

Bayesian Latent Variable Models for Mixed Discrete Outcomes

David B. Dunson^{1,*} and Amy H. Herring²

¹Biostatistics Branch,

National Institute of Environmental Health Sciences

MD A3-03, P.O. Box 12233

Research Triangle Park, NC 27709, U.S.A.

² Department of Biostatistics

The University of North Carolina

Chapel Hill, NC

* dunson1@niehs.nih.gov

SUMMARY. In studies of complex health conditions, mixtures of discrete outcomes (event time, count, binary, ordered categorical) are commonly collected. For example, studies of skin tumorigenesis record latency time prior to the first tumor, increases in the number of tumors at each week, and the occurrence of internal tumors at the time of death. Motivated by this application, we propose a general underlying Poisson variable framework for mixed discrete outcomes, accommodating dependency through an additive gamma frailty model for the Poisson means. The model has log-linear, complementary log-log, and proportional hazards forms for count, binary and discrete event time outcomes, respectively. Simple closed form expressions can be derived for the marginal expectations, variances, and correlations. Following a Bayesian approach to inference, conditionally-conjugate prior distributions are chosen that facilitate posterior computation via an MCMC algorithm. The methods are illustrated using data from a Tg.AC mouse bioassay study.

Key Words: Discrete time survival; Latent variables; Joint model; Multiple binary outcomes; Poisson counts; Proportional hazards; Random effects; Tumor multiplicity

1. Introduction

Discrete outcomes having a variety of measurement scales (event time, count, binary) are commonly collected in studies of complex health conditions. For example, animal studies of skin and breast tumorigenesis collect data on the time to first tumor (*discrete event time*), increases in tumor burden over time (*repeated counts*), and the occurrence of internal tumors at death (*multiple binary*). Motivated by this application, this article proposes a latent variable model, which links the different discrete measurements to underlying counts that are assigned an additive gamma frailty model.

In tumor studies, these underlying counts are interpretable as the number of tumors in different categories defined by time, site, and pathology. However, the methodology can be used in general applications involving mixed discrete outcomes. In the univariate case, we place a standard log-linear gamma frailty model on the underlying Poisson mean to allow for possible over-dispersion due to unmeasured genetic and environmental factors. This structure results in complementary log-log and proportional hazards random effects models for binary and discrete event time data, respectively. To allow for general dependency structures in the multivariate case, we replace the gamma frailty multiplier on the underlying Poisson mean with a linear combination of gamma latent variables.

This formulation is motivated by Tg.AC transgenic mouse bioassays (Spalding et al., 1999). Tg.AC mice have an oncogene inserted, enhancing their susceptibility to tumorigenesis. Although lines of transgenic mice are periodically genotyped to verify transgene activation, the number of activated copies varies among animals. For this reason, there is a known source of extra-Poisson variability in the number of tumors per animal. Investigators are interested in distinguishing effects of test chemicals on (1) time to detection of the first papilloma (*latency*); (2) rates of increase in papilloma burden after onset of the first tumor (*multiplicity*); and (3) lifetime risk of developing malignant tumors (*malignancy*). The extent to which these different factors are sensitive to transgene activity is unknown, but it

is important to allow differential effects and to accommodate other sources of extra-Poisson variability in the papilloma counts.

By incorporating shared latent variables in regression models for the different outcomes, our model flexibly accommodates dependency among outcomes having different measurement scales using a related approach to that used in underlying normal variable models for mixed categorical and continuous data (Muthén, 1984; Regan and Catalano, 1999; Shi and Lee, 2000; Gueorguieva and Agresti, 2001; Lee and Shi, 2001; Dunson, Chen, and Harry, 2003). However, the underlying normal variable framework does not naturally accommodate count data, and there are some difficulties in parameter interpretation due to the lack of closed form expressions relating regression coefficients to the category probabilities. In addition, tumor counts are more appropriately modeled as Poisson or over-dispersed Poisson, since tumor cells arise from a series of mutations in genes involved in regulation of cell division and death, each occurring with low frequency in the cell population.

As an alternative to underlying normal variable models, previous authors have defined multivariate distributions for mixed outcomes by incorporating shared normally distributed random effects in generalized linear mixed models (Moustaki, 1996; Sammel, Ryan, and Legler, 1997; Moustaki and Knott, 2000; Dunson, 2000, 2003). Although models of this type are very flexible, the lack of simple expressions for the marginal mean and variance makes parameter interpretation difficult. In addition, model fitting tends to be highly computationally intensive, particularly when more than a few random effects are incorporated.

Our proposed latent variable model has practical advantages in terms of simplicity of parameter interpretation and computation. The category probabilities can be characterized as simple analytic expressions of the model parameters. In addition, closed form expressions can be derived for the marginal expectations, variances, and correlations. Finally, conditionally-conjugate prior distributions can be chosen, leading to simple conditional posterior distributions and closed forms for normalizing constants used in model selection.

Alternative Poisson latent variable models have been proposed in the literature, motivated by applications to studies of malformations (Legler and Ryan, 1997) and tumorigenesis and cure (Yakovlev and Tsodikov, 1996; Dunson and Baird, 2002). These models have fundamentally different structures from the model proposed in this article, which is more closely related to Poisson-gamma models for longitudinal counts (Crouchley and Davies, 1999; Jorgensen et al., 1999; Henderson and Shimakura, 2003) and to gamma frailty models for survival data (Clayton, 1991). In addition to allowing for mixtures of counts and other discrete outcomes, the primary innovative feature of our model is the additive factor analytic structure of the frailty. Although additive gamma frailty models have been proposed for survival data (Korsgaard and Andersen, 1998; Petersen, 1998; Li, 2002), the frailty term in our model is more flexible due to the incorporation of unknown factor loadings.

Section 2 proposes the model and discusses properties. Section 3 describes a Bayesian approach to model fitting and inference. Section 4 contains results from a simulated data example. Section 5 applies the methods to Tg.AC data, and Section 6 discusses the results.

2. Poisson-Gamma Latent Variable Framework

2.1 Univariate Case

We first describe Poisson underlying variable models for univariate count, binary, ordered categorical, and discrete time to event data, motivated by applications to tumor studies. First suppose that data on subject i consist of a single tumor count $y_i = z_i$. We focus initially on the simple Poisson log linear model:

$$z_i \sim \text{Poisson}(\eta_i), \quad \eta_i = \exp(\mathbf{x}_i' \boldsymbol{\beta}), \quad (1)$$

where $\mathbf{x}_i = (x_{i1}, \dots, x_{iq})'$ is a vector of predictors, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)'$ are unknown regression coefficients. Related Poisson models have been widely used for tumor counts.

