

## Statistical Analysis of Skin Tumor Data from Tg.AC Mouse Bioassays

David B. Dunson,<sup>\*1</sup> Joseph K. Haseman,<sup>\*</sup> Angélique P. J. M. van Birgelen,<sup>‡</sup>  
Stanley Stasiewicz,<sup>†</sup> and Raymond W. Tennant<sup>†</sup>

<sup>\*</sup>Biostatistics Branch, <sup>†</sup>Laboratory of Environmental Carcinogenesis and Mutagenesis, and <sup>‡</sup>Toxicology Operations Branch,  
National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709

Received September 29, 1999; accepted January 11, 2000

New strategies for identifying chemical carcinogens and assessing risk have been proposed based on the Tg.AC (zetaglobin promoted v-Ha-ras) transgenic mouse. Preliminary studies suggest that the Tg.AC mouse bioassay may be an effective means of quickly evaluating the carcinogenic potential of a test agent. The skin of the Tg.AC mouse is genetically initiated, and the induction of epidermal papillomas in response to dermal or oral exposure to a chemical agent acts as a reporter phenotype of the activity of the test chemical. In Tg.AC mouse bioassays, the test agent is typically applied topically for up to 26 weeks, and the number of papillomas in the treated area is counted weekly. Statistical analyses are complicated by within-animal and serial dependency in the papilloma counts, survival differences between animals, and missing data. In this paper, we describe a statistical model for the analysis of skin tumor data from a Tg.AC mouse bioassay. The model separates effects on papilloma latency and multiplicity and accommodates important features of the data, including variability in expression of the transgene and dependency in the tumor counts. Methods are described for carcinogenicity testing and risk assessment. We illustrate our approach using data from a study of the effect of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) exposure on tumorigenesis.

**Key Words:** carcinogenicity; dose-response; drug safety; historical controls; papilloma; risk assessment; skin-painting study; transgenic mouse; tumorigenesis.

Genetically altered mice are widely used in studying mechanisms of carcinogenesis, and in recent years transgenic mouse models have been developed that can potentially discriminate between carcinogens and noncarcinogens (Eastin, 1998; Spalding *et al.*, 1999; Tennant *et al.*, 1998; Yamamoto *et al.*, 1998). The Center for Drug Evaluation and Research of the FDA has recently approved the use of transgenic animal models in screening for drug-induced carcinogenicity, and transgenic animals have become widely used in drug evaluation. There is a great deal of interest in developing methods for rapid carcinogen identification, since thousands of new drugs and chemi-

cals are developed each year that potentially modify the risk of cancer in humans. Conventional 2-year rodent bioassays are expensive and time consuming to conduct. Test systems that utilize animals susceptible to carcinogens do not require as many animals and can evaluate chemicals within shorter periods. Also, genetically altered mice that incorporate human protooncogenes may be better animal surrogates for human cancer than the wild-type rodents used in conventional studies (Contrera and DeGeorge, 1998). Transgenic mice have been developed that employ *ras* oncogenes that are known to function in both human and animal cancers (Leder *et al.*, 1990; Yamamoto *et al.*, 1996).

Preliminary studies suggest that the Tg.AC mouse, which carries an activated v-Ha-ras oncogene, may be a good model for rapid carcinogen identification (Spalding *et al.*, 1999; Tennant *et al.*, 1995, 1996). The Tg.AC mouse has genetically initiated skin, and the epidermal cells serve as targets for tumorigenesis. Although the incidence of spontaneous papillomas is very low, both genotoxic and non-genotoxic carcinogens can cause prompt epithelial proliferation and papilloma formation (Spalding *et al.*, 1993). Thus, unlike the conventional bioassay, in which the majority of tumors are occult and are not detectable until necropsy, the primary observation in Tg.AC mouse bioassays consists of weekly counts of the number of detectable skin papillomas.

The current standard for statistical analysis of skin papilloma data from Tg.AC mouse bioassays separately tests for differences between each experimental group and the control group with respect to: (1) percent animals with tumors; (2) average latency time to appearance of the first skin tumor; (3) average number of tumors per animal at risk; (4) average number of tumors per tumor-bearing animal; and (5) average latency to development of maximal number of tumors observed (Tennant *et al.*, 1998).

There are several limitations to this approach. First, the five measures are closely related and it is unlikely that a chemical has one effect and not others. Repeated testing drives the experiment-wise false positive rate above 0.05, and it is necessary to correct for multiple comparisons. A more powerful approach would compare groups with respect to fewer measures (perhaps just tumor latency and multiplicity). Second,

<sup>1</sup> To whom correspondence should be addressed at Biostatistics Branch, National Institute of Environmental Health Sciences, PO Box 12233, MD A3-03, Research Triangle Park, NC 27709. Fax: (919) 541-4311. E-mail: dunson1@niehs.nih.gov.

animals that die early are not as likely to develop papillomas or to achieve a maximum. Thus, tests based on the above measures can be extremely sensitive to group-specific differences in animal survival. Third, there is often interest not only in testing for differences between groups but also in characterizing the response at different exposure levels. Current methods for dose-response estimation in carcinogenicity bioassays are based on the proportion of animals with one or more tumors, and do not account for the actual number of tumors. Clearly, new methods are needed to better characterize exposure effects in Tg.AC bioassays.

