

# Semiparametric Bayes Multiple Testing: Applications to Tumor Data

Lianming Wang and David B. Dunson

*Biostatistics Branch, MD A3-03*

*National Institute of Environmental Health Sciences*

*U.S. National Institutes of Health*

*P.O. Box 12233, RTP, NC 27709*

*\*Correspondence: wangl3@niehs.nih.gov*

*Abstract:* In National Toxicology Program (NTP) studies, investigators want to assess whether a test agent is carcinogenic overall and specific to certain tumor types, while estimating the dose-response profiles. Because there are potentially correlations among the tumors, joint inference is preferred to separate univariate analyses for each tumor type. In this regard, we propose a random effect logistic model with a matrix of coefficients representing log-odds ratios for the adjacent dose groups for tumors at different sites. We propose appropriate nonparametric priors for these coefficients to characterize the correlations and to allow borrowing of information across different dose groups and tumor types. Global and local hypotheses can be easily evaluated by summarizing the output of a single MCMC chain. Two multiple testing procedures are applied for testing local hypotheses based on the posterior probabilities of local alternatives. Simulation studies are conducted and a NTP tumor data set is analyzed illustrating the proposed approach.

*Key Words:* Dirichlet process; Logistic model; Mixture prior; Multiple testing; Nonparamet-

ric Bayes; Order constraint; Tumorigenicity.

## 1. Introduction

In a typical 2-year toxicology and carcinogenesis study conducted by the National Toxicology Program (NTP), male and female mice and rats are randomly assigned to a control group and three different dose groups having an increasing concentration of a test agent. These rodents are examined for tumors in different organ sites in a necropsy examination occurring after these animals die naturally or are sacrificed at 2 years. The main purpose of the study is to assess the evidence that the test agent is carcinogenic overall and specific to some types of tumors. The current standard analysis considers each tumor type separately and uses the Poly-3 adjustment to the survival time (Bailer and Portier, 1988) modifying the Cochran-Armitage trend test (Cochran, 1954; Armitage, 1955). However, separate analyses for many tumor types introduces multiplicity in testing and causes an inflated type I error rate. In this article, we propose to analyze the multiple tumor types jointly using a Bayesian approach to assess overall and site-specific evidence of carcinogenicity.

Multiple testing has been a popular research area in statistics, with the primary goal being to control the type I error rate without greatly reducing power. Many procedures have been proposed by either controlling familywise error rate (FWER) (e.g., Holm, 1979; Hochberg, 1988; Hommel, 1988; Westfall and Young, 1993 among others) or controlling false discovery rate (FDR) (Benjamini and Hochberg, 1995; Genovese and Wasserman, 2001; Storey, 2002, 2003; Storey et al. 2004 among others) at a nominal level. For reviews on multiple hypothesis testing, we refer to Shaffer (1995) and Dudoit et al. (2003).

As noted by Westfall and Wolfinger (1997), incorporating correlations into multiple-

ity adjustments could greatly improve the power of the test. Bayesian methods have been strongly recommended for multiple testing (Freedman 1996) because Bayesian methods automatically characterize the dependence structure among the hypotheses. Berry and Hochberg (1999) and Scott and Berger (2006) provided reviews of Bayesian perspectives on multiple comparisons. Gopalan and Berry (1998), Dahl and Newton (2007) and MacLehose et al. (2007) proposed use of Dirichlet process priors (Ferguson, 1973, 1974) to accommodate multiple comparisons. Müller et al. (2004) investigated the optimal sample size for multiple testing problems and developed Bayesian decision rules under different loss functions.

While there are numerous papers in the literature about multiple testing focusing on the case of continuous responses, there are rather limited papers considering the same problem for binary responses. We refer to Knoke (1976), Bristol (1993), and Chuang-Stein and Tong (1995) in frequentist literature and Chen and Sarkar (2004) in Bayesian literature. Chen (1996) proposed an estimating equations-based global test for assessing dose effects on multiple tumor sites. In this paper, we study correlated multivariate binary responses.

One difficulty of using Bayesian methods for multiple testing lies in specifying priors to characterize the dependence among the hypotheses (Freedman 1996). In this paper we overcome this difficulty through the following two steps. (1) We specify local null hypotheses as corresponding to elements in a matrix being set to zero. (2) We assign appropriate priors to the matrix of parameters so that these parameters have a positive probability to be equal to zero. The proposed priors induce correlations among the matrix of parameters and also allow borrowing information across the rows and columns of the matrix, which correspond to different dose groups and tumor types in the motivating application.

Logistic regression has been widely used for modeling binary responses. To model all

the tumor types jointly, we consider the following nested and crossed random effect logistic model,

$$\text{logit}(p_{ij}) = \tilde{\alpha}_j + \sum_{l=1}^k \beta_{lj} 1_{(l \leq d_i)} + \tilde{\phi}_i, \quad i = 1, \dots, n, \quad j = 1, \dots, J, \quad (1)$$

where  $p_{ij}$  is the probability of having the  $j$ th tumor for animal  $i$ ,  $d_i \in \{1, \dots, k\}$  is the dose group of animal  $i$ , and  $\tilde{\phi}_i$  is an animal-specific tumor susceptibility latent variable. In addition,  $\tilde{\alpha}_j$  is the log-odds of the  $j$ th tumor in the control group, and  $\beta_{lj}$  is the change in log-odds of tumor  $j$  attributable to increasing from dose group  $l - 1$  to dose group  $l$ . The frailties  $\tilde{\phi}_i$ 's characterize the heterogeneity among subjects, and are restricted to have mean zero for identifiability. Assuming that the test agent does not have a beneficial effect, we restrict  $\beta_{lj} \geq 0$ . Generalizations to allow down-turns at higher doses, which is common when dose affects survival times of animals, are straightforward using the ideas of Hans and Dunson (2005).

Model (1) is related to the random effects models of Meng and Dempster (1987) and Coull et al. (2001), which use normal prior distributions to borrow strength across dose effects on related outcomes, while also allowing dependence in the multiple outcomes. The normality assumption is questionable when a test agent can have no effect for some tumors, while having dramatic effects for other tumors. Hence, one of the goals of this article is to more flexibly borrow information across animals, tumor types and dose group comparisons using nonparametric Bayes machinery. For articles on parametric multiple tumor modeling, we refer to Lu and Malani (1995) and Dunson and Dinse (2002).

To test whether the test agent is carcinogenic overall, we can formulate the following

global hypothesis:

$$H_0 : \quad \beta_{lj} = 0, \quad l = 1, \dots, k, \quad j = 1, \dots, J$$

$$H_a : \quad \text{At least one } \beta_{lj} > 0, \quad \forall j, l.$$

To test whether the test agent is carcinogenic to specific tumor types, we construct the following local hypotheses,

$$H_{0j} : \quad \beta_{1j} = \dots = \beta_{kj} = 0$$

$$H_{aj} : \quad \text{At least one } \beta_{lj} > 0, \quad \forall l$$

for  $j = 1, \dots, J$ . Clearly, the global null hypothesis is the intersection of all the local null hypotheses. Rejecting any local hypothesis results in rejecting the global null hypothesis. In addition to global and local hypothesis testing, our proposed method allows estimation of tumor site-specific dose response and identification of posterior distributions for low observed adverse effect levels (LOAEL). LOAELs are often of interest from a regulatory perspective, as the lowest dose for which there is a change relative to control.