In many cases, the exact number of tumors per animal (z_i) is unknown, and data consist instead of a binary indicator variable y_i , which equals one if animal i has any tumors and

zero otherwise. Hence, we have $y_i = I_{(z_i > 0)}$ which implies under expression (1) that

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = 1 - \exp \{ - \exp(\mathbf{x}'_i \boldsymbol{\beta}) \}, \quad (2)$$

a complementary log-log model. It is straightforward to extend this model to the ordered categorical case, in which the exact count is unknown but categories are available.

Alternatively, the outcome may be the time of tumor onset. Let $t_i \in \{1, \dots, J\}$ denote the minimum of the tumor onset time and the death (censoring) time, let $\delta_{ij} = I_{(t_i \geq j)}$ be an at risk indicator, and let $y_{ij} = 1$ if tumor onset occurs at time j and $y_{ij} = 0$ otherwise. If z_{ij} denotes the number of tumors on animal i at time j , then $y_{ij} = I_{(z_{ij} > 0)}$ for $j = 1, \dots, t_i$. Assuming that $z_{ij} \sim \text{Poisson}(\eta_{ij})$, where $\eta_{ij} = \exp(\mathbf{x}'_{ij} \boldsymbol{\beta})$ and \mathbf{x}_{ij} is a vector of predictors,

$$\Pr(y_{ij} = 1 \mid \delta_{ij} = 1, \mathbf{x}_{ij}) = 1 - \exp \{ - \exp(\mathbf{x}'_{ij} \boldsymbol{\beta}) \}, \quad (3)$$

which is a discrete-time version of the proportional hazards model.

In each of these cases, the observed outcome y_{ij} is linked to an underlying Poisson variable z_{ij} , and integrating out z_{ij} results in a regression model for y_{ij} . Although this underlying Poisson structure is motivated by the tumor application, defining binary response and discrete-time hazards models in this manner has computational advantages, which will be discussed in Section 3. In addition, relating mixed discrete outcomes $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})'$ to underlying variables $\mathbf{z}_i = (z_{i1}, \dots, z_{in_i})'$ having a common scale (in this case, Poisson counts) has advantages in defining multivariate distributions for these mixed outcomes.

2.2 Additive Gamma Frailty Structure

In this Section, we propose an additive gamma frailty model for the underlying Poisson variables, defining a multivariate distribution for discrete outcomes having mixed measurement scales. First, we generalize expression (1) to allow over-dispersion by using the standard Poisson-gamma shared frailty model, $\eta_{ij} = \xi_i \exp(\mathbf{x}'_{ij} \boldsymbol{\beta})$, where ξ_i is a frailty having a gamma

$\mathcal{G}(\phi^{-1}, \phi^{-1})$ density with variance ϕ . The marginal expectation and variance of z_{ij} are

$$\mathbb{E}(z_{ij} | \mathbf{x}_{ij}) = \exp(\mathbf{x}'_{ij}\boldsymbol{\beta}) \quad \text{and} \quad \mathbb{V}(z_{ij} | \mathbf{x}_{ij}) = \exp(\mathbf{x}'_{ij}\boldsymbol{\beta}) + \phi \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})^2, \quad (4)$$

so that ϕ is a measure of over-dispersion relative to the Poisson distribution.

We generalize this model to allow for multivariate $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})'$ by letting

$$\eta_{ij} = \mathbb{E}(z_{ij} | \boldsymbol{\xi}_i, \mathbf{x}_{ij}) = (\boldsymbol{\xi}'_i \boldsymbol{\lambda}_j) \exp(\mathbf{x}'_{ij}\boldsymbol{\beta}), \quad (5)$$

where $\boldsymbol{\eta}_i = (\eta_{i1}, \dots, \eta_{ip})' = \mathbb{E}(\mathbf{z}_i | \boldsymbol{\xi}_i, \mathbf{X}_i)$, $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{ip})'$ is a $p \times q$ matrix of covariates, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)'$ are unknown regression coefficients, $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{ir})'$ is a vector of independent gamma distributed latent variables with $\xi_{ik} \sim \mathcal{G}(\phi_k^{-1}, \phi_k^{-1})$, for $k = 1, \dots, r$, and $\boldsymbol{\lambda}_j = (\lambda_{j1}, \dots, \lambda_{jr})'$ are factor loadings for the j th outcome, $j = 1, \dots, p$. By linking the underlying Poisson variables \mathbf{z}_i to the observed outcomes \mathbf{y}_i as described in Section 2.1, we can accommodate various mixtures of discrete outcomes.

The frailty multiplier, $\boldsymbol{\xi}'_i \boldsymbol{\lambda}_j$, in expression (5) is structured in a general manner to accommodate dependency in the multiple outcomes as well as over-dispersion relative to the Poisson distribution for count outcomes. The frailty is defined as a weighted sum of gamma latent variables, a structure related to additive gamma frailty models proposed previously in the literature (Korsgaard and Andersen, 1998; Petersen, 1998; Li, 2002). However, the factor analytic structure of our model, in which multipliers on the gamma latent variables are assumed to be unknown, differs from previous models which assume known weights (e.g., 0 or 1). This structure is motivated by studies in which dependency arises due to the presence of unmeasured factors (e.g., transgene activity), each of which may have a differential effect on the outcomes.

Under expression (5), the marginal expectation and variance of z_{ij} , integrating out the latent gamma variables are

$$\mathbb{E}(z_{ij} | \mathbf{x}_{ij}) = \left(\sum_{k=1}^r \lambda_{jk} \right) \exp(\mathbf{x}'_{ij}\boldsymbol{\beta}) = \exp(\nu_j + \mathbf{x}'_{ij}\boldsymbol{\beta}),$$

$$V(z_{ij} | \mathbf{x}_{ij}) = E(z_{ij} | \mathbf{x}_{ij}) + \left(\sum_{k=1}^r \phi_k \lambda_{jk}^2 \right) \left\{ \exp(\mathbf{x}'_{ij} \boldsymbol{\beta}) \right\}^2, \quad (6)$$

where $\nu_j = \log(\sum_k \lambda_{jk})$. The expectation follows by simply plugging in $E(\xi_{ik}) = 1$ for each of the latent variables in expression (5). The variance is derived in Appendix A.