Kokoska *et al.* (1993) proposed an approach for the statistical analysis of tumor multiplicity data from initiation/promotion experiments. Under their approach, the number of induced tumors and the individual tumor appearance times are assigned parametric distributions, and inference is based on the mean number of tumors per group and the mean time to tumor appearance. The Kokoska *et al.* model requires data on the individual tumor onset times. To obtain such data, the papillomas need to be individually monitored to determine for each study week the number of papillomas that appear for the first time in that week. Even when substantial time and effort is invested in monitoring the individual tumors, the onset times are subject to substantial measurement error, particularly when the tumor burden is moderate to large. For this reason, individual tumor data are typically not collected in Tg.AC bioassays. An additional drawback of the Kokoska *et al.* model is that it does not account for dependency between the appearance times for multiple tumors on the same animal, or for variability between animals in the propensity to develop tumors. In Tg.AC bioassays, the appearance times for multiple papillomas on the same mouse tend to be highly correlated and the papilloma response can vary substantially between mice, possibly due to heterogeneity in expression of the transgene.

In previous work, we developed flexible statistical models for skin papilloma data (Dunson, 2000; Dunson and Haseman, 1999). These models characterize the change in the papilloma burden at each observation time using underlying variables that relate to different features of the tumor response, including latency, susceptibility, multiplicity, and regression. Such an approach is extremely useful in characterizing differences in mechanistic studies and, unlike the Kokoska *et al.* approach, the models account for both dependency in the tumor appearance times and variability between animals. However, due to the complexity of the models, special software is needed to implement the analysis and it can be difficult to reliably estimate all the parameters when the papilloma incidence is low.

Statistical methods have also been proposed based on a two stage clonal expansion model of carcinogenesis, in which initiated cells multiply and regress via a stochastic birth and death process (Dewanji *et al.*, 1999). Such models are appealing, but have had limited application in testing for carcinogenic effects, due to the complexity of the likelihood. To simplify estimation and to accommodate variation in the response

among individual animals, generalized estimating equations can potentially be used for model fitting (see, for example, Burnett *et al.*, 1995). However, this approach relies heavily on large sample approximations that may not be appropriate in Tg.AC studies, which typically have a low spontaneous tumor incidence and a small to moderate sample size.

In this paper, we describe an alternative approach for the statistical analysis of skin papilloma data from a Tg.AC bioassay. We characterize the effect of exposure on the papilloma response using a mixed-effects Poisson transition model. Our model is a type of generalized linear mixed model (GLMM), and the reader is referred to Zeger and Karim (1991) and Breslow and Clayton (1993) for technical details related to GLMMs. In recent years, GLMMs have become widely used for analyzing correlated and overdispersed data (see, for example, Fung *et al.*, 1998; Piepho, 1999). Under our model, the increase in the papilloma burden from one week to the next has a Poisson sampling distribution. During a latency period prior to the appearance of any papillomas, the Poisson mean is assumed to depend on a mouse-specific susceptibility variable, on duration of exposure, and on dose through a log-linear model. After appearance of the first skin tumor, there is a shift in the Poisson mean, and the subsequent rate of increase in the papilloma burden is assumed to depend on dose through a second log-linear model. The proposed statistical model can be used for testing of exposure effects on papilloma incidence, latency and multiplicity, or for dose-response estimation. Analyses can be implemented easily within standard statistical packages, such as SAS. We illustrate the methods through application to a National Toxicology Program (NTP) study of TCDD (van Birgelen *et al.*, 1999).

## MATERIALS AND METHODS

**Modeling skin tumor counts.** In a Tg.AC mouse bioassay, each animal is randomly assigned to a dose group and is exposed throughout the 26 week duration of the study. Skin papillomas on the back of each animal are counted once per week for 26 weeks or until the animal dies. Natural deaths tend to be rare due to the short duration of the study. However, there may be treatment-induced mortality in the higher dose groups for some test chemicals. Animals that appear to be suffering, either due to toxicity or to a high tumor burden, are sometimes sacrificed for humane reasons prior to completing the study.

Let  $Z_{ij}$  be the number of detectable papillomas on the back of mouse  $i$  at week  $j$ . On a given animal, the change in the tumor burden from one week to the next equals the number of new papillomas that appear minus the number of old papillomas that regress. Thus, if papillomas are not individually tracked, we cannot determine with certainty the number of new skin tumors that appear in a given week. Data typically consist of weekly counts of the number of detectable tumors for each mouse, since tracking of individual tumors can be difficult when the papilloma burden is high. Therefore, we assume that the individual tumor onset times are unknown, and we model the rate of increase in the papilloma burden.

Let  $M_{ij} = \max\{Z_{i1}, \dots, Z_{i,j-1}\}$  be the maximum papilloma burden observed for mouse  $i$  prior to week  $j$ , and let  $Y_{ij} = M_{i,j+1} - M_{ij}$  be the increase in the maximum papilloma burden for mouse  $i$  between week  $j - 1$  and week  $j$ . We assume that the random variable  $Y_{ij}$  has a Poisson sampling distribution with the following mean:

$$\begin{aligned} \mu_{ij} &= E(Y_{ij}|M_{ij}, b_i, t_j, d_i) \\ &= \begin{cases} \exp\{\beta_1 + (b_i + \gamma_1)t_j d_i\} & \text{if } M_{ij} = 0, \\ \exp(\beta_2 + \gamma_2 d_i) & \text{if } M_{ij} > 0, \end{cases} \quad (1) \end{aligned}$$

where  $b_i$  is a mouse-specific susceptibility variable,  $t_j = j/T$ ,  $T$  is the duration of the study,  $d_i$  is the dose level for mouse  $i$  on the log scale,  $\beta_1$ ,  $\beta_2$  are intercept parameters related to the rate of appearance of spontaneous papillomas, and  $\gamma_1$ ,  $\gamma_2$  are slope parameters associated with exposure. The mouse-specific variable  $b_i$  is assumed to have a normal sampling distribution with mean zero and variance  $\sigma^2$ .