## 2. Model and method

Let  $y_{ij}$  denote the binary indicator for tumor type  $j$  and subject  $i$ , with 1 indicating that the  $j$ th type tumor is observed in animal  $i$  and 0 otherwise. Denote  $x_{il} = 1_{(l \leq d_i)}$  for  $l = 1, \dots, k$ ,  $i = 1, \dots, n$ ,  $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})'$ ,  $\boldsymbol{\beta}_j = (\beta_{1j}, \dots, \beta_{kj})'$  for  $j = 1, \dots, J$ , and  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_J)$  is a  $k \times J$  matrix of dose effect coefficients. Then we can rewrite (1) as

$$\Pr(y_{ij} = 1 \mid \mathbf{x}_i) = \frac{\alpha_j \phi_i \exp(\mathbf{x}_i' \boldsymbol{\beta}_j)}{1 + \alpha_j \phi_i \exp(\mathbf{x}_i' \boldsymbol{\beta}_j)}, \quad (2)$$

where  $\alpha_j = \exp(\tilde{\alpha}_j)$  for  $j = 1, \dots, J$  and  $\phi_i = \exp(\tilde{\phi}_i)$  for  $i = 1, \dots, n$ . The global null

hypothesis  $H_0$  is equivalent to letting  $\boldsymbol{\beta} = \mathbf{0}_{k \times J}$ , while the local null hypothesis  $H_{0j}$  sets the  $j$ th column of  $\boldsymbol{\beta}$  equal to zero.

Since our main interests focus on testing hypotheses, which amounts to checking whether the corresponding  $\beta_{lj}$ 's equal zero, the priors for  $\beta_{lj}$ 's play an essential role. The desirable priors should induce general dependence structures among the  $\beta_{lj}$ 's while also allowing efficient posterior computation.

### 2.1. Prior structure for $\beta_{lj}$ 's

We consider the following hierarchical priors for the elements of  $\boldsymbol{\beta} = \{\beta_{lj}\}$

$$\begin{aligned} \beta_{lj} | (\tilde{G}_l, \pi_l) &\stackrel{iid}{\sim} G_l = \pi_l \delta_0 + (1 - \pi_l) \tilde{G}_l, \quad j = 1 \cdots, J, \\ \tilde{G}_l &\stackrel{iid}{\sim} DP(M_l \tilde{G}_{0l}), \quad \pi_l \stackrel{iid}{\sim} Beta(1, c), \quad c \sim \pi(c | c_0), \end{aligned} \quad (3)$$

where  $\delta_0$  is a point mass probability measure at 0,  $\pi(c|c_0)$  is a general prior for  $c$  with parameter  $c_0$ , and  $\tilde{G}_l \stackrel{iid}{\sim} DP(M_l \tilde{G}_{0l})$  denotes that distribution  $\tilde{G}_l$  is assigned a Dirichlet process (DP) prior with base distribution  $\tilde{G}_{0l}$  and precision parameter  $M_l$ .

DP priors were originally proposed for use in multiple testing by Gopalan and Berry (1998), who used the DP to allow equalities in the distribution in different treatment groups. Dahl and Newton (2007) recently proposed a closely-related specification to Gopalan and Berry (1998), but they used the DP to borrow information across multiple outcomes in performing treatment group comparisons, motivated by gene expression applications. MacLehose et al. (2007) instead used a mixture of a DP and a point mass at zero as a flexible shrinkage prior for the coefficients in a high-dimensional logistic regression model. In the multiple tumor testing application, such specifications are not appropriate, since instead of having a high-dimensional vector of coefficients for exchangeable predictors, we have a

matrix of coefficients characterizing treatment group contrasts for each tumor.

The primary difference between (3) and the specification of MacLehose et al. (2007) is the inclusion of a separate zero-inflated DP (ZI-DP) prior on the distribution of the coefficients characterizing the increase in log-odds of tumor attributable to each increment in dose. In particular, we have  $G_l \sim \text{ZI-DP}(\pi_l, M_l, \tilde{G}_{0l})$ , for  $l = 1, \dots, k$ , with  $l$  indexing the comparison between dose groups  $l$  and  $l+1$ . The prior distribution,  $G_l$ , controls borrowing of information across the different tumor types in estimating the increase in log odds of tumor attributable to increasing dose from level  $l$  to  $l+1$ , with the mass at zero corresponding to tumor types with no difference in tumor incidence in these two adjacent dose groups. By avoiding the assumption of  $G_l = G$ , we allow the magnitude of the coefficients and proportion of zero coefficients to vary between each of the dose group contrasts, which seems well justified. We borrow information in modeling the  $G_l$ 's through use of a hyperprior for a parameter  $c$  that is shared in the priors for the point mass probabilities,  $\pi_1, \dots, \pi_k$ . For example, if there is no evidence in the data of a carcinogenic effect on any of the tumor types, we would expect the posterior of  $c$  to be concentrated on values close to zero, favoring values close to one for  $\pi_1, \dots, \pi_k$ . Hence, the prior allows borrowing of information both across rows (dose group contrasts) and columns (tumor types).

Note that (3) assigns equal prior probability to all of the local null hypotheses, which is reasonable given that it is seldom the case that one has prior knowledge of tumor sites that are more likely to be affected by the chemical exposure. Because of the assumption of equal prior probabilities, differences in the estimated posterior probabilities are entirely driven by the data. The prior specification (3) assigns a positive dependence among all  $\beta_{lj}$ 's. For example, knowing the event  $\beta_{lj} = 0$  ( $\beta_{lj} > 0$ ) will enlarge the chance of  $\beta_{l'j'} = 0$  ( $\beta_{l'j'} > 0$ )

since it is easy to verify

$$\Pr(\beta_{l'j'} = 0 | \beta_{lj} = 0) > \Pr(\beta_{l'j'} = 0)$$

and

$$\Pr(\beta_{l'j'} > 0 | \beta_{lj} > 0) > \Pr(\beta_{l'j'} > 0)$$

for any  $l' \neq l$ ,  $j' \neq j$ . Such a positive dependence structure is well motivated in the multiple tumor application, since the  $\beta_{lj}$ s all correspond to increments on the log odds of tumor development given increments on dose in a common exposure. Knowledge that one  $\beta_{lj}$  is high will naturally lead to an increased prior expectation that the chemical is carcinogenic and hence other  $\beta_{lj}$ s may be high as well.

Under the prior specification in (3), the probabilities of local and global null hypotheses are in the following forms:

$$\Pr(H_{0j}) = E_c \left[ \frac{1}{1+c} \right]^k \quad \text{and} \quad \Pr(H_0) = E_c \left[ \frac{J!}{\prod_{j=1}^J (c+j)} \right]^k. \quad (4)$$

In (4),  $E_c$  denotes expectation with respect to  $c$ . Since  $c_0$  is the only parameter in the prior distribution of  $c$ , we can easily assign the prior probability of the global null hypothesis to be approximately 0.5 by choosing an appropriate value of  $c_0$ . We also specify priors for the other parameters in (2). For example, we assign a DP prior to  $\alpha_j$ 's in order to borrow information across tumor types in a related manner to Dahl and Newton (2007). Posterior computation is efficient relying on the data augmentation approach of Dunson and Stanford (2005) combined with the blocked Gibbs sampler of Ishwaran and James (2001). All these details are shown in Appendices A and B.