2.3 Parameter Interpretation and Properties

For a count outcome, z_{ij} is observed directly and (6) provides a closed form for the marginal expectation and variance. It is clear that the subject-specific and population-averaged regression models have the same multiplicative form, with only the intercept varying. Hence, the regression coefficients, β_h , have both conditional and marginal interpretations in terms of logarithms of ratios of expectations. The distribution of z_{ij} is over-dispersed relative to the Poisson distribution when $\sum_{k=1}^r \phi_k \lambda_{jk}^2 > 0$. Hence, when y_{ij} is a count, the $\boldsymbol{\lambda}_j$ parameters measure not only the correlation between y_{ij} and the other outcomes but also over-dispersion in the marginal distribution of y_{ij} relative to the Poisson distribution. To separate these two attributes, one can include a latent variable specific to each count outcome, along with latent variables that load on more than one outcome and hence accommodate correlation. For example, if y_{ij} is a count and ξ_{ik} is included to allow over-dispersion, then we would restrict $\lambda_{j'k} = 0$ for all $j' \neq j$. This strategy and issues in model identifiability are discussed in detail in Section 3.1.

For binary outcomes, $y_{ij} = I_{(z_{ij} > 0)}$, a simple closed form expression can be derived for the marginal probability of a response (e.g., one or more tumors), integrating out the underlying Poisson and latent gamma variables:

$$\begin{aligned} \Pr(y_{ij} = 1 | \mathbf{x}_{ij}) &= \Pr(z_{ij} > 0 | \mathbf{x}_{ij}) = 1 - \Pr(s_{ij1} = 0, \dots, s_{ijr} = 0 | \mathbf{x}_{ij}) \\ &= 1 - \int \prod_{k=1}^r \Pr(s_{ijk} = 0 | \xi_{ik}, \mathbf{x}_{ij}) \mathcal{G}(\xi_{ik}; \phi_k^{-1}, \phi_k^{-1}) d\xi_{ik} \\ &= 1 - \prod_{k=1}^r \int_0^\infty \exp(-\lambda_{jk} \xi_{ik} e^{\mathbf{x}'_{ij} \boldsymbol{\beta}}) \frac{(\phi_k^{-1})^{\phi_k^{-1}}}{\Gamma(\phi_k^{-1})} \xi_{ik}^{\phi_k^{-1}-1} \exp(-\xi_{ik} \phi_k^{-1}) d\xi_{ik} \end{aligned}$$

$$= 1 - \prod_{k=1}^r \left(\frac{1}{1 + \phi_k \lambda_{jk} \exp(\mathbf{x}'_{ij} \boldsymbol{\beta})} \right)^{\phi_k^{-1}}, \quad (7)$$

which simplifies to a logistic regression model when $r = 1$ and $\phi_1 = 1$. This simplification also occurs when $r > 1$, and there is a single factor loading on the j th outcome ($\lambda_{jk} > 0$, $\lambda_{jk'} = 0$ for $k' \neq k$) having variance one ($\phi_k = 1$). Note that by characterizing discrete event time data using the underlying Poisson formulation described in Section 2.1 with the gamma frailty model extension of Section 2.2, one can use expression (7) directly for the marginal hazard of an event at time j , integrating out the frailty term.

A useful closed form expression for characterizing the dependency between multiple binary outcomes is $\Pr(y_{ij} = 0 \mid y_{ij'} = 0, \mathbf{X}_i) / \Pr(y_{ij} = 0 \mid \mathbf{X}_i)$

$$\begin{aligned} &= \frac{\int \Pr(y_{ij} = 0 \mid \boldsymbol{\xi}_i, \mathbf{X}_i) \Pr(y_{ij'} = 0 \mid \boldsymbol{\xi}_i, \mathbf{X}_i) \pi(d\boldsymbol{\xi}_i)}{\Pr(y_{ij} = 0 \mid \mathbf{X}_i) \Pr(y_{ij'} = 0 \mid \mathbf{X}_i)} \\ &= \frac{\prod_{k=1}^r \int \exp\{-\xi_{ik}(\lambda_{jk} e^{\mathbf{x}'_{ij} \boldsymbol{\beta}} + \lambda_{j'k} e^{\mathbf{x}'_{ij'} \boldsymbol{\beta}})\} \mathcal{G}(\xi_{ik}; \phi_k^{-1}, \phi_k^{-1}) d\xi_{ik}}{\prod_{k=1}^r \{(1 + \phi_k \lambda_{jk} e^{\mathbf{x}'_{ij} \boldsymbol{\beta}})(1 + \phi_k \lambda_{j'k} e^{\mathbf{x}'_{ij'} \boldsymbol{\beta}})\}^{-\phi_k^{-1}}} \\ &= \prod_{k=1}^r \left(\frac{(1 + \phi_k \lambda_{jk} e^{\mathbf{x}'_{ij} \boldsymbol{\beta}})(1 + \phi_k \lambda_{j'k} e^{\mathbf{x}'_{ij'} \boldsymbol{\beta}})}{1 + \phi_k (\lambda_{jk} e^{\mathbf{x}'_{ij} \boldsymbol{\beta}} + \lambda_{j'k} e^{\mathbf{x}'_{ij'} \boldsymbol{\beta}})} \right)^{\phi_k^{-1}}. \end{aligned} \quad (8)$$

This expression is interpretable as the multiplicative increase in the probability of $y_{ij} = 0$ given $y_{ij'} = 0$. For example, in chronic bioassay studies that record the occurrence of tumors in different organ sites, one could use this expression to estimate the multiplicative increase in the probability of being free of liver tumors given that the animal is free of kidney tumors.