The first expression in Model 1 relates to tumor onset, and involves three parameters:  $\beta_1$ ,  $\gamma_1$ , and  $\sigma^2$ . In a control animal having no tumors,  $\exp(\beta_1)$  can be regarded as the rate of appearance of the first papilloma. In a given week, the probability that a control animal having no tumors develops its first papilloma is  $1 - \exp\{-\exp(\beta_1)\}$ . Typically  $\beta_1 \ll 0$ , since the spontaneous tumor incidence is low during a 26 week study. The probability of detecting the first skin tumor increases with increased time of exposure and dose of a carcinogen. The expression  $(b_i + \gamma_1)t_j d_i$  models this process. For example, if  $\gamma_1 = 0$ , then the dose of the chemical does not affect the probability of an animal developing its initial papilloma during the study. Alternatively, a large value for  $\gamma_1$  implies that the exposed mice develop papillomas more rapidly, or equivalently, a higher proportion of exposed mice develop papillomas during the study. The mouse-specific variable  $b_i$  accounts for possible heterogeneity in response among mice. For example,  $\sigma^2 = 0$  would imply that all animals have the same underlying probability of developing papillomas during the course of the study given equivalent survival. Alternatively,  $\sigma^2 > 0$  implies that mice have different susceptibilities to the development of papillomas. Highly susceptible mice will tend to develop more papillomas and develop them earlier than less susceptible mice.

The second expression in Model 1 relates to tumor multiplicity, and involves two parameters:  $\beta_2$  and  $\gamma_2$ . Once a tumor has appeared, the development of additional tumors in a given animal may proceed at a different rate than the development of the initial tumor. The term  $\exp(\beta_2)$  can be regarded as the spontaneous rate of development of additional papillomas in a control mouse that already has at least one papilloma. The parameter  $\beta_2$  may or may not equal  $\beta_1$ . The parameter  $\gamma_2$  represents the effect of dose on papilloma multiplicity. For example, if  $\gamma_2 = 0$ , then the dose of the test chemical does not affect the probability of developing additional papillomas during the study, once an initial tumor has occurred.

Model 1 follows a simple form that tends to provide a good fit to papilloma data from the few Tg.AC bioassays that we have examined to this point. As more data become available, it may be necessary to refine the model or to include additional parameters to more accurately represent the underlying process that generated the data. For example, the rate of development of papillomas may depend on the body weight of the mouse, on the current tumor burden, or on the age of the mouse. Several studies have demonstrated a positive correlation between body weight and tumor incidence for some tissue sites (Haseman *et al.*, 1997; Turturro *et al.*, 1993), and in some cases it may be necessary to adjust for body weight within Model 1 to avoid biases caused by weight differences across dose groups. Also, as the papilloma burden increases, the rate of developing new tumors may be slowed (and existing tumors may fuse) due to limited space on the animal and/or the inability to provide sufficient nutrients for new tumors to grow and develop. To account for such an effect, we could incorporate a  $\beta_3 M_{ij}$  term in the second expression of Model 1. However, in our experience, this term does not appreciably improve the fit of the model unless the papilloma burden is extremely high. Including a  $\beta_4$  age term to account for an increase in the incidence of spontaneous tumors with age also tends to have little effect on model fit.

**Dose-response modeling.** In carcinogenicity bioassays, there is interest not only in identifying carcinogens, but also in characterizing the magnitude of the tumor response as a function of dose. Estimates of dose-response are useful in comparing compounds, in quantifying risk, and in setting guidelines for

acceptable levels of human exposure. In conventional studies, where tumor multiplicity is rare at most sites, estimates of dose-response are typically based on the proportion of animals with tumors.

In Tg.AC bioassays, the cumulative proportion of mice with detectable papillomas can be estimated for each study week. Under Model 1, the probability that a mouse develops papillomas by week  $j$  of the study varies with dose and between animals according to the model:

$$\begin{aligned} P_{ij} &= 1 - \prod_{k=1}^j \Pr(Y_{ik} = 0 | M_{ik} = 0, b_i, t_k, d_i) \\ &= 1 - \exp\left[-\sum_{k=1}^j \exp\{\beta_1 + (b_i + \gamma_1)t_k d_i\}\right]. \quad (2) \end{aligned}$$

Suppose that  $T_i$  is the number of observations for mouse  $i$  prior to death. The probability that mouse  $i$  develops papillomas during the study is  $P_{iT_i}$ . Mice dying prior to terminal sacrifice will have less opportunity to develop papillomas than mice that survive the duration of the study. Model 2 accounts for variability between mice in survival and in sensitivity to exposure. The expected proportion of animals with papillomas can be estimated for any given study week by integrating the mouse-specific probability,  $P_{ij}$ , across the distribution of the susceptibility variable  $b_i$ . This can be done easily using numerical integration (Shampine *et al.*, 1997), and an S-PLUS program is available at our website ([dir.niehs.nih.gov/dirlecm/transgen/tgac.html](http://dir.niehs.nih.gov/dirlecm/transgen/tgac.html)).

Since Tg.AC mice commonly develop multiple papillomas in response to exposure to a chemical carcinogen, it may be of interest to estimate the effect of dose not only on the proportion of mice with papillomas but also on the mean papilloma burden. Under model (1), the expected maximum papilloma burden achieved for mouse  $i$  by week  $j$  is:

$$\begin{aligned} E\left\{\sum_{k=1}^j Y_{ik}\right\} &= \sum_{k=1}^j \{E(Y_{ik} | M_{ik} = 0, b_i, t_k, d_i)(1 - P_{i,k-1}) \\ &\quad + E(Y_{ik} | M_{ik} > 0, b_i, t_k, d_i) P_{i,k-1}\}. \end{aligned}$$

An average across animals can be calculated for any given study week by integrating the mouse-specific papilloma burden across the distribution of the susceptibility variable  $b_i$ . An S-PLUS program to implement this calculation is available at our website.