Based on draws from the Gibbs sampler, we can calculate the posterior probability of

the global null hypothesis as

$$\Pr(H_0|Data) \approx \frac{1}{m - m_0} \sum_{h=m_0+1}^m \mathbf{1}_{\{\beta_{lj}^{(h)}=0, l=1,\dots,k, j=1,\dots,J\}},$$

and the posterior of the local null hypothesis as

$$\Pr(H_{0j}|Data) \approx \frac{1}{m - m_0} \sum_{h=m_0+1}^m \mathbf{1}_{\{\beta_{lj}^{(h)}=0, l=1,\dots, k\}}, \quad j = 1, \dots, J,$$

where  $m$  is the total number of iterations,  $m_0$  is the number of iterations in burn-in period, and  $\beta_{lj}^{(h)}$  is the realization of  $\beta_{lj}$  in the  $h$ th iteration. The posterior probabilities of the global and local alternative hypotheses are  $1 - \Pr(H_0|Data)$  and  $1 - \Pr(H_{0j}|Data)$ ,  $j = 1, \dots, J$ , respectively.

## 2.2. Bayesian multiple testing procedures

To determine which tumor types are significantly affected, we need to conduct local hypotheses testing. Posterior probabilities can be used for decision-making since these probabilities measure the evidence of true alternatives. A natural method is to reject each local null hypothesis if the posterior probability of the corresponding local alternative is larger than a pre-specified large value, e.g., 0.90. Notice that 0.90 corresponds to 9 for using Bayes factor in the case that the prior probabilities of null and alternative are equal.

An alternative method is to adopt the Bayesian multiple testing procedure proposed in Müller et al. (2004). We use the posterior probability  $v_j = \Pr(H_{aj}|Data)$  as a test statistic and for each  $j$  reject  $H_{0j}$  if  $v_j \geq r$ , where  $r$  is a common threshold for all the local hypotheses. We calculate the posterior expected FDR and false negative rate (FNR) as functions of  $r$  in the following forms:

$$\overline{\text{FDR}}(r) = \frac{\sum_j \mathbf{1}_{(v_j \geq r)}(1 - v_j)}{\sum_j \mathbf{1}_{(v_j \geq r)} + \epsilon} \quad \text{and} \quad \overline{\text{FNR}}(r) = \frac{\sum_j \mathbf{1}_{(v_j < r)}v_j}{J - \sum_j \mathbf{1}_{(v_j < r)} + \epsilon},$$

where  $\epsilon$  is a small positive constant to avoid zero denominator. In Müller et al. (2004), several goals are considered based on  $\overline{\text{FDR}}$  and/or  $\overline{\text{FNR}}$ . Here, we choose to minimize  $\overline{\text{FNR}}$  subject to  $\overline{\text{FDR}} \leq \alpha_D$ , where  $\alpha_D$  is a pre-specified value, say 5%. In this case, the optimal threshold  $r^*$  is determined by  $r^* = \min(r \in [0, 1] : \overline{\text{FDR}}(r) \leq \alpha_D)$ . For each  $j$  we then reject  $H_{0j}$  if  $v_j \geq r^*$ .

In this paper, the prior structure (3) essentially models a matrix of dependent coefficients  $\beta_{lj}$ 's and allows us to test  $\beta_{lj} = 0$ 's. The testing methods mentioned above can also apply for testing

$$H_{0lj} : \beta_{lj} = 0 \text{ vs. } H_{alj} : \beta_{lj} > 0$$

for all  $l = 1, \dots, k$  and  $j = 1, \dots, J$  if these hypotheses are of primary interest.

### 3. Application to NTP Study of Isoprene

#### 3.1. Description and Frequentist Results

The data come from a 2-year inhalation study of isoprene in male rats conducted by NTP. Isoprene was selected for toxicologic evaluation because of its potential harm for human exposure due to its large annual production volume. In the study, groups of 50 male rats were exposed to 0, 220, 700, or 7000 ppm isoprene by inhalation, 6 hours per day, 5 days per week, for 105 weeks. These rats either died naturally or were sacrificed at the end of the study, and were examined for tumor status at various organs at their death times. We concentrate on the data with tumor status at 24 different organs. Let  $y_{ij}$  denote a binary response indicating whether at least one tumor was detected at organ  $j$  for animal  $i$  at its death time. The third to six columns in Table 1 show the number of male rats that were detected to have at least one tumor in the 24 organs, in the control, low-, medium- and high-

dose groups respectively. It appears from Table 1 that there is an increasing dose response for kidney, testes and mammary gland, respectively. However, it is not obvious based on a simple subjective examination of the data that isoprene is carcinogenic.

The NTP analyzed each organ separately as seen in the NTP technical report No. 486 available at <http://ntp.niehs.gov>. A poly-3 trend test modifying the Cochran-Armitage linear trend test to take survival differences into account was used for testing each organ type and reported that three organs (kidney, testes and mammary gland) showed clear evidence of carcinogenicity of isoprene based on the p-values of the poly-3 trend test. These p-values are 0.009,  $< 0.001$ , and  $< 0.001$  for kidney, testes and mammary gland respectively (NTP Tr-486, pp. 84-87). The p-values of poly-3 tests for other organs are larger than 0.10.

Considering that there are 24 organs to evaluate, one needs to adjust for multiple testing to avoid an inflated type 1 error. Controlling  $\text{FWER} \leq 0.05$  in this application, Bonferroni adjustment takes  $0.05/24 = 0.002$  as a cutoff point and rejects any null hypothesis if the corresponding p-value is less than 0.002. Clearly kidney is no longer significant in this multiple testing setup under Bonferroni adjustment. As we know, Bonferroni adjustment is too conservative. Another widely used approach dealing with multiplicity is the Benjamini and Hochberg (1995) algorithm, which controls FDR at a pre-specified level, e.g., 0.05. Their procedure is as follows for this application: First label the p-values in ascending order such that  $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(J)}$  and denote by  $H_{(i)}$  the null hypothesis corresponding to  $p_{(i)}$ . Second define  $i_0$  to be the largest  $i$  for which  $p_{(i)} \leq 0.05i/J$ . The decision rule is to reject all  $H_{(i)}$  for  $i = 1, \dots, i_0$ . Applying this algorithm, the three cutoff points to be compared with the three smallest p-values are 0.002, 0.004 and 0.006, and one only rejects the two hypotheses with the smallest p-values. Again, kidney is not significant using this multiple

testing procedure. In the following we reanalyze the data and conduct multiple testing with the proposed approach. One advantage of the proposed approach over Bonferroni adjustment and Benjamini and Hochberg (1995) method is that the proposed approach characterizes and utilizes the dependency among the tested hypotheses.

### 3.2. Analysis and Results for the Proposed Approach

We reanalyze the multiple tumor data with the proposed approach. We assign a  $\mathcal{G}(1, 1)$  prior to  $\nu$  and DP precision parameters  $M_l$  for  $l = 1, \dots, k$  and  $M_0$ . The DP base measures  $P_0$  and  $G_{0l}$  are taken to be  $\mathcal{G}(1, 1)$  and  $\mathcal{G}^+(1, 1)$ , a truncated  $\mathcal{G}(1, 1)$  above 1, respectively. The definitions of  $\nu$ ,  $M_l$ ,  $G_{0l}$ ,  $P_0$  are provided in the Appendix. We set  $c_0$  to be 0.287 to allow  $\Pr(H_0) = 0.5$  under the prior structure (3) or (7) as shown in Appendix. These priors are pretty general and realistic, so we recommend these priors as default values in other analyses of multi-site tumorigenicity data. We collect 10000 samples from the MCMC output after a 2000-iteration burn-in.