3. Bayesian Inference

3.1 Prior Specification

A Bayesian specification of the model is completed with prior distributions for the factor loadings, $\boldsymbol{\lambda} = (\boldsymbol{\lambda}'_1, \dots, \boldsymbol{\lambda}'_p)$, the frailty variances, $\boldsymbol{\phi} = (\phi_1, \dots, \phi_r)'$, and the regression parameters, $\boldsymbol{\beta}$. We focus here on the prior for $\boldsymbol{\lambda}$ and $\boldsymbol{\phi}$, and allow $\pi(\boldsymbol{\beta})$ to follow an arbitrary form (e.g., multivariate normal). For simplicity in prior elicitation and computation, we assume *a priori* independence between $\boldsymbol{\beta}$ and the elements of $\boldsymbol{\lambda}$ and $\boldsymbol{\phi}$. For the frailty precision param-

eters, ϕ_k^{-1} , we choose gamma $\mathcal{G}(a_k, b_k)$ priors following common practice (Clayton, 1991). For the factor loadings parameters, λ_{jk} , we also choose gamma priors, $\pi(\lambda_{jk}) = \mathcal{G}(\lambda_{jk}; c_{jk}, d_{jk})$. However, for purposes of identifiability and interpretability, some of the elements of $\boldsymbol{\lambda}$ must be constrained.

For example, one can apply the following strategy. First, without loss of generality, assume that the first p_1 elements of \mathbf{y}_i are counts and the remaining elements are binary (or indicators of event occurrence for discrete event time data). Then, following the strategy proposed in Section 2.3, we let $\phi_1, \dots, \phi_{p_1}$ measure over-dispersion in outcomes $j = 1, \dots, p_1$, respectively. For the corresponding latent variables, $\xi_{i1}, \dots, \xi_{i,p_1}$, the factor loadings are fixed in advance by letting $\lambda_{jk} = 1$ for $j = k$ and $\lambda_{jk} = 0$ for $j \neq k$, $k = 1, \dots, p_1$, $j = 1, \dots, p$. The remaining latent variables, $\xi_{i,p_1+1}, \dots, \xi_{ir}$, load on more than one outcome and hence accommodate correlation. Letting $\boldsymbol{\Lambda}_2$ denote the $p \times (r - p_1)$ factor loadings matrix with row vectors $(\lambda_{j,p_1+1}, \dots, \lambda_{jr})'$, for $j = 1, \dots, p$, the standard identifiability conditions can be placed on $\boldsymbol{\Lambda}_2$ to ensure identifiability. In particular, the number of factors should be chosen so that $p_2 = p - p_1 > r - p_1$ (often much greater). In addition, one element in each column can be set to one and sufficient structural zeros incorporated based on the data application to avoid ambiguity. As we discuss in Section 4, an alternative is to avoid fixing factor loadings at one by instead fixing the frailty variances: $\phi_{p_1+1}, \dots, \phi_r = 1$.

3.2 Data Augmentation MCMC Algorithm

We propose a data augmentation MCMC algorithm for posterior computation, taking advantage of the underlying Poisson formulation of our model as Albert and Chib (1993) utilized the underlying normal structure of probit models. In particular, our model assumes that $y_{ij} = g_j(z_{ij})$, where y_{ij} is the observed outcome, z_{ij} is an underlying Poisson variable, and $g_j(\cdot)$ is a known link function (identity for counts, threshold for binary or categorical data). Potentially, $g_j(\cdot)$ can depend on a vector of unknown thresholds $\boldsymbol{\tau}_j$ to allow for ordered

categorical y_{ij} , but we focus on the simple case where the thresholds are known.

We express z_{ij} as the following sum of independent Poisson latent variables:

$$z_{ij} = \sum_{k=1}^r s_{ijk}, \quad s_{ijk} \sim \text{Poisson}(\lambda_{jk} \xi_{ik} \exp(\mathbf{x}'_{ij} \boldsymbol{\beta})), k = 1, \dots, r. \quad (9)$$

Following standard algebraic routes, it is straightforward to show that the resulting full conditional posterior distributions for s_{ijk} , ξ_{ik} , and λ_{jk} follow simple conjugate forms. After specifying initial values, our algorithm alternates between:

1. Imputing the underlying Poisson variables $\mathbf{s}_{ij} = (s_{ij1}, \dots, s_{ijr})'$ by sampling from their full conditional posterior distribution (for all i, j).
2. Updating the latent variables, $\boldsymbol{\xi}_i$, by sampling one at a time from their gamma full conditional posterior distributions.
3. Updating the unknown factor loadings, λ_{jk} , by sampling one at a time from their gamma full conditional posterior distributions.
4. Updating the frailty variances, $\boldsymbol{\phi}$, and regression parameters, $\boldsymbol{\beta}$.

In Step 4, Metropolis-Hastings or adaptive rejection sampling can be used. Note that gamma priors are conditionally-conjugate for the exponentiated intercept, $\exp(\beta_1)$, and for the exponentiated regression coefficients for binary predictors. Thus, for some elements of $\boldsymbol{\beta}$, one can sample directly from the conditional distributions if gamma priors are used. The full conditional distributions needed to implement this algorithm are outlined in Appendix B.

4. Simulation Example

To assess the robustness of inferences to the prior specification and to the choice of factor structure, we first analyzed simulated data in which the true values of the parameters were known. These data were chosen to have the same structure as the Tg.AC mouse bioassay study of PETA described in Section 5. In particular, outcomes for each mouse consisted of

the age at first papilloma detection, weekly increases in the papilloma burden after onset, and the occurrence of malignant tumors. Mice are at risk of papillomas starting at week 9 and are examined each week up to week 27, at which time animals are sacrificed and examined for malignant tumors.

Let T_i denote the number of weeks at which mouse i is examined starting at week 9 and going up to the censoring time. Let $\delta_i = 1$ if the mouse developed papillomas by T_i and $\delta_i = 0$ otherwise, and let $t_i \in \{1, 2, \dots, T_i\}$ denote the minimum of T_i and the week of first papilloma occurrence. The data for the three aspects of the tumor response are (i) *binary indicators* ($y_{ij} = I_{(j=t_i)}\delta_i$ for $j \leq t_i$) of papilloma onset; (ii) *repeated counts* ($y_{i,t_i+1}, \dots, y_{i,T_i+1}$) of weekly increases in papilloma burden from onset of the first tumor; and (iii) a *binary indicator* (y_{i,T_i+2}) of the presence of malignant tumors.