**Fitting the model.** Model 1 is in the form of a Markov generalized linear mixed model, and the SAS procedure NLMIXED can be used to obtain approximate maximum likelihood estimates of the parameters. The NLMIXED procedure uses adaptive Gaussian quadrature, which has been found to be one of the most reliable methods of estimation for nonlinear mixed effects models (Pinheiro and Bates, 1995). An example SAS program that uses the NLMIXED procedure to analyze Tg.AC mouse papilloma data can be found at our website ([dir.niehs.nih.gov/dirlecm/transgen/tgac.html](http://dir.niehs.nih.gov/dirlecm/transgen/tgac.html)). Alternatively, Model 1 can be fit using the SAS macro GLIMMIX (Wolfinger, 1993), which uses penalized quaslikelihood for parameter estimation (Breslow and Clayton, 1993).

Another possibility is to follow a Bayesian approach to inference (Carlin and Louis, 1996; Gelman *et al.*, 1996). In Bayesian models, prior uncertainty in the parameters is quantified through the use of prior probability distributions. Inference is based on the posterior distribution of the parameters conditional on the prior and on the data from the current study. In recent years, Bayesian approaches have become widely used (Malakoff, 1999), due in part to the ability to incorporate prior information from previous studies (see, for exam-

ple, Dempster *et al.*, 1983; Dunson and Dinse, 2000; Ibrahim *et al.*, 1998). The Gibbs sampler (Gelfand and Smith, 1990) can be used to fit Model 1 within BUGS, a freely available software package for Bayesian inference Using Gibbs Sampling (Best *et al.*, 1996, [www.mrc-bru.cam.ac.uk/bugs](http://www.mrc-bru.cam.ac.uk/bugs)). An example program is available at our website, and methods for choosing prior distributions based on historical control data are described in Appendix A.

Although the BUGS software is not as widely used or as familiar as SAS, the Bayesian approach has several advantages over maximum likelihood estimation in this setting. First, Bayesian point and interval estimates are appropriate regardless of the sample size, while maximum likelihood estimates rely on large sample approximations. Second, information from previous studies can be included in a Bayesian analysis through the prior distributions, as we illustrate in Appendix A. When the tumor incidence is low, as is the case in Tg.AC mouse bioassays, information from historical studies can improve the sensitivity of statistical tests (Haseman, Huff, and Boorman, 1984). Also, it may be difficult to obtain maximum likelihood estimates when few animals get any tumors in a study. Bayesian analyses that incorporate prior information are not subject to this estimability problem. Third, it is trivial to fit an extended version of Model 1 in BUGS that accommodates extra-variability relative to the Poisson distribution. Such variability may occur in Tg.AC studies, but can be difficult to account for within current software for maximum likelihood estimation. The Bayesian approach has been used previously to analyze data from conventional tumorigenicity studies (Dunson and Dinse, 2000).

**Statistical tests.** Using Model 1, the response to a given chemical can be assessed based on papilloma incidence, latency, and multiplicity. If there is no effect of exposure on the incidence of papillomas then  $\gamma_1 = \gamma_2 = 0$ ; that is, dose has no effect on the rate of appearance of new papillomas prior to or after the appearance of the first papilloma. If there is no effect of exposure on the latency time from the start of the study to the appearance of the first papilloma, then  $\gamma_1 = 0$ . If there is no effect of exposure on papilloma multiplicity, adjusting for animal-to-animal differences in the latency time, then  $\gamma_2 = 0$ . Thus, the null hypotheses corresponding to incidence, latency, and multiplicity are

$$H_{01}: \gamma_1 = \gamma_2 = 0, H_{02}: \gamma_1 = 0, \text{ and } H_{03}: \gamma_2 = 0,$$

respectively.

Within the maximum likelihood approach, we first test  $H_{01}$  to assess an overall dose-response trend in papilloma incidence. This can be done by rejecting  $H_{01}$  if  $2\{L - L(H_{01})\} \geq \chi^2_2(0.05) = 6$ , where  $L$  is the log likelihood under Model 1 and  $L(H_{01})$  is the log likelihood under Model 1 with  $\gamma_1 = \gamma_2 = 0$ . If we fail to reject  $H_{01}$  we conclude there is no evidence of a dose-response trend in papilloma incidence. However, if we reject  $H_{01}$ , then we would like to know whether the trend is due to a shortening of the latency time and/or to an increase in papilloma multiplicity. If  $z_1 = \hat{\gamma}_1/\text{se}(\hat{\gamma}_1) > 1.64$ , we reject  $H_{02}$  and conclude that there is a significant decrease in papilloma latency with increasing dose. If  $z_2 = \hat{\gamma}_2/\text{se}(\hat{\gamma}_2) > 1.64$ , we reject  $H_{03}$  and conclude that there is a significant increase in papilloma multiplicity with increasing dose. All of the information required to conduct these tests is given in the SAS output.

Within the Bayesian approach, samples will be available from the joint distribution of the parameters conditional on the papilloma data from the current study and on the prior, which can potentially be chosen based on historical control data as described in Appendix A. We conclude that there is evidence of an increasing dose-response trend in incidence if

$$\widehat{\Pr} \{ \gamma_1/\widehat{\text{se}}(\gamma_1) + \gamma_2/\widehat{\text{se}}(\gamma_2) > 0 \} \geq 0.95,$$

where this test statistic can be estimated based on a large number of Monte Carlo samples of  $\gamma_1$  and  $\gamma_2$ . Increasing dose-response trends in latency and multiplicity can be assessed by examining 95% intervals for  $\gamma_1$  and  $\gamma_2$ ,

respectively. Strictly positive intervals are suggestive of increasing dose-response trends.

**Large papilloma responses.** Quantification of the papilloma response can be difficult for animals with a high tumor burden. As the number of papillomas on an animal becomes large, it becomes difficult to accurately distinguish individual tumors and papillomas frequently coalesce and continue to grow as a single mass. In such cases, the papilloma count is clearly not the best way to quantify the response. A better measure of effect may be the volume occupied by the skin tumors. However, this volume can be extremely difficult to estimate accurately for live animals.