The estimated posterior probability of the global alternative is  $\Pr(H_a | Data) = 1$ , indicating that isoprene is carcinogenic to male rats and has a significant effect on at least one organ in the sense that the tumor occurrence rates increase significantly due to the effect of isoprene at these organs. We then conduct the multiple local hypotheses testing in order to identify the affected organs. As mentioned in subsection 2.2, we make decisions based on the posterior probabilities of local alternatives. These probabilities are plotted in Figure 1 and also presented in the last column in Table 1 as the evidence of dose effect between the high-dose group and the control group. For the first method, we take 0.90 as a cutoff point and reject  $H_{0j}$  if  $v_j > 0.90$ . The choice of 0.90 corresponds to the following equivalent

decision rule based on Bayes factor: rejecting  $H_{0j}$  if the corresponding Bayes factor is larger than 20.12 in this case. Based on this decision rule, we obtain three positive discoveries as clearly seen in Figure 1 and Table 1. These positive discoveries correspond to kidney, testes, and mammary gland. Applying the second method (Muller et al., 2004), we obtain the optimal threshold  $r^* = 0.83$  by controlling  $\overline{\text{FDR}} \leq 0.05$  and thus reject  $H_{0j}$  if  $v_j \geq 0.83$  for each  $j$ . This decision rule gives the same conclusion as the first method. Notice that the NTP technical report gave the same conclusion but was based on univariate analysis. Figure 6 shows the estimated trajectories of the tumor occurrence rates for kidney, testes, and mammary gland.

After concluding that tumor incidence is significantly increased due to isoprene exposure at a particular organ site, it is of interest to identify which dose levels differ from control in that site. The last three columns in Table 1 show the evidence of dose effect between each dose group and the control group in terms of posterior probability  $\Pr(\sum_{l'=1}^l \beta_{l'j} > 0 | \text{Data})$ . The posterior probabilities in different columns are not comparable since the corresponding prior probabilities are different. To provide a uniform basis for comparison, we consider Bayes factor  $BF_{lj} = \frac{v_{lj}/(1-v_{lj})}{\tilde{v}_{lj}/(1-\tilde{v}_{lj})}$  for  $l = 1, 2, 3$  and  $j = 9, 15, 23$ , where  $v_{lj}$  and  $\tilde{v}_{lj}$  are the posterior and prior probabilities of the event  $\{\sum_{l'=1}^l \beta_{l'j} > 0\}$  respectively. Table 2 shows the Bayes factors for kidney, testes, and mammary gland, and these results suggest that the medium dose (700ppm) of isoprene starts to show an increase in tumors at the testes and mammary gland, while kidney tumors are increased only by the high-dose (7000ppm) of isoprene.

In this study, the survival times of rats were also recorded. Potentially, inferences on chemical effects can be biased by differences in survival between the dose groups. To adjust

for survival, we introduce a weight term  $w_{ij}$  in the mean of the Poisson latent variable  $z_{ij}$  defined in (5) in Appendix A:

$$z_{ij} \sim \text{Poisson}(w_{ij}\alpha_j\phi_i e^{\mathbf{x}'_i\boldsymbol{\beta}_j}\xi_{ij}),$$

where  $w_{ij} = 1$  if  $y_{ij} = 1$ , i.e., at least one tumor was found at the  $j$ th organ for animal  $i$ ; Otherwise  $w_{ij} = (t_i/T_0)^3$ , where  $t_i$  is the survival time for animal  $i$  and  $T_0$  is the duration of the study, which is 105 weeks in this example. This adjustment is similar to the poly-3 trend test in Bailer and Portier (1988) modifying the Cochran-Armitage trend test. We then implement the proposed approach incorporating the survival adjustment and identify the same positive discoveries. The survival adjustment does not improve much the performance of the proposed method in this example. This is not surprising since survival times are very similar among 4 groups (NTP Tr-486, pp 25-26).

Our proposed approach combines joint analysis of all tumors (organs) and a multiple testing procedure. The joint model is important in that it characterizes the correlation among multiple tumors and further the correlation among the local hypotheses. To see the importance of the joint model, we consider the following analysis for comparison with the proposed approach. We conduct Bayesian univariate analyses on each organ data separately with a logistic model closely related to model (1) in order to obtain the posterior probability of each local alternative  $v_j$ , and then apply the multiple testing procedures mentioned in subsection 2.2. Both methods fail to reject the local hypothesis corresponding to kidney.

#### 4. Simulation studies

To evaluate the proposed approach, we mimic the real data for the NTP experiments with simulations focusing on the following three cases: (1) null case:  $H_0$  is true, (2) sparse case:

only a few local null hypotheses are true, and (3) general case. When generating data, we take all  $\beta_{lj} = 0$  in the null case, all  $\beta_{lj} = 0$  except that  $\beta_{lj} = 0.3$  for  $l = 1, 2, 3$  and  $j = 9, 15, 23$  in the sparse case, and the following matrix as the true values of  $\beta_{lj}$ 's, for  $l = 1, 2, 3$  and  $j = 1, \dots, 24$ ,

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} * 0.3$$

in the general case. Note that  $\beta_{lj} = 0.3$  indicates a moderate effect between the  $(l - 1)$ th and  $l$ th dose groups for tumor type  $j$ . We generate each  $\log(\alpha_j)$  from  $N(-1, 1)$  to allow a large variation across occurrence rates of different types of tumors in the control group and set  $\nu = 1/20$  to allow small heterogeneity among tumors similarly to case of the real data considered since  $\nu$  is the variance of the frailties. To match the NTP study in Section 3, we let  $n = 200$  resulting from 50 animals per group, with one control and three groups with increasing dose levels giving  $k = 3$ , and set  $J = 24$  indicating that 24 different tumor types are to be considered. Note that the sample size of 200 is quite modest given that there are  $24 \times 3 = 72$   $\beta_{lj}$  dose effect parameters to be estimated, with an additional 24 intercept parameters and a tumor susceptibility random effect specific to each animal. However, to realistically characterize the real data we cannot discard any of these parameters *a priori*.

We use the same priors as we recommended in Section 3 and set  $c_0 = 0.287$  to ensure the prior probability of the global null hypothesis  $\Pr(H_0) = 0.5$  in all cases. For each setup, we compare our approaches with the Bonferroni adjustment (CA-B) and Benjamini and Hochberg procedure (CA-BH) based on the individual p-values from the Cochran-Armitage trend tests for the local hypotheses. Note that the Poly-3 test can not be used here since there are no survival times available. To evaluate the accuracy of parameter estimation, we

also calculate the MSE of all the estimates of  $\beta_{ij}$ 's from the joint analysis and compare it to the MSE of the maximum likelihood estimates (MLEs) from separate univariate analyses. The simulation results are based on 100 repeated data sets.

In testing the global hypothesis, the proposed approach rejects the global null if the posterior probability of the global null  $P(H_0|Data) > 0.95$ , and the multiplicity adjusted Cochran-Armitage trend tests (both CA-B and CA-BH) reject the global null if any of the p-values from the Cochran-Armitage trend tests for individual local hypotheses is smaller than  $0.05/24$ . Table 3 summarizes size (type-1 error) in the null case and power in the sparse and general cases for both methods. As seen in Table 3, both of the two methods give a type 1 error equal to 0.05 in the null case and show a comparable power in the sparse case (0.47 vs. 0.45). However, the proposed method show a much larger power than the adjusted CA test in the general case (0.92 vs. 0.66).