We considered various models for these data, each of which incorporated two latent variables, ξ_{i1} and ξ_{i2} , with ξ_{i1} measuring activity of the transgene and other factors predicting overall tumorigenic response for mouse i and ξ_{i2} measuring factors specific to the count outcomes leading to overdispersion. In particular, choosing models within the class described in Section 2, we let

$$\Pr(y_{ij} = 1 \mid x_i, \boldsymbol{\xi}_i) = 1 - \exp \{ - (\boldsymbol{\xi}'_i \boldsymbol{\lambda}_1) \exp(\beta_1 + \beta_2 x_i) \}, \quad j = 1, \dots, t_i, \quad (10)$$

$$E(y_{ij} \mid x_i, \boldsymbol{\xi}_i) = (\boldsymbol{\xi}'_i \boldsymbol{\lambda}_2) \exp(\beta_3 + \beta_4 x_i), \quad j = t_i + 1, \dots, T_i + 1, \quad (11)$$

$$\Pr(y_{ij} = 1 \mid x_i, \boldsymbol{\xi}_i) = 1 - \exp \{ - (\boldsymbol{\xi}'_i \boldsymbol{\lambda}_3) \exp(\beta_5 + \beta_6 x_i) \}, \quad j = T_i + 2, \quad (12)$$

where $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2})'$, and models (10) - (12) characterize the discrete hazard of papilloma onset, increases in papilloma burden after onset, and the risk of malignancy, respectively.

We considered the following possibilities for the frailty multiplier:

Model	ϕ^\dagger	Outcome Type		
		Latency ($\xi'_i \lambda_1$)	Multiplicity ($\xi'_i \lambda_2$)	Malignancy ($\xi'_i \lambda_3$)
I	(ϕ_1, ϕ_2)	ξ_{i1}	$\lambda_1 \xi_{i1} + \xi_{i2}$	$\lambda_2 \xi_{i1}$
II	$(\phi_1 = 1, \phi_2)$	$\lambda_1 \xi_{i1}$	$\lambda_2 \xi_{i1} + \xi_{i2}$	$\lambda_3 \xi_{i1}$
III	$(\phi_1 = 1, \phi_2)$	$1 + \lambda_1 \xi_{i1}$	$\lambda_2 \xi_{i1} + \xi_{i2}$	$1 + \lambda_3 \xi_{i1}$
IV	$(\phi_1 = 1, \phi_2 = 1)$	$1 + \lambda_1 \xi_{i1}$	$1 + \lambda_2 \xi_{i1} + \lambda_4 \xi_{i2}$	$1 + \lambda_3 \xi_{i1}$

\dagger - The frailty variances are either set equal to one or estimated

These choices for the frailty multiplier are equally reasonable biologically, and, in cases we have considered, lead to essentially identical inferences about covariate effects and the degree of dependency in the outcomes. In each case, different choices are made to ensure identifiability. For example, in model I, the factor loading on ξ_{i1} is set equal to one in expression (10) to ensure identifiability of the variance parameter ϕ_1 . However, in models II-IV, we instead set the variance equal to $\phi_1 = 1$ and leave the factor loadings unspecified.

Although models I-IV lead to similar inferences given sufficiently long MCMC chains, computational efficiency varies substantially. Slow mixing is a well known problem in Bayesian computation of random effects models in general, and our model is no exception. For models I and II, there tends to be very high autocorrelation in the MCMC algorithm. To illustrate a primary cause of this problem, consider a scenario with a fixed probability of malignancy and a very low correlation between other papilloma outcomes and malignancy. In order to characterize these data, the factor loading on ξ_{i1} in the malignancy component must be small. However, this will force the marginal probability of malignancy to be small unless β_5 is large, causing high correlation in these parameters. Gelfand, Sahu and Carlin (1995) addressed a related problem in normal linear mixed models by using a hierarchical centering reparameterization to improve mixing. Motivated by this idea, we add one to the frailty multiplier in Models III and IV to reduce a posteriori dependency in the frailty multipliers and regression coefficients. This leads to dramatic improvements in mixing, particularly for model IV.

The improvement in computational efficiency has a real practical impact, decreasing

numbers of iterations needed for a given level of precision in an estimated posterior quantile by an order of magnitude or even more. These gains are not specific to the model structures considered in the example, but should hold broadly for frailty multipliers incorporating an intercept of one and including unknown factor loadings instead of frailty variances. Practically, it is often difficult to distinguish between the fit obtained by estimating a factor loading for a particular frailty and fixing the variance instead of estimating the variance and fixing the loading. However, by estimating the factor loadings instead of the variances, one can take advantage of the availability of simple conditionally-conjugate forms for the posterior distributions, simplifying efficient computation. This strategy also facilitates extensions for Bayesian model selection due to the availability of closed forms for normalizing constants (Section 6 provides additional details). For these reasons, and since there is seldom information in the prior or current data that would lead one to prefer one form over another, we recommend using frailty structures having the form of model IV as the default in general applications.

Therefore, we focus on our results for model IV. We simulated data under expressions (10)-(12) for $n = 200$, $\boldsymbol{\beta} = (-2, 1, -2, 1, -1, 0)'$, and $\boldsymbol{\lambda} = (1, 2, 0.05, 0.25)'$, with x_i a 0/1 predictor generated from a Bernoulli(0.5) distribution. To complete a Bayesian specification of the model, we choose $\mathcal{G}(0.1, 0.1)$ priors for the regression parameters, and $\mathcal{G}(1, 1)$ priors for the factor loadings. Starting at the prior means, we ran our MCMC algorithm for 220,000 iterations, discarding the first 20,000 iterations as a burn-in and collecting every 10th sample to thin the chain. For each of the parameters, 95% credible intervals included the true value, and the posterior means did not differ systematically from the true values. Thus, it appears that the results are driven by the data and not by the choice of hyperparameters. Figure 1 illustrates these results, and provides evidence of adequate mixing. Section 5 applies this same approach to data from a tumorigenicity study.

5. Tumorigenicity Application

We analyze data from a Tg.AC mouse bioassay study of pentaerythritol triacrylate (PETA), a chemical used in the production of inks, coatings, glues, polyester, and fiberglass, as well as in many other industrial processes. In Tg.AC mouse bioassays conducted by the National Toxicology Program (NTP), animals are randomized to a control or one of several dose groups. The number of skin papillomas on the back of each mouse is counted weekly for 26 weeks, at which time the animals are sacrificed and examined for internal tumors. In the PETA study, there were 30 animals in each of six treatment groups, including the control group (0 mg/kg) and five dose groups (0.75, 1.5, 3, 6, or 12 mg/kg).