Since the spontaneous papilloma incidence is low in Tg.AC mice, a high tumor burden typically occurs only with exposure to a clear chemical carcinogen, such as 12-O-tetradecanoyl-phorbol-13-acetate (TPA) or benzene. We recommend discontinuing weekly clinical observations of the papilloma burden on a mouse once the count for that mouse exceeds a threshold of 20 tumors. In our experience, papillomas can be counted with reasonable accuracy and papillomas rarely coalesce when the tumor burden is below this threshold. Mice dying or exceeding the threshold prior to the last observation time will contribute fewer observations to the analysis than mice that complete the study with a small to moderate papilloma burden. Due to the likelihood-based structure of Model 1, the missing observations are ignorable and can be excluded from consideration in the analysis (Laird, 1988).

#### Example: TCDD Study

Data from a National Toxicology Program study of the effect of TCDD exposure on papilloma development are used to illustrate the proposed statistical methodology (van Birgelen *et al.*, 1999). In this study, female hemizygous Tg.AC mice were housed individually and were randomly assigned to treatment groups. Groups of 20 mice received 0, 5, 17, 36, 76, 121, 166, 355, or 760 ng/kg of TCDD topically in acetone three times a week for 26 weeks. On a weekly basis, the number of skin papillomas were recorded for each animal. No TCDD-induced alterations in body weight gain or mortality were observed. However, 38 of the 180 mice died prior to completing the study. This survival rate is similar to the 85% average survival rate that has been reported for vehicle control Tg.AC mice in 26 week studies (Eastin *et al.*, 1998). Early death could not be attributed to any one cause, including the occurrence of odontogenic tumors, which have been observed to cause early mortality in Tg.AC mice. The large number of dose groups in this study facilitates evaluation of the statistical model.

A Kaplan and Meier (1958) estimate of the cumulative proportion of mice with papillomas is plotted in Figure 1 for each study week and dose group. Animals dying prior to terminal sacrifice are not fully at risk of developing skin tumors, and the Kaplan-Meier approach adjusts the proportions for animal survival. Figure 1 shows a clear dose-dependent decrease in papilloma latency. The survival-adjusted average maximum papilloma burden is plotted in Figure 2 for each dose group and study week. Animals that die early have less opportunity to develop papillomas. To account for animal survival, we estimated the average increase in the maximum papilloma burden at each week among surviving animals, and we summed these averages to estimate the average maximum papilloma burden at each week. Figure 2 shows a dose-dependent increase in the maximum papilloma burden.

We first fit Model 1 to the TCDD data using NLMIXED in SAS. The approximate maximum likelihood estimates of the parameters are shown in Table 1, along with standard errors and confidence intervals. The SAS program that was used to obtain these estimates is available at our web site ([dir.niehs.nih.gov/dirlecsm/transgen/tgac.html](http://dir.niehs.nih.gov/dirlecsm/transgen/tgac.html)), and researchers can easily modify this program to analyze their own data sets.

The estimated probability that a vehicle control animal gets one or more papillomas during the course of the study is very small ( $1 - \exp\{-26 \exp(\hat{\beta}_1)\} = 2.3e - 7$ ), which is not surprising since no papillomas were detected in the vehicle control group. The estimated spontaneous rate of developing additional papillomas in a control mouse that already has at least one papilloma is also small ( $\exp(\hat{\beta}_2) = 0.029$ ) as is the expected maximum papilloma burden in a control mouse surviving the duration of the study:

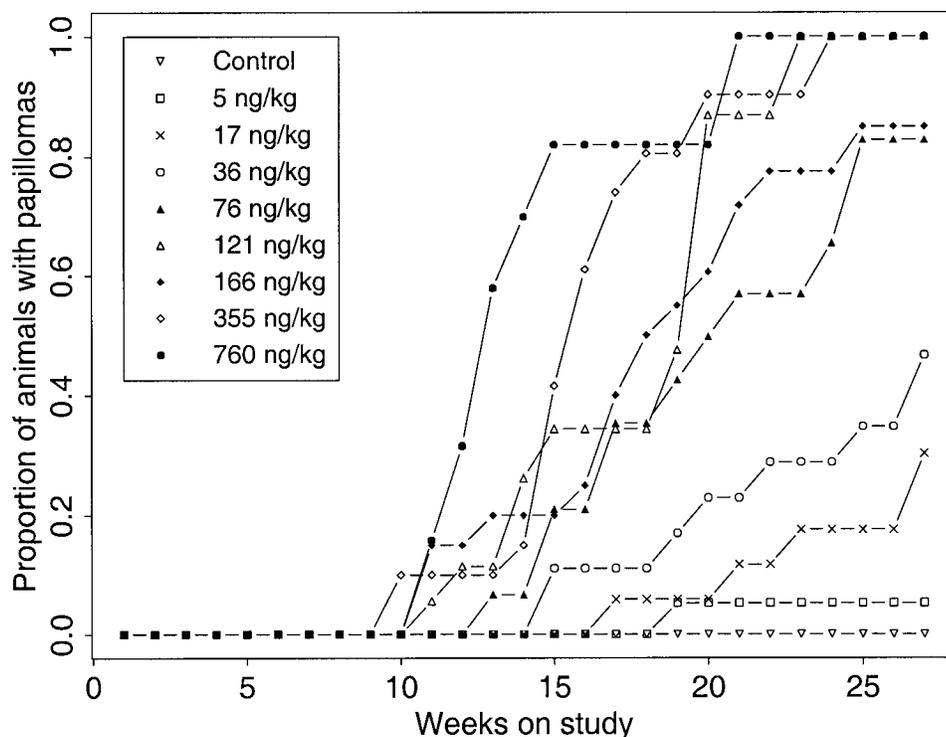


FIG. 1. The cumulative proportion of Tg.AC mice with a detectable papilloma during the course of the study after dermal exposure to TCDD. A Kaplan-Meier approach is used to adjust estimates for animals dying early prior to developing any skin tumors.

$$\sum_{k=1}^{26} (\exp(\hat{\beta}_1) \exp\{(1-k)\exp(\hat{\beta}_1)\} + \exp(\hat{\beta}_2) [1 - \exp\{(1-k)\exp(\hat{\beta}_1)\}]) = 3.1e - 7.$$

Estimation of these values relies on extrapolation downwards from the values in the dosed groups. Historical control data could potentially be included, as described in Appendix A, to improve the reliability of these estimates.

