so forth. In such situations the existing methodologies are not helpful for answering questions of practical interest, such as whether a set of responses changes (increases or decreases) monotonically with the dose of a chemical.

Comparison of two or more experimental conditions on the basis of multivariate binary response data is an important problem. Much of the existing methodology is designed to compare experimental conditions that are not ordered. However, in many instances the experimenter is interested in comparing treatment conditions that are ordered as in dose–response studies, time-course experiments, tumor stages in oncology, and so forth. In such situations the existing methodologies are not helpful for answering questions of practical interest, such as whether a set of responses changes (increases or decreases) monotonically with the dose of a chemical.

In this article we extend the existing literature in two important directions. First, we formulate the scientific question as a multivariate stochastic ordering problem among multiple groups. By doing so, we provide a framework for comparing two or more experimental conditions in terms of all n binary response variables jointly. In dose–response studies, such as those conducted by toxicologists, a fundamental assumption is that the response is monotonic in dose. We demonstrate that this assumption reduces the number of order restrictions substantially and allows the development of practical estimation and testing procedures for MVB data of any dimension. This allows true multivariate inference on the hypothesized ordering of the experimental conditions. Furthermore, our methodology does not require the specification or verification of any complicated modeling assumptions regarding the dependence among the binary variables. Our simulations suggest that the constrained estimators proposed in this article enjoy a substantial increase in efficiency (i.e., reduction in TMSE) relative to the standard unconstrained estimators. Similarly, our proposed test for multivariate stochastic ordering is more powerful than the unrestricted Hotelling $T^2$ type test (Agresti and Klinenberg 2005) and the Bonferroni correction-based methodology in the cases that we investigated. When the binary outcomes are dependent, the proposed test is more powerful than a host of UIT type tests. In particular, it improves considerably over the test of Finos, Salmaso, and Solari (2007), currently the only available

<table>
<thead>
<tr>
<th>Tumor type</th>
<th>Estimator</th>
<th>Control</th>
<th>100 ppm</th>
<th>300 ppm</th>
<th>600 ppm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mammary Gland</td>
<td>Unconstrained</td>
<td>0.050</td>
<td>0.052</td>
<td>0.023</td>
<td>0.130</td>
</tr>
<tr>
<td></td>
<td>Constrained</td>
<td>0.042</td>
<td>0.042</td>
<td>0.042</td>
<td>0.130</td>
</tr>
<tr>
<td>Pituitary Gland</td>
<td>Unconstrained</td>
<td>0.607</td>
<td>0.822</td>
<td>0.696</td>
<td>0.756</td>
</tr>
<tr>
<td></td>
<td>Constrained</td>
<td>0.609</td>
<td>0.758</td>
<td>0.758</td>
<td>0.758</td>
</tr>
</tbody>
</table>

6. DISCUSSION

We performed a simulation study to evaluate the type I error rate of our procedure under the null hypothesis that two correlated tumors had incidence rates of 0.06 and 0.73 for each of the four dose groups. These incidence rates were obtained by averaging the observed incidence rates of mammary gland and pituitary tumors in the current example. We used the same survival rates in our simulation study as in the NTP study. As before, we estimated the type I errors using 1000 bootstraps and 1000 simulation runs. In each case the type I error was close to the nominal level of 0.05. Thus we believe that the significant result obtained in the foregoing example is not a false-positive finding.
test in the literature. The gain in power is substantial in many cases. We found that when the binary outcomes are independent, there usually exists a test (among the tests $M_n$, $S_n$, $W_n$, or $F_n$) with power comparable to that of $D_n$. However, the best test among the foregoing varies according to the underlying pattern.

Our proposed methodology can be modified to handle the competing-risks problem commonly encountered in animal cancer bioassays. Essentially, in long-term animal cancer studies, an animal may die before the end of the study without developing tumors. In such cases, the estimates of tumor incidence rate must be corrected for the length of time that the animals were on the study. As routinely done by toxicologists in the case of univariate binary response, we used the Poly-3 methodology to estimate tumor incidence rates. As expected, the proposed methodology performs accurately (i.e., achieves the nominal type I error rate) when there are no survival differences among the dose groups. It also performs accurately (or conservatively) if the shape parameter in the Weibull model is 3 or greater, but may be liberal when the true shape parameter is 1.5, the tumors are somewhat common, and there are survival differences among groups. We note in this situation, the Poly-3 correction commonly used by toxicologists is incorrect. Thus it seems that inflated type I error rates are intrinsic to the methodology used by toxicologists when the model is misspecified. If it is of paramount importance, the type I error can be controlled by using larger values of the ridging constant $\delta$. Note that the constrained estimator is consistent for all values of $\delta \geq 0$. Perhaps, borrowing ideas from the well-established ridge regression theory, the ridge constant $\delta$ may be derived empirically using cross-validation or empirical Bayes methodology.

Finally, the framework provided in this article may be generalized in several directions. Our focus has been on MVB distributions, but the theory can be extended to multivariate ordinal and even multivariate continuous data. The method also could be extended to situations where a scientist is interested in comparing experimental conditions after adjusting for confounders or covariates. Such extensions may be feasible by couching the proposed methodology in generalized multivariate linear models, although this might not be trivial. In some applications, such as reproductive toxicology, MVB responses may be observed over time or generations. Such experimental designs induce correlations not only among the binary responses at given time (generation), but also over time.