For comparison, we first fitted separate generalized linear models to each of the outcome types using models (10) - (12), with the frailty multipliers excluded and with x_i equal to the $\log(\text{dose}+1)$ standardized to the unit interval. Log transforming dose resulted in clear improvements in model fit. The maximum likelihood estimates, standard errors, and p-values from one-sided Wald tests are as follows:

Component	Parameter	MLE	Std. Error	p-value
Latency	Intercept (β_1)	-6.15	0.35	—
	Slope (β_2)	4.98	0.43	< 0.01
Multiplicity	Intercept (β_3)	-1.40	0.16	—
	Slope (β_4)	2.27	0.19	< 0.01
Malignancy	Intercept (β_5)	-3.57	0.63	—
	Slope (β_6)	1.79	0.85	0.02

Hence, considering the outcomes separately, there is a significant dose response trend at the 0.05 level in each case. A clear limitation of this analysis is that it does not account for or provide information on within-animal dependency in the different aspects of the tumor response. By accounting for dependency, we can potentially gain efficiency, while limiting the possibility of an inflated type I error rate. In addition, the dependency structure is of interest biologically.

We repeated the analysis incorporating the frailty multipliers to account for dependency in the different outcomes and using the proposed Bayesian approach to inference. To assess

robustness, we considered each of the forms I-IV for the frailty multipliers. As in the simulation study, we obtained identical conclusions and essentially identical model fit in each case. Therefore, for the reasons discussed in Section 4, we focus on model IV, which is preferred due to its favorable computational properties.

The model incorporates separate intercept and slope parameters for the three different types of tumorigenic responses. Within animal dependency in these responses is accommodated through the shared latent variable, ξ_{i1} , which is allowed to affect the outcomes differentially by varying the factor loadings. The marginal probability that a mouse exposed to dose x develops a papilloma by week t is $1 - \exp\{-\exp(\beta_1 + \beta_2 x)t\} / [1 + \lambda_1 \exp(\beta_1 + \beta_2 x)t]$, a useful expression for estimating marginal survival curves. Following papilloma onset, $(1 + \lambda_1) \exp(\beta_3)$ is the expected rate of increase in the papilloma burden, and $\exp(\beta_4 x)$ is the multiplicative increase in this rate attributable to a unit increase in dose (on the standardized log scale). The marginal probability of malignancy at dose x is $1 - \exp\{-\exp(\beta_5 + \beta_6 x)\} / [1 + \lambda_3 \exp(\beta_5 + \beta_6 x)]$.

We considered a variety of prior specifications including (i) the prior described in Section 4; (ii) an informative prior for $\beta_1, \beta_3, \beta_5$ chosen to yield marginal expectations in the control group consistent with results from previous Tg.AC bioassays, and (iii) a prior equivalent to (ii) but with the variance increased five-fold. Focusing on prior (ii), marginal summaries of the low and high dose group tumor response, including the probability of developing a papilloma given survival to the end of the study, the expected weekly increase in the papilloma burden, and the probability of malignancy, are plotted in Figure 2 for each iteration of the Gibbs sampler after thinning. In addition, Figure 3 plots estimated dose response curves for each marginal summary.

Based on Figure 3 and on comparing the empirical and estimated variances in the tumor counts as a function of dose, the model fit is very good. For each type of tumor response, there was evidence of an increasing dose response trend, with $\Pr(\beta_2 > 0 | \text{data}) > 0.99$,

$\Pr(\beta_4 > 0 \mid \text{data}) > 0.99$, and $\Pr(\beta_6 > 0 \mid \text{data}) = 0.91$. Posterior summaries are presented in Table 2, along with ranges of values across the different prior specifications. The probability of developing a papilloma during the study, the rate of increase in the burden, and the probability of malignancy varied only slightly between mice in the 10th percentile of the distribution of ξ_{i1} and those in the 90th percentile. Hence, the within-animal dependency in these outcomes is low. This is an interesting and unexpected finding, since we would have anticipated a moderate correlation in the different tumor outcomes. The low correlation may reflect the different processes involved in tumor initiation, promotion, and progression to malignancy.

As is apparent from Table 2, the conclusions are robust to the prior specification and to the model structure, and we obtained essentially identical results for each prior and factor model structure we considered. The results for the tumor multiplicity component of the model were the most sensitive, which is as expected given the small number of malignant tumors observed in the study.

6. Discussion

The methodology developed in this article has been motivated by application to studies that collect information on the onset and proliferation of skin or breast tumors, which can be detected in live animals, in addition to the occurrence of occult tumors of different types. Current methodology considers data from clinical observations on palpable tumors separately from necropsy data collected on occurrences of malignancies and occult tumors. Although animal experiments commonly rely on inbred mice with minimal genetic differences, there may still be within-animal dependency in the different tumor outcomes, and it is important to consider these outcomes jointly in the analysis.

Motivated by this problem, we have proposed an underlying Poisson model for mixed discrete outcome data, with an additive gamma frailty term included to flexibly accommo-

date dependency. The structure of the model leads to simple expressions for the marginal expectations, variances, and correlations. In addition, efficient posterior computation is simplified by the availability of conditionally-conjugate prior distributions. Unlike frequentist methods based on marginally-specified random effects models (e.g., Zeger, 1998; Heagerty, 1999), the proposed approach does rely on parametric distributions for the latent variables and informative prior distributions for the parameters. However, given that asymptotic justifications are not appropriate in the tumorigenicity application and that prior information is available, our Bayesian approach has many advantages, which also hold in other application areas. In addition, marginal models for mixed discrete outcomes remain to be developed. Potentially, one could limit sensitivity to parametric assumptions within the Bayesian approach by using a Dirichlet process mixture of gamma distributions for the latent factors, as in Dunson (2004).

Although we have focused on Bayesian methods, the underlying Poisson framework is also useful in frequentist analyses. For example, the data augmentation strategy can be used to develop a Monte Carlo EM algorithm (Wei and Tanner, 1990) for maximum likelihood estimation. One advantage of the Bayesian approach is that it provides a more natural framework for accounting for model uncertainty, a challenging problem in latent variable analyses (Lopes and West, 2004). For example, our methods can be generalized for covariance selection and graphical modeling of mixed discrete outcomes. One can assess whether any two outcomes are conditionally independent by testing whether the loadings on common factors are 0. To assign a prior to the set of possible conditional independence relationships (i.e., graphs), mixture priors can be chosen for the λ 's that incorporate point masses at 0. The resulting full conditional posterior distributions for the λ 's will maintain a conjugate form, and covariance selection can proceed by a stochastic search Gibbs sampling algorithm (George and McCulloch, 1997). Such a method would automatically account for uncertainty in the factor structure and allow inferences on the covariance.