It also appears that there is a strong dose-response trend in papilloma incidence, latency, and multiplicity. The p-value from a likelihood ratio test for a trend in incidence is  $P < 0.001$ , and maximum likelihood based tests for trends in latency and multiplicity are also highly significant ( $P < 0.001$ ). The estimated probability of getting at least one papilloma during 26 weeks of treatment with 5 ng/kg of TCDD for an animal with average susceptibility to TCDD ( $b_i = 0$ ) is

$$1 - \exp\left\{-\sum_{k=1}^{26} \exp(\hat{\beta}_1 + \hat{\gamma}_1 t_k)\right\} = 0.0006.$$

This probability increases to close to one for animals with average susceptibility that are treated with at least 36 ng/kg of TCDD. Less susceptible animals ( $b_i < 0$ ) will have a lower risk of developing papillomas. For example, an animal with susceptibility in the 10th percentile ( $b_i = -12.06$ ), based on the estimated level of animal-to-animal variability ( $\hat{\sigma}^2 = 88.62$ ), has only a 0.017 probability of developing papillomas in the 36 ng/kg group. Thus, it appears that there is high animal-to-animal variability in sensitivity to TCDD, though an animal would have to be in the lower 2.7th percentile in susceptibility to have lower than a 95% risk of developing papillomas with 26 weeks of exposure to 760 ng/kg of TCDD.

We also fit Model 1 using a Bayesian approach implemented with the BUGS

software. We specified non-informative priors for the parameters, in order for the results to be comparable to the results from NLMIXED. However, an informative prior could be chosen as described in Appendix A based on the spontaneous papilloma response observed in previous studies of individually housed Tg.AC mice (e.g., Mahler *et al.*, 1998). Data from studies with group housed animals should not be used to choose the prior, since wounds caused by fighting between cage mates can cause papilloma development in Tg.AC mice (Tennant *et al.*, 1998). The posterior means, standard errors, and confidence limits from the Bayesian analysis are shown in Table 2, and the BUGS program that was used to obtain these estimates is available at our web site. The posterior means from the Bayesian analysis are very similar to the approximate maximum likelihood estimates from NLMIXED.

Since TCDD clearly affects papilloma incidence, latency and multiplicity, the primary objective of analyzing this particular dataset is to assess the fit of the proposed model to the data. Based on Model 1, we estimated the expected proportion of animals with papillomas for each dose group and study week by plugging the parameter estimates from Table 2 into Model 2, and integrating across the distribution of the mouse-specific susceptibility variable  $b_i$ , using the S-PLUS program available at our website. The resulting estimates are plotted in Figure 3. The model-based estimates in Figure 3 approximate the Kaplan-Meier estimates in Figure 1 for each study week. Based on Model 1, we also estimated the expected maximum papilloma burden for each dose group and study week. The resulting estimates are plotted in Figure 4. The model-based estimates in Figure 4 approximate the empirical estimates in Figure 2 for each study week.

## DISCUSSION

Due to recent advances in molecular biology and pharmacology, the rate of development of new drugs has increased substantially. Conventional rodent studies for evaluating drug safety are expensive and timing consuming, and are limited in

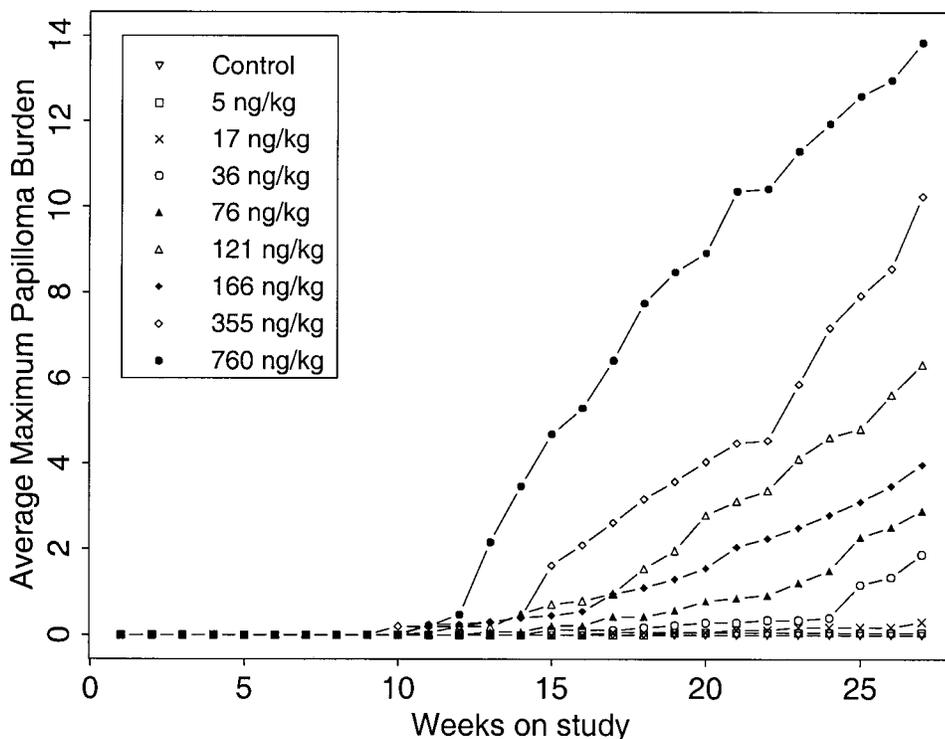


FIG. 2. The average maximum papilloma burden per mouse during the course of the study after dermal exposure to TCDD. Estimates are adjusted for animal survival.