To evaluate the local hypotheses, we calculate the averages of the observed FDRs and powers based on 100 data sets using the two proposed approaches mentioned in Section 2.2, CA-B, and CA-BH respectively. Comparison results are listed in Table 4. As seen in Table 4, the proposed methods give a FDR around 0.05 in all the cases. The proposed approach performs similarly to the CA-B and CA-BH in the null and sparse cases but the proposed approach shows a much larger power in the general case considered.

Regarding the accuracy of estimating the  $\beta$ , the average of MSE from the proposed method is much smaller than that from the maximum likelihood method, with 0.029 vs. 0.527 in the null case, 0.111 vs. 0.532 in the sparse case, and 0.175 vs. 0.540 in the general case considered above.

In all the simulation cases, the MCMC samples converge very quickly and the mixing

is good. This is due to the efficient Gibbs samplers shown in the Appendix, in which all the parameters except  $\nu$  have a closed form in their full conditional distributions. We apply a Metropolis-Hasting algorithm to update  $\nu$  using a normal random walk with a step size 0.05, which gives an acceptance rate close to 30% - 40% in all cases. Our priors are not overly concentrated, but assign high probability to a wide range of plausible values in the tumor applications. We find results are robust to moderate changes in the hyperparameter choice. For example, changing  $\mathcal{G}(1, 1)$  to  $\mathcal{G}(0.1, 0.1)$  as a prior for  $\nu$  or  $M_l$  does not change the conclusions in the above simulations, though we note that  $\mathcal{G}(0.1, 0.1)$  is overly-diffuse.

## 5. Concluding remarks

In NTP studies, a committee of experts makes a decision about the weight of evidence that a chemical is carcinogenic. This decision is based primarily on the results of many different univariate analyses conducted for each tumor separately, without adjustment for multiple comparisons or within-animal dependence in the multiple tumors. The decision of the committee is quite important, forming the basis for later regulatory decisions, and hence is often criticized by industry and special interest groups. The practice of making a call based on toxicologists and other experts subjective synthesis of the weight of evidence in the data naturally opens up the NTP to criticism, particularly in cases in which it is not obvious that the significant dose response trends in individual tumors are not false positives. Hence, it is important to have automated statistical procedures available, which provides an overall weight of evidence in the data, adjusting for the complicating features of multiple testing, within-animal dependence and differential survival.

The approach proposed in this article provides such a methodology, combining a Bayesian

joint analysis of all tumor types and a Bayesian multiple testing procedure. In addition, to avoid criticism of sensitivity to subjectively-chosen priors, we have used a highly flexible semiparametric Bayes approach, with hyperparameters carefully chosen to assign high probability to a wide range of plausible values given our experience in analyzing tumorigenicity data. These choices were validated through simulation studies, which assess frequentist operating characteristics under a variety of scenarios. These simulations demonstrated that our Bayesian multiple testing procedure produces a valid and powerful frequentist test for overall carcinogenicity, while also doing a good job of identifying sites of action and estimating dose response parameters. Implementation of the proposed approach involves a straightforward and efficient MCMC algorithm, which can be routinely used in the data analysis phase for NTP studies and other dose response studies. The program will be made publicly available and is straightforward to implement for non-experts.

Although our focus was the multiple tumor application, the proposed approach provides a general methodology for Bayesian multiple testing of *matrix hypotheses*. For matrix hypotheses, one is interested in testing for equalities in the columns of the matrix separately for each of the rows. For example, in gene expression studies the rows may correspond to genes and the columns to treatment groups. In clinical trials or epidemiologic studies, the rows may instead correspond to different subgroups of individuals, defined by genes, ethnic groupings, age or other factors, while the columns correspond to levels of a treatment or environmental exposure. There is increasing interest in personalized medicine and gene x environment interaction studies in allowing differences in treatment or exposure effects between sub-groups. Our proposed approach allows efficient testing of the large numbers of closely-related hypotheses that are considered in such studies through flexible borrowing of

information across the rows and columns in the matrix.

## Acknowledgement

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

## Appendix

### A. More prior specifications

We rely on a data augmentation for efficient posterior computation. First note that model (2) can be equivalently rewritten in the following form containing an underlying Poisson variable (Dunson and Stanford, 2005):

$$\begin{aligned} y_{ij} &= 1_{(z_{ij}>0)} \\ z_{ij} &\sim \text{Poisson}(\alpha_j \phi_i e^{\mathbf{x}'_i \boldsymbol{\beta}_j} \xi_{ij}) \\ \xi_{ij} &\sim \text{Exp}(1) \end{aligned} \tag{5}$$

Then the likelihood conditional on frailties based on (5) can be written as

$$L = \prod_i \left\{ \prod_j y_{ij}^{1_{(z_{ij}>0)}} (1 - y_{ij})^{1_{(z_{ij}=0)}} \text{Poisson}(z_{ij} | \alpha_j e^{\mathbf{x}'_i \boldsymbol{\beta}_j} \phi_i \xi_{ij}) \exp(-\xi_{ij}) \right\},$$

where  $\text{Poisson}(z|\mu)$  denotes the Poisson distribution with mean  $\mu$ .

In order to borrow information among  $\alpha_j$ 's, we consider the following prior for  $\alpha_j$ 's:

$$\alpha_j | P \sim P, \quad P \sim DP(M_0 P_0), \tag{6}$$

where  $P_0 = \mathcal{G}(a_\alpha, b_\alpha)$  is the DP base distribution and takes the gamma distribution with mean  $a_\alpha/b_\alpha$  and  $M_0$  is the precision parameter measuring the closeness of  $P$  to the base

distribution  $P_0$ . Since  $\phi_i$  is a positive random variable and has mean 1, the natural prior is  $\mathcal{G}(\nu^{-1}, \nu^{-1})$  with  $\nu$  characterizing the prior variance of  $\phi_i$ ,  $i = 1, \dots, n$ .

Let  $\eta_{lj} = \exp(\beta_{lj})$  for each  $l$  and  $j$ . For computational purposes, instead of sampling  $\beta_{lj}$ 's directly, we sample  $\eta_{lj}$ 's from their full conditionals and update  $\beta_{lj} = \log(\eta_{lj})$  in each MCMC step. We take the following priors for  $\eta_{lj}$ 's:

$$\begin{aligned} \eta_{lj} | (G_l, \pi_l) &\stackrel{iid}{\sim} \pi_l \delta_1 + (1 - \pi_l) G_l, \quad j = 1 \dots, J, \\ G_l &\stackrel{iid}{\sim} DP(M_l G_0), \quad \pi_l \stackrel{iid}{\sim} Beta(1, c), \quad c \sim \pi(c | c_0). \end{aligned} \quad (7)$$

Since the prior for  $\eta_{lj}$ 's in (7) is equivalent to that for  $\beta_{lj}$ 's in (3), the dependence structure between  $\beta_{lj}$ 's and the results in (4) still hold. We take  $\pi(c|c_0) = \mathcal{G}(c_0, 1)$ , a conjugate prior for  $c$ . The parameter  $c_0$  controls the prior probabilities of global and local null hypotheses in (4) for given  $k$  and  $J$ . Even though we can allow different  $G_{0l}$ 's for different  $l$ 's, we take  $G_{0l} = G_0$  for all  $l$  in the following for computational purpose. Since  $\beta_{lj}$ 's are nonnegative, we have  $\eta_{lj} \geq 1$ . We take  $G_0 = \mathcal{G}^{+1}(a_G, b_G)$ , a truncated  $\mathcal{G}(a_G, b_G)$  above 1 (having support  $(1, +\infty)$ ). We assign independent priors  $\mathcal{G}(1, 1)$  to  $M_0, M_1, \dots, M_k$ .