**APPENDIX: PROOFS**

**Proof of Theorem 1**

Theorem 1 states that: (i) If $X_i \preceq X_j$ and $X_{il} \preceq X_{jl}$ for some $1 \leq l \leq m$ then $X_{il} \preceq X_{jl}$; and (ii) If $X_i \preceq X_j$ and $X_{il} \preceq X_{jl}$ then $X_{il} \preceq X_{jl}$ for some $1 \leq l \leq m$. We start with (i). Taken together the relationships $X_i \preceq X_j$ and $X_{il} \preceq X_{jl}$ for some $1 \leq l \leq m$ imply that $E[\phi(X_j)] \leq E[\phi(X_i)]$ for all increasing functions $\phi : \mathbb{R}^m \rightarrow \mathbb{R}$ with a strict inequality for some functions $\phi$. The validity of statement I is invariant to monotone transformations. Therefore, we may assume that the RVs $X_i$ and $X_j$ are non-negative. Suppose by contradiction that (ii) does not hold. The relationship $X_i \preceq X_j$ implies that $X_{il} \preceq X_{jl}$ for all $1 \leq l \leq m$. Because $X_{il} \preceq X_{jl}$ does not hold, we must have $X_{il} = X_{jl}$ for all $1 \leq l \leq m$. Let $\theta \in \mathbb{R}^m$ be nonnegative, that is, $\theta_l \geq 0$ for all $1 \leq l \leq m$. Because $X_i \preceq X_j$, we have $\theta^T X_i \preceq \theta^T X_j$, but the equalities $X_{il} = X_{jl}$ imply that $\theta^T X_{il} = \theta^T X_{jl}$. The positivity of $X_i$ and $X_j$ guarantees the existence of their Laplace transforms, so $L_{X_i}(\theta) = \mathbb{E}[\exp(-\theta^T X_i)] = \mathbb{E}[\exp(-\theta^T X_j)] = L_{X_j}(\theta)$. The equality of the Laplace transforms implies that $X_i \equiv X_j$, contradicting the fact that $X_i \preceq X_j$ and proving (ii).

**Proof of Proposition 1**

Let $P_0 \in P_1$ be the true value of $P$. Recall that $\tilde{P} = \Pi_{\Sigma_0}^\perp (\tilde{P})$ is the $\Sigma_0$-projection of the unconstrained MLE onto $P_1$, where $\Sigma_0$ denotes the true scaled variance of the MLE. By definition, $0 \leq (\tilde{P} - P_0)^T \Sigma_0^{-1}(\tilde{P} - P_0) \leq (\tilde{P} - P_0)^T \Sigma_0^{-1}(\tilde{P} - P_0) = o_P(1)$, because $n \rightarrow \infty$ implies the consistency of $\tilde{P}$. Thus $\tilde{P}$ and $\hat{P}$ have the same limit $P_0$, as required.

**Proof of Theorem 2**

Let $P_0 \in P_0$ be the true value of $P$. Note that (2.6) may be rewritten as $H_0 : RP = 0$ versus $H_1 : RP > 0$, where $R = (r_{ij})$ is a $(K-1) \times m$ matrix with elements $r_{ij} = -1$ if $i = j$, $r_{ij} = 1$ if $i = j + m$ and $r_{ij} = 0$ otherwise. By the multivariate central limit theorem,

$$\sqrt{n}(\hat{P} - P_0) \Rightarrow N(0, \Sigma_0).$$

The second equality follows by the consistency of $\Sigma_0$. Convergence follows from (A.1), where $Z$ follows a $\chi^2(K-1)$ distribution. The third equality follows from theorem 3.7.1 of Silvapulle and Sen (2005), where $P_1 \cap P_0^\perp$ is a convex cone ($P_0^\perp$ is the linear space orthogonal to $P_0$). Thus in Silvapulle and Sen’s (2005) notation, $D_n \Rightarrow \tilde{F}_0^\perp(\Sigma_0, P_1 \cap P_0^\perp)$ is a chi-bar square distributed RV. Because $\Sigma_0$ is of full rank and $R$ is a restriction matrix [a matrix whose rows are permutations of $(-1, 1, 0, \ldots, 0)$], it follows from their corollary 3.7.2 that

$$\lim_{n \rightarrow \infty} P(D_n \geq x) = \sum_{k=0}^{m(K-1)} w_k(\Sigma_0, R)\mathbb{P}(X_k^2 \geq x),$$

establishing (3.4).

**SUPPLEMENTARY MATERIALS**

**Appendix A:** The relationship between the orthant orders and the multivariate stochastic order, discussed in Remark 2, is investigated and clarified. (o-MVBs.SM.pdf)

**Appendix B:** Additional simulations for higher-dimensional data (i.e., $m = 5$) are reported on along with simulation results for some situations where the ordering assumptions are violated. (o-MVBs.SM.Tables.pdf)

[Received May 2010. Revised May 2011.]
REFERENCES