their relevance to human populations. Fortunately, biological advances have also led to the development of transgenic animals that have been shown to result in a carcinogenic response. These animal models have been approved by the U.S. FDA for use in evaluating drug safety. Transgenic mouse bioassays may soon be widely used for rapid carcinogen identification and risk assessment. In order for the results from these studies to be properly interpreted, there is a critical need for the development of new statistical methods.

We have proposed a new approach for the analysis of skin papilloma data from Tg.AC studies. A Poisson mixture model is used to describe the effect of exposure on the rate of increase in the maximum papilloma burden. The model accommodates distinct effects on papilloma latency and multiplicity, as well as variability between mice in sensitivity to exposure. Due to

the structure of the model, it is straightforward to incorporate additional factors to account for body weight, age of the mouse, and overdispersion relative to the Poisson distribution.

Our goal was to develop a method for routine analysis of skin papilloma data from Tg.AC studies. Since the incidence of spontaneous papillomas is very low, a compound that has a weak carcinogenic effect may induce only a few papillomas in a Tg.AC bioassay. Therefore, we have used a simple statistical model that can be fit even if no papillomas are detected for the control animals and only a few are detected for the exposed animals. The model provided an excellent fit to data from a Tg.AC study of TCDD, based on examination of plots of the observed and predicted proportion of mice with papillomas and the average maximum papilloma burden at each study week. We used the predictive log-likelihood approach of Dempster (1974), as described in Karim and Zeger (1992), to further

TABLE 1  
Results of Modeling Increases in Papilloma Response  
Using NLMIXED

Parameter	MLE	Standard error	95% Confidence interval
$\beta_1$	-18.56	1.98	(-22.47, -14.66)
$\beta_2$	-3.534	0.328	(-4.180, -2.888)
$\gamma_1$	37.08	4.35	(28.50, 45.65)
$\gamma_2$	3.379	0.367	(2.656, 4.103)
$\sigma^2$	88.62	26.90	(35.53, 141.7)

Note. Data from NTP study of TCDD (van Birgelen *et al.*, 1999).

TABLE 2  
Results of Modeling Increases in Papilloma Response  
Using BUGS

Parameter	Posterior mean	Standard error	95% Credible interval
$\beta_1$	-18.50	1.98	(-23.40, -15.49)
$\beta_2$	-3.578	0.325	(-4.225, -2.965)
$\gamma_1$	36.85	4.27	(30.49, 47.16)
$\gamma_2$	3.437	0.364	(2.750, 4.160)
$1/\sigma^2$	0.012	0.003	(0.006, 0.019)

Note. Data from NTP study of TCDD (van Birgelen *et al.*, 1999).

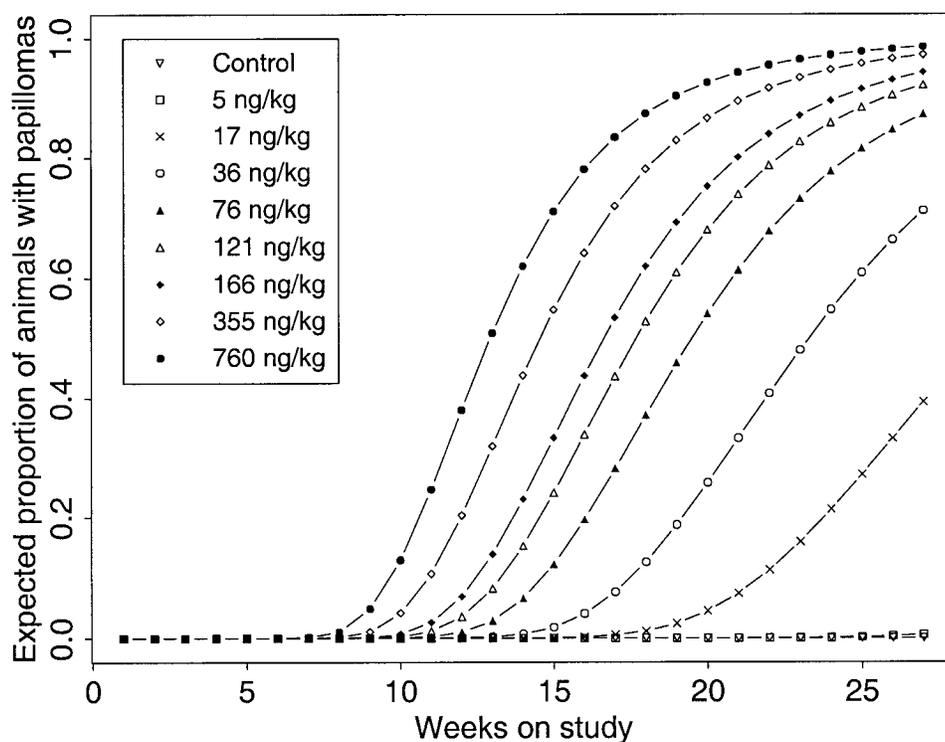


FIG. 3. The estimated proportion of Tg.AC mice with a detectable papilloma during the course of the study after dermal exposure to TCDD. This estimate is based on the fitted parameters shown in Table 2, and is for mice surviving the duration of the experiment.

verify the adequacy of the model. We recommend routinely checking model fit based on plots of the observed and predicted values, and an S-PLUS program for calculating the fitted values is available at our website. There is need for further evaluation of model fit based on data from multiple studies involving smaller sample sizes and a variety of test agents. The TCDD study utilized a relatively high number of dose groups and animals, which made it a good study for model evaluation. In future work we plan to evaluate the operating characteristics of the proposed test procedures for designs that use fewer dose groups and animals per group.