## B. Posterior computation

We use the blocked Gibbs sampler of Ishwaran and James (2001), which relies on a truncation of the stick-breaking representation of the DPs in (6) and (7) (Sethuraman, 1994):

$$\begin{aligned} P &= \sum_{h=1}^N q_h \delta_{\alpha_h^*}, \quad q_h = W_h \prod_{h' < h} (1 - W_{h'}) \\ W_h &\sim Beta(1, M_0) \quad \forall h = 1, \dots, N-1, \quad \text{and } W_N = 1 \end{aligned} \quad (8)$$

and

$$G_l = \sum_{h=1}^N p_{lh} \delta_{\theta_{lh}}, \quad \theta_{lh} \sim G_0, \quad p_{lh} = V_{lh} \prod_{h' < h} (1 - V_{lh'}),$$

$$V_{lh} \sim \text{Beta}(1, M_l), \quad \forall h = 1, \dots, N-1, \quad \text{and} \quad V_{lN} = 1 \quad (9)$$

for each  $l$ . Here  $N$  determines the number of random atoms. The truncated DP will approximate DP with arbitrary small error when  $N$  is sufficiently large. However, in practice a moderate value of  $N$  is adequate as there are not many clusters occupied, especially in the sparse case.

Let  $K_j$  denote the label variable of  $\alpha_j$  with  $K_j = h$  if  $\alpha_j = \alpha_h^*$ . Let  $R_{lj} = h$  if  $\eta_{lj} = \theta_{lh}$  for  $h = 1, \dots, N$  and  $R_{lj} = N+1$  if  $\eta_{lj} = 1$ . Taking random values as starting points for all the parameters, the MCMC algorithm proceeds through iterating the following steps:

1. Sample  $z_{ij}$ 's,  $\xi_{ij}$ 's,  $\phi_i$ 's, and  $\nu$ .

(a) For each  $i$  and  $j$ , assign  $z_{ij}$  to be 0 if  $y_{ij} = 0$ ; Otherwise, sample  $z_{ij}$  from  $\text{Poisson}(\alpha_j e^{\mathbf{x}'_i \boldsymbol{\beta}_j} \phi_i \xi_{ij})$  such that  $z_{ij} > 0$ .

(b) Sample  $\xi_{ij}$ 's from their full conditionals,

$$\xi_{ij} | \cdot \sim \mathcal{G}(1 + z_{ij}, 1 + e^{\mathbf{x}'_i \boldsymbol{\beta}_j} \alpha_j \phi_i).$$

(c) Sample  $\phi_i$ 's from their full conditionals,

$$\phi_i | \cdot \sim \mathcal{G}(\nu^{-1} + \sum_j z_{ij}, \nu^{-1} + \sum_j \alpha_j e^{\mathbf{x}'_i \boldsymbol{\beta}_j} \xi_{ij}).$$

(d) Sample  $\nu$  with Metropolis-Hasting algorithm

$$\nu | \cdot \propto \pi(\nu) \prod_{i=1}^n \phi_i^{\nu^{-1}-1} \exp(-\nu^{-1} \phi_i) / (\nu^{\nu^{-1}} \Gamma(\nu^{-1})).$$

2. Update  $\alpha_j$ 's and related parameters including  $\alpha_h^*$ 's,  $K_j$ 's,  $w_h$ 's, and  $M_0$ .

(a) Sample DP atoms  $\alpha_h^*$ 's in (8) from their full conditionals,

$$\alpha_h^* | \cdot \sim \mathcal{G}(a_\alpha + \sum_{i,j} z_{ij} 1_{(K_j=h)}, b_\alpha + \sum_{i,j} e^{\mathbf{x}'_i \boldsymbol{\beta}_j} \xi_{ij} \phi_i 1_{(K_j=h)})$$

(b) Sample  $K_j$ 's from their full conditionals. For  $h = 1, \dots, N$ ,

$$\Pr(K_j = h | \cdot) = \frac{(\alpha_h^*)^{\sum_i z_{ij}} \exp(-\alpha_h^* \sum_i e^{\mathbf{x}'_i \boldsymbol{\beta}_j} \phi_i \xi_{ij}) q_h}{\sum_{h'=1}^N (\alpha_{h'}^*)^{\sum_i z_{ij}} \exp(-\alpha_{h'}^* \sum_i e^{\mathbf{x}'_i \boldsymbol{\beta}_j} \phi_i \xi_{ij}) q_{h'}}.$$

(c) Let  $W_N = 1$  and for  $h = 1, \dots, N-1$ , sample  $W_h$  from its full conditional,

$$W_h | \cdot \sim \text{Beta}(1 + \sum_j 1_{(K_j=h)}, M_0 + \sum_j 1_{(K_j>h)}).$$

(d) Let  $\alpha_j = \alpha_{K_j}^*$  for  $j = 1, \dots, J$ . Update  $q_h$  with

$$q_h = W_h \prod_{h' < h} (1 - W_{h'}) \quad \text{for } h = 1, \dots, N.$$

(e) Sample DP precision parameter  $M_0$  from its full conditional,

$$M_0 | \cdot \sim \mathcal{G}(N, 1 - \sum_{h=1}^{N-1} \log(1 - W_h)).$$

3. Update  $\beta_{lj}$ 's and related parameters for each fixed  $l$  in this block.

(a) Sample  $\theta_{lj}$  from truncated  $\mathcal{G}(a', b')$  below by 1, with  $a' = a_G + \sum_{i,j} x_{il} z_{ij} 1_{(R_{lj}=h)}$

and

$$b' = b_G + \sum_{i,j} \alpha_j e^{\mathbf{x}'_{i(-l)} \boldsymbol{\beta}_{j(-l)}} \xi_{ij} \phi_i x_{il} 1_{(R_{lj}=h)},$$

where  $\mathbf{x}_{i(-l)}$  is the vector of  $\mathbf{x}_i$  deleting the  $l$ th element  $x_{il}$ , and similarly for  $\boldsymbol{\beta}_{j(-l)}$ .

(b) Sample  $R_{lj}$ 's from their full conditionals,

$$\Pr(R_{lj} = N + 1 | \cdot) = r_{lj} \pi_l \exp(-\sum_i e^{\mathbf{x}'_{i(-l)} \boldsymbol{\beta}_{j(-l)}} \alpha_j \xi_{ij} \phi_i x_{il})$$

$$\Pr(R_{lj} = h | \cdot) = r_{lj}(1 - \pi_l)p_{lh}\theta_{lh}^{\sum_i x_{il}z_{ij}} \exp\left(-\sum_i e^{\mathbf{x}'_{i(-l)}}\boldsymbol{\beta}_{j(-l)}\alpha_j\xi_{ij}\phi_i x_{il}\theta_{lh}\right)$$

for  $h = 1, \dots, N$ , where  $r_{lj}$  is a normalizing constant such that

$$\sum_{h=1}^{N+1} \Pr(R_{lj} = h | \cdot) = 1.$$

- (c) Update  $\eta_{lj}$ 's and  $\beta_{lj}$ 's with  $\eta_{lj} = \theta_{lR_{lj}}$  and  $\beta_{lj} = \log(\eta_{lj})$  for  $j = 1, \dots, J$ . Sample  $V_{lh}$  from

$$V_{lh} | \cdot \sim \text{Beta}\left(1 + \sum_j 1_{(R_{lj}=h)}, M_l + \sum_j 1_{(R_{lj}>h)}\right)$$

for  $h = 1, \dots, N - 1$  and update  $p_{lh} = V_{lh} \prod_{h'<h} (1 - V_{lh'})$  for  $h = 1, \dots, N$ .