ACKNOWLEDGEMENTS

The authors thank Wendell Jones for providing the data and the Associate Editor and anonymous referee for useful suggestions.

References

- Albert, J.H. and Chib, S. (1993). Bayesian-analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669-679.
- Clayton, D.G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**, 467-485.
- Crouchley, R. and Davies, R.B. (1999). A comparison of population average and random effects models for the analysis of longitudinal count data with baseline information. *Journal of the Royal Statistical Society A* **162**, 331-347.
- Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society B* **62**, 355-366.
- Dunson, D.B. (2003). Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association* **98**, 555-563.
- Dunson, D.B. and Baird, D.D. (2002). A proportional hazards model for incidence and induced remission of disease. *Biometrics* **58**, 71-78.
- Dunson, D.B., Chen, Z., and Harry, J. (2003). A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* **59**, 521-530.
- Dunson, D.B. (2004). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association*, in press.

- Gelfand, A.E., Sahu, S.K. and Carlin, B.P. (1995). Efficient parameterizations for normal linear mixed models. *Biometrika* **82**, 479-488.
- George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339-373.
- Gueorguieva, R.V. and Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association* **96**, 1102-1112.
- Heagerty, P.J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* **55**, 688-698.
- Jorgensen, B., Lundbye-Christiansen, S., Song, P.X.K., and Sun, L. (1999). A state space model for multivariate longitudinal count data. *Biometrika* **86**, 169-191.
- Korsgaard, I.R. and Andersen, A.H. (1998). The additive genetic gamma frailty model. *Scandinavian Journal of Statistics* **25**, 255-269.
- Henderson, R. and Shimakura, S. (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika* **90**, 355-366.
- Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.
- Lee, S.Y. and Shi, J.Q. (2001) Maximum likelihood estimation of two-level latent variable models with mixed continuous and polytomous data. *Biometrics* **57**, 787-794.
- Legler, J.M. and Ryan, L.M. (1997). Latent variable models for teratogenesis using multiple binary outcomes. *Journal of the American Statistical Association* **92**, 13-20.

- Li, H.Z. (2002). An additive genetic gamma frailty model for linkage analysis of diseases with variable age of onset using nuclear families. *Lifetime Data Analysis* **8**, 315-334.
- Lopes, H.F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41-67.
- Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology* **49**, 313-334.
- Moustaki, I. and Knott, M. (2000). Generalized latent trait models. *Psychometrika* **65**, 391-411.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika* **49**, 115-132.
- Petersen, J.H. (1998). An additive frailty model for correlated life times. *Biometrics* **54**, 646-661.
- Regan, M.M. and Catalano, P.J. (1999) Likelihood models for clustered binary and continuous outcomes: Applications to developmental toxicology. *Biometrics* **55**, 760-768.
- Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society B* **59**, 667-678.
- Shi, J.Q. and Lee, S.Y. (2000) Latent variable models with mixed continuous and polytomous data. *Journal of the Royal Statistical Society B* **62**, 77-87.
- Spalding, J.W., French, J.E., Tice, R.R., Furedi-Machacek, M., Haseman, J.K., and Tennant, R.W. (1999). Development of a transgenic mouse model for carcinogenesis bioassays: evaluation of chemically induced skin tumors in Tg.AC mice. *Toxicological Sciences* **49**, 241-254.

Wei, G. and Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85**, 699-704.

Yakovlev, A.Y. and Tsodikov, A.D. (1996), *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. Singapore: World Scientific.

Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621-629.

APPENDIX A

Derivation of Marginal Variance and Correlation Coefficient

First note that the random variable, z_{ij} , can be expressed as a sum of independent Poisson random variables: $z_{ij} = \sum_{k=1}^r s_{ijk}$, where $s_{ijk} \sim \text{Poisson}(\lambda_{jk}\xi_{ik} \exp(\mathbf{x}'_{ij}\boldsymbol{\beta}))$, for $k = 1, \dots, r$. Therefore, $V(z_{ij} | \mathbf{x}_{ij}) = \sum_{k=1}^r V(s_{ijk} | \mathbf{x}_{ij})$, where the marginal variance of s_{ijk} is

$$\begin{aligned}
 V(s_{ijk} | \mathbf{x}_{ij}) &= E(s_{ijk}^2 | \mathbf{x}_{ij}) - E(s_{ijk} | \mathbf{x}_{ij})^2 = E\{E(s_{ijk}^2 | \xi_{ik}, \mathbf{x}_{ij})\} - E\{E(s_{ijk} | \xi_{ik}, \mathbf{x}_{ij})\}^2 \\
 &= E\left[\lambda_{jk}\xi_{ik} \exp(\mathbf{x}'_{ij}\boldsymbol{\beta}) + \{\lambda_{jk}\xi_{ik} \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})\}^2\right] - E\{\lambda_{jk}\xi_{ik} \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})\}^2 \\
 &= \lambda_{jk} \exp(\mathbf{x}'_{ij}\boldsymbol{\beta}) + \lambda_{jk}^2 \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})(1 + \phi_k) - \lambda_{jk}^2 \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})^2 \\
 &= \lambda_{jk} \exp(\mathbf{x}'_{ij}\boldsymbol{\beta}) + \phi_k \lambda_{jk}^2 \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})^2.
 \end{aligned}$$

Expression (6) follows directly after substituting back into the expression for $V(z_{ij} | \mathbf{x}_{ij})$.