- (d) Sample  $M_l$  from its full conditional,

$$M_{lh} | \cdot \sim \mathcal{G}\left(1 + \sum_{j=1}^J 1_{(R_{lj}=h)}, 1 - \sum_{h=1}^{N-1} \log(1 - V_{lh})\right).$$

- (e) Sample  $\pi_l$  from its full conditional,

$$\pi_l | \cdot \sim \text{Beta}\left(1 + \sum_j 1_{(\beta_{lj}=0)}, c + \sum_j 1_{(\beta_{lj}>0)}\right).$$

4. Sample  $c$  from its full conditional,

$$c | \cdot \sim \mathcal{G}\left(c_0 + k, 1 - \sum_{l=1}^k \log(1 - \pi_l)\right).$$

With the exception of  $\nu$ , which is updated by using a Metropolis-Hasting algorithm in step 1(d), all other parameters have a closed form in their full conditionals, simplifying implementation.

## References

- Armitage, P. (1955), "Tests for Linear Trends in Proportions and Frequencies," *Biometrics*, 11, 375-386.
- Bailer, A.J. and Portier, C.J. (1988), "Effects of Treatment-Induced Mortality and Tumor-Induced Mortality on Tests for Carcinogenicity in Small Samples," *Biometrics*, 44, 417-431.
- Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of Royal Statistical Society: B*, 57, 289-300.
- Berry, D.A. and Hochberg, Y. (1999), "Bayesian Perspectives on Multiple Comparisons ," *Journal of Statistical Planning and Inference*, 82, 215-227.
- Bristol, D.R. (1993), "One-sided Multiple Comparisons of response rates with a Control," *In: Hope, F.M. (Ed.), Multiple Comparison, Selection, and Applications in Biometry*, Marcel Dekker, Inc., New York, 77-96.
- Chen, J.J. (1996), "Global Tests for Analysis of Multiple Tumour Data from Animal Carcinogenicity Experiments," *Statistics in Medicine*, 15, 1217-1225.
- Chen, J. and Sarkar, S.K. (2004), "Multiple testing of Response Rates with a Control: a Bayesian Stepwise Approach, " *Journal of Statistical Planning and Inference*, 125, 3-16.
- Cochran, W.G. (1954), "Some Methods for Strengthening the Common  $\chi^2$  tests," *Biometrics*, 10, 417-451.

- Chuang-Stein, C. and Tong, D.M. (1995), "Multiple Comparisons Procedures for Comparing Several Treatments with a Control Based on Binary Data," *Statistics in Medicine*, 14, 2509-2522.
- Coull, B.A., Hobert, J.P., Ryan, L.M. and Holmes, L.B. (2001), "Crossed Random Effect Models for Multiple Outcomes in a Study of Teratogenesis," *Journal of the American Statistical Association*, 96, 1194-1204.
- Dahl, D.B. and Newton, M.A. (2007), "Multiple Hypothesis Testing by Clustering Treatment Effects," *Journal of the American Statistical Association*, 102, 517-526.
- Dudoit, S., Shaffer, J.P. and Boldrick, J.C. (2003), "Multiple Hypothesis Testing in Microarray Experiments", *Statistical Science*, 1, 71-103.
- Dunson, D.B. and Dinse, G.E. (2002), "Bayesian Models for Multivariate Current Status Data with Informative Censoring," *Biometrics*, 58, 79-88.
- Dunson, D. B. and Stanford, J.B. (2005), "Bayesian Inference on Predictors of Conceptions Probabilities," *Biometrics*, 61, 126-133.
- Ferguson, T.S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics*, 1, 209-230.
- Ferguson, T.S. (1974), "Prior Distributions on Spaces of Probability Measures," *Annals of Statistics*, 2, 615-629.
- Freeman, L. (1996), "Bayesian Statistical Methods - A Natural Way to Access Clinical Evidence," *British Medicine Journal*, 313, 569-570.

- Genovese, C. and Wasserman, L. (2001), "Operating Characteristics and extensions of the FDR procedure," *Technical report 737*, Department of Statistics, Carnegie Mellon University.
- Gopalan, R. and Berry, D. A. (1998), "Bayesian Multiple Comparisons Using Dirichlet Process Priors" *Journal of the American Statistical Association*, 1130-1139.
- Gutman, R. and Hochberg, Y. (2007), "Improved Multiple Test Procedures for Discrete Distributions: New Ideas and Analytical Review," *Journal of Statistical Planning and Inference*. 2380-2393
- Hans, C. and Dunson, D.B. (2005), "Bayesian Inferences on Umbrella Orderings," *Biometrics*, 61, 1018-1026.
- Hochberg, Y. (1988), "A Sharper Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 75, 800-802.
- Holm, S. (1979), "A Simple Sequentially Rejective multiple test Procedure," *Scandinavia Journal of Statistics*, 6, 65-70.
- Hommel, G. (1988), "A Stagewise Rejective multiple test Procedure Based on a Modified Bonferroni Test," *Biometrika*, 75, 383-386.
- Ishwaran, H. and James, L.F. (2001), "Gibbs Sampling Methods for Stick-breaking Priors," *Journal of the American Statistical Association*, 101, 179-194.
- Knoke, J.D. (1976), "Multiple Comparisons with Dichotomous Data," *Journal of the American Statistical Association*, 71, 849-853.

- Lu, Y. and Malani, H.M. (1995), "Analysis of Animal Carcinogenicity Experiments with Multiple Tumor Types," *Biometrics*, 51, 73-86.
- MacLehose, R.F., Dunson, D.B., Herring, A.H. and Hoppin, J.A. (2007), "Bayesian Methods for Highly Correlated Exposure Data," *Epidemiology*, 18, 199-207.
- Meng, C.Y.K. and Dempster, A.P. (1987), "A Bayesian-Approach to the Multiplicity Problem for Significance Testing with Binomial Data," *Biometrics*, 43, 301-311.
- Miranda-Moreno, L.F., Labbe, A. and Fu, L. (2007), "Bayesian Multiple Testing Procedures for Hotpot Identification," *Accident Analysis and prevention* 39, 1192-1201.
- Müller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004) "Optimal Sample Size for Multiple Testing: The Case of Gene Expression Microarrays," *Journal of the American Statistical Association*, 99, 990-1001.
- Scott, J.G. and Berger, J.O. (2006). An Exploration of Aspects of Bayesian Multiple Testing. *Journal of Statistical Planning and Inference*, 136, 2144-2162.
- Sethuraman, J. (1994) "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639-650.
- Shaffer, J.P. (1986), "Modified Sequentially Rejective Multiple Test Procedures," *Journal of the American Statistical Association*, 81, 826-831.
- Shaffer, J.P. (1995), "Multiple Hypothesis Testing: A Review," *Annual Review of Psychology*, 46, 561-584.

- Storey, J.D. (2002), "A Direct Approach to False Discovery Rates," *Journal of Royal Statistical Society: B*, 64, 479-498.
- Storey J.D. (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the Q-value." *Annals of Statistics*, 31, 2013-2035.
- Storey, J.D., Taylor, J.E. and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation, and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of Royal Statistical Society: B*, 66, 187-205.
- Westfall, P.H. and Wolfinger, R.D., (1997), "Multiple tests with discrete distributions", *American Statistician* 51, 38.
- Westfall, P.H. and Young, S.S. (1993), "Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment," Wiley, New York.

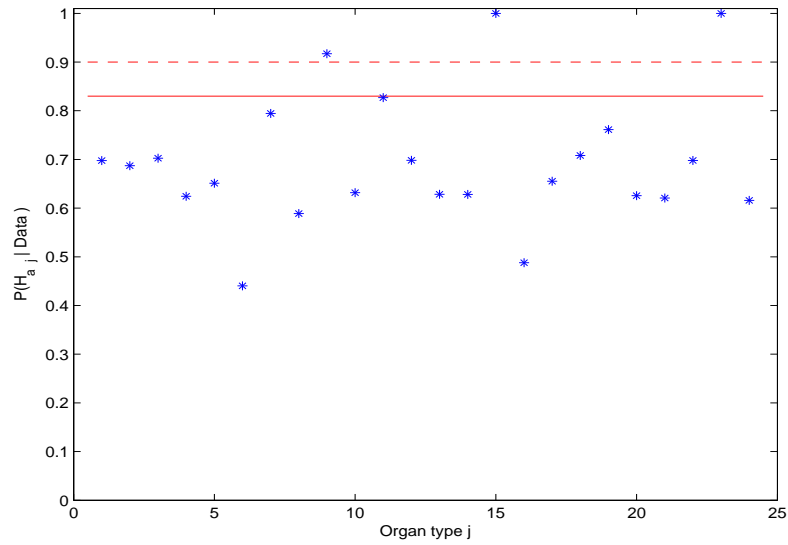


Figure 1: Posterior probabilities of local alternative hypotheses. The broken and solid lines correspond to 0.90 and 0.83 respectively.

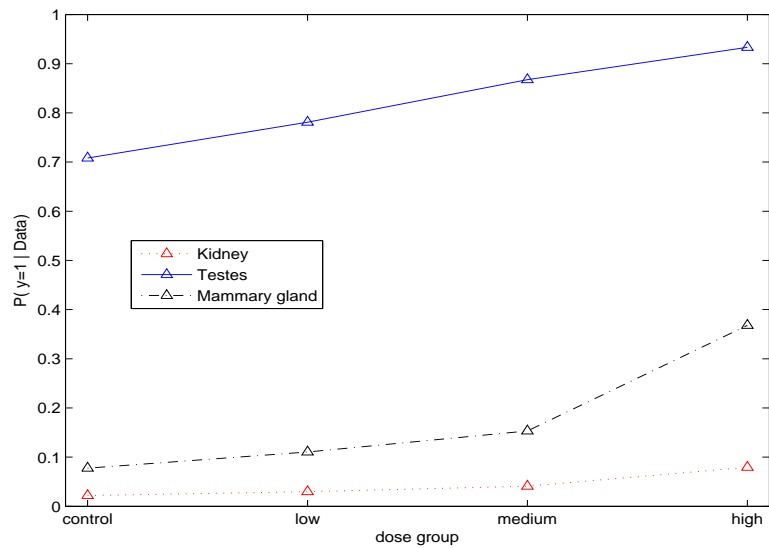


Figure 2: Dose-response curve for tumors at kidney, testes, and mammary gland.

Table 1: For each of the 24 combinations of organs and tumor types in the NTP study of isoprene, we show the number of male rats with a tumor at that organ out of 50 in each of the four groups in columns 3 – 6, along with the evidence of dose effect between each dose group and the control group in terms of posterior probability in columns 7 – 9.

$j$	Tumor/Organ type	Dose group (ppm)				Posterior probability		
		C	L	M	H	C vs. L	C vs. M	C vs. H
1	Adrenal Cortex	1	0	0	1	0.298	0.492	0.702
2	Intestine Large	0	1	1	0	0.393	0.562	0.691
3	Intestine Small	0	0	0	1	0.306	0.499	0.708
4	Stomach, Glandular	0	1	0	0	0.314	0.467	0.631
5	Stomach, Forestomach	0	0	1	0	0.320	0.507	0.655
6	Adrenal Medulla	20	16	20	18	0.149	0.320	0.447
7	Zymbal's Gland	0	1	1	1	0.476	0.662	0.800
8	Oral Mucosa	1	0	0	0	0.250	0.421	0.596
9	Kidney	0	2	2	6	0.629	0.805	<b>0.918</b>
10	Spleen	0	1	0	0	0.314	0.470	0.638
11	Skin	4	8	6	8	0.601	0.715	0.828
12	Tongue	0	0	0	1	0.306	0.499	0.708
13	Tissue NOS (I)	1	1	0	0	0.325	0.478	0.636
14	Tissue NOS (II)	0	1	0	0	0.319	0.474	0.635
15	Testes	33	37	44	48	0.605	0.940	<b>1.000</b>
16	Preputial Gland	5	3	5	3	0.197	0.364	0.499
17	Prostate	1	0	1	0	0.309	0.505	0.659
18	Liver	0	3	0	0	0.499	0.600	0.715
19	Brain	0	0	1	1	0.384	0.560	0.764
20	Bone (I)	0	1	0	0	0.320	0.474	0.634
21	Bone (II)	0	1	0	0	0.318	0.469	0.628
22	Bone (III)	0	0	0	1	0.309	0.492	0.704
23	Mammary Gland	2	5	7	21	0.593	0.861	<b>1.000</b>
24	Pancreas	3	1	3	0	0.328	0.506	0.611

Note: C, L, M, and H denote the control, low-, medium, and high-dose groups with concentration of isoprene at 0, 220, 700, and 7000ppm respectively.

Table 2: Evidences of dose effect between each dose group and control group in terms of Bayes factors for kidney, testes, and mammary gland respectively.

$j$	Organ	C vs. L	C vs. M	C vs. H
9	Kidney	9.822	13.292	26.186
15	Testes	8.896	50.320	$\infty$
23	Mammary gland	8.434	19.864	$\infty$

Table 3: Type 1 errors and powers of the global hypothesis tests with the proposed approach (PA) and the adjusted Cochran-Armitage test in the null, sparse and general cases based on 100 repeated data sets.

	PA	Adjusted CA
Null	0.05	0.05
Sparse	0.47	0.45
General	0.92	0.66

Table 4: The average of observed FDRs and powers for multiple local hypotheses testing from 100 data sets with the two multiple testing procedures based on the proposed method: PA1 and PA2, and two adjusted Cochran-Armitage tests dealing with multiplicity: Bonferroni adjustment (CA-B) and Benjamini and Hochberg method (CA-BH).

		PA1	PA2	CA-B	CA-BH
Null	FDR	0.040	0.030	0.050	0.050
	Power	-	-	-	-
Sparse	FDR	0.079	0.053	0.020	0.030
	Power	0.180	0.150	0.170	0.197
General	FDR	0.050	0.059	0.036	0.049
	Power	0.289	0.290	0.113	0.201