APPENDIX B

Derivation of Conditional Posterior Distributions

The joint posterior distribution of the parameters $\boldsymbol{\theta} = (\boldsymbol{\phi}', \boldsymbol{\lambda}', \boldsymbol{\beta}')$, the underlying variables $\{z_{ij}, \mathbf{s}_{ij}\}$, and the latent variables $\{\boldsymbol{\xi}_i\}$ is proportional to

$$\left[\prod_{i=1}^n \prod_{j:\delta_{ij}=1}^p I_{(y_{ij}=g_j(z_{ij}))} I_{(z_{ij}=\sum_{j=1}^r s_{ijk})} \prod_{k=1}^r \{\lambda_{jk} \xi_{ik} \exp(\mathbf{x}'_{ij} \boldsymbol{\beta})\}^{s_{ijk}} \exp\{-\lambda_{jk} \xi_{ik} \exp(\mathbf{x}'_{ij} \boldsymbol{\beta})\} \right. \\ \left. \times \frac{(\phi_k^{-1})^{\phi_k^{-1}}}{\Gamma(\phi_k^{-1})} (\xi_{ik})^{\phi_k^{-1}} \exp(-\xi_{ik} \phi_k^{-1}) \right] \pi(\boldsymbol{\theta}). \quad (13)$$

From this expression, the full conditional posterior distributions can be derived following standard algebraic routes. Firstly, the conditional posterior distribution of z_{ij} given y_{ij} but integrating out \mathbf{s}_{ij} is

$$[z_{ij} \mid \boldsymbol{\xi}, \boldsymbol{\theta}, \mathbf{y}, \mathbf{X}] = \text{Poisson}(z_{ij}; \eta_{ij}) \quad \text{s.t. } y_{ij} = g_j(z_{ij}),$$

where $\boldsymbol{\xi} = (\boldsymbol{\xi}'_1, \dots, \boldsymbol{\xi}'_n)'$, $\mathbf{y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_n)'$, $\mathbf{X} = (\mathbf{X}'_1, \dots, \mathbf{X}'_n)'$, and $\mathbf{z} = (\mathbf{z}'_1, \dots, \mathbf{z}'_n)'$. The full conditional posterior distribution of \mathbf{s}_{ij} is a point mass at $s_{ij1} = \dots = s_{ijr} = 0$ when $z_{ij} = 0$, and is otherwise

$$[s_{ij1}, \dots, s_{ijr} \mid \mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\theta}, \mathbf{y}, \mathbf{X}] = \text{Multinomial}\left(\mathbf{s}_{ij}; z_{ij}, \frac{\lambda_{jk} \xi_{ik} \exp(\mathbf{x}'_{ij} \boldsymbol{\beta})}{\eta_{ij}}, k = 1, \dots, r\right),$$

The full conditional posterior distribution of ξ_{ik} is

$$[\xi_{ik} \mid \mathbf{z}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{y}, \mathbf{X}] = \mathcal{G}\left(\xi_{ik}; \phi_k^{-1} + \sum_{j:\delta_{ij}=1}^p s_{ijk}, \phi_k^{-1} + \sum_{j:\delta_{ij}=1}^p \lambda_{jk} \exp(\mathbf{x}'_{ij} \boldsymbol{\beta})\right).$$

The full conditional posterior distribution of λ_{jk} (assuming λ_{jk} is not fixed in advance) is

$$[\lambda_{jk} \mid \mathbf{z}, \mathbf{s}, \boldsymbol{\xi}, \boldsymbol{\theta}_{(-\lambda_{jk})}, \mathbf{y}, \mathbf{X}] = \mathcal{G}\left(\lambda_{jk}; a_{jk} + \sum_{i:\delta_{ij}=1}^n s_{ijk}, b_{jk} + \sum_{i:\delta_{ij}=1}^n \xi_{ik} \exp(\mathbf{x}'_{ij} \boldsymbol{\beta})\right).$$

Finally, the posterior distribution of $\gamma_h = \exp(\beta_h)$ is also gamma when γ_h is assigned a gamma prior and x_{ih} is either 0 or 1 for all i .

Table 1*Tumor data for the first male and female mouse in each dose group of the Tg.AC mouse**bioassay study of PETA.*

Dose (mg/kg)	Weekly papilloma counts from study week 9 [‡]																			MT^{\dagger}	
	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	x	x	x	x	x	x	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	2	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	2	2	2	2	6	6	7	9	17	20	20	20	20	20	20	20	x	x	x	x	0
6	0	0	0	1	0	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	0
12	4	4	4	6	7	7	11	10	18	20	20	20	20	20	20	20	20	20	20	20	1
12	0	5	4	5	6	10	13	13	17	20	20	20	20	20	20	20	20	20	20	20	0

[†] $MT = 1$ if a malignant tumor is detected at necropsy and $MT = 0$ otherwise[‡] Number of papillomas is censored at 20

x: animal died

Table 2*Posterior summaries under different prior specification.*

Response	Unknown	Dose Group	
		Control	High Dose
Latency	Marginal Pr(Papillomas)	0.08 [‡] _[0.06–0.09]	0.99 _[0.99–0.99]
	Pr(Papillomas Low Activity [†])	0.06 _[0.05–0.06]	0.97 _[0.97–0.98]
	Pr(Papillomas High Activity)	0.11 _[0.07–0.11]	> 0.99 _[0.99–1.00]
Multiplicity	Marginal Rate of Increase	0.22 _[0.20–0.22]	2.89 _[2.89–3.08]
	Rate Low Activity	0.18 _[0.17–0.18]	2.37 _[2.32–2.60]
	Rate High Activity	0.27 _[0.24–0.27]	3.65 _[3.65–3.77]
Malignancy	Marginal Pr(Malignancy)	0.05 _[0.05–0.05]	0.12 _[0.12–0.13]
	Pr(Malignancy Low Activity)	0.03 _[0.03–0.05]	0.09 _[0.09–0.10]
	Pr(Malignancy High Activity)	0.07 _[0.06–0.08]	0.18 _[0.17–0.18]

[†] ξ_{i1} is in 10th percentile for low activity and 90th percentile for high transgene activity

[‡] subscript is range in values from main analysis and sensitivity analyses

Figure Captions:

Figure 1. Values of the regression coefficients $(\beta_2, \beta_4, \beta_6)$ and factor loadings $(\lambda_1, \lambda_2, \lambda_3)$ at each iteration of the Gibbs sampling algorithm for the simulated data example. Posterior means (dashed lines), 95% credible intervals (dotted lines), and true values (solid lines) are also shown.

Figure 2. Values of the marginal probability of developing papillomas, the expected rate of increase in the papilloma burden after onset, and the marginal probability of developing malignant tumors for mice in the low and high dose group at each iteration of the Gibbs sampling algorithm. Posterior means (solid lines) and 95% credible intervals (dotted lines) are also shown.

Figure 3. Estimated marginal dose response curves for the three different aspects of tumor response. Points are empirical averages. Dose units are mg/kg PETA.