Although we have focused on Tg.AC studies, the statistical methods are applicable to other model systems where tumors are detectable in live animals. These include most animal models of skin and breast cancer (e.g. Boorman *et al.*, 1999). Such models are widely used for assessing the tumorigenic potential of test compounds, for exploring mechanisms of tumor induction, and for identifying agents with chemopreventive attributes.

#### APPENDIX A

##### Choosing the Prior Parameters for the Bayesian Analysis

To fit Model 1 using a Bayesian approach, it is necessary to specify prior distributions for each of the model parameters. Following the standard approach (see, for example, Gilks *et al.*, 1993), we assign a gamma ( $a_{01}, a_{02}$ ) prior for  $1/\sigma^2$ , where  $a_{01}/a_{02}$  is the prior mean and  $a_{01}/a_{02}^2$  is the prior variance. In

addition, we choose normal priors for the parameters  $\beta_1, \beta_2, \gamma_1$ , and  $\gamma_2$ :

$$\beta_1 \sim N(m_{01}, v_{01}), \beta_2 \sim N(m_{02}, v_{02}), \gamma_1 \sim N(m_{03}, v_{03}),$$

$$\gamma_2 \sim N(m_{04}, v_{04}), \quad (1)$$

where  $m_{01}, m_{02}, m_{03}, m_{04}$  are prior means representing the investigators best “guess” at the parameter values based on all information that is available prior to running the current study, and  $v_{01}, v_{02}, v_{03}, v_{04}$  are prior variances which are chosen to reflect the uncertainty in this choice. To choose a noninformative prior, set  $a_{01} = 0.001, a_{02} = 0.001, m_{01} = m_{02} = m_{03} = m_{04} = 0$  and  $v_{01} = v_{02} = v_{03} = v_{04} = 1000$ . A Bayesian analysis that uses non-informative priors often gives similar results to a maximum likelihood analysis.

While there is typically limited prior information about  $\gamma_1$  and  $\gamma_2$ , the parameters representing the effect of dose on tumor latency and multiplicity, respectively, historical control data are informative about  $\beta_1$  and  $\beta_2$ , the parameters related to the spontaneous tumor incidence rate prior to and after the appearance of the first papilloma. If weekly papilloma counts are available for historical control animals, the prior means  $m_{01}$  and  $m_{02}$  can be set equal to the estimates for  $\beta_1$  and  $\beta_2$ , respectively, from an analysis of the historical control data. Similarly, the prior variance  $v_{01}$  and  $v_{02}$  can be set equal to the estimated variance from such an analysis. This approach is referred to as coherent Bayesian updating, and is a standard

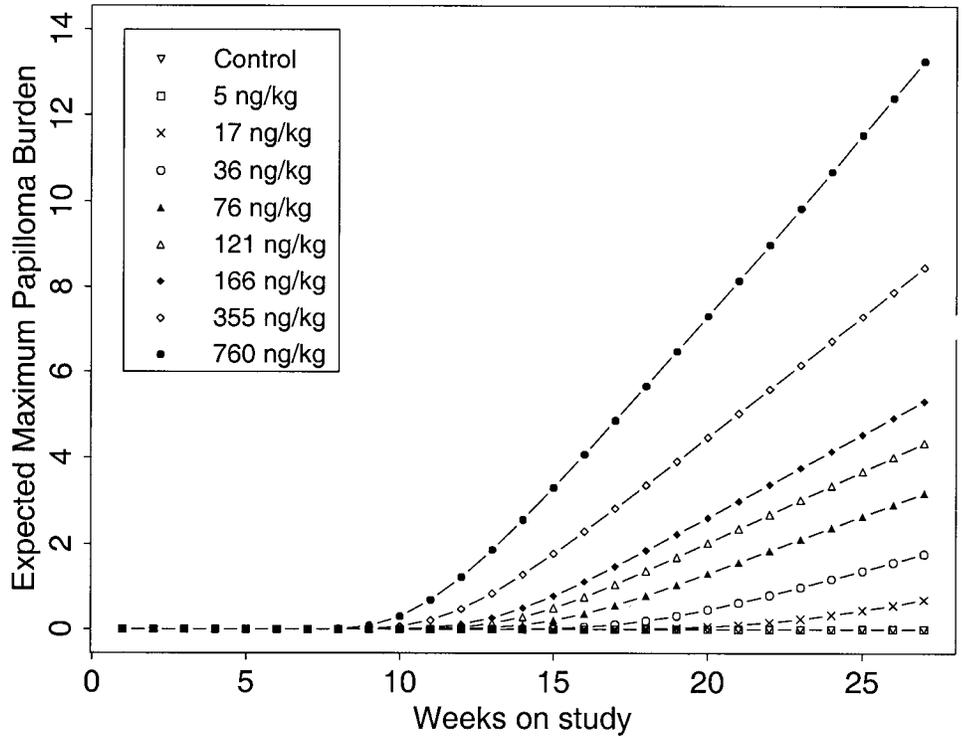


FIG. 4. The estimated maximum papilloma burden per mouse during the course of the study after dermal exposure to TCDD. These estimates are based on the fitted parameters shown in Table 2, and are for mice surviving the duration of the experiment.

method in Bayesian analysis (see, for example, Gelman *et al.*, 1996).

If weekly papilloma counts are not available for the historical studies, we can choose  $m_{01}$ ,  $m_{02}$ ,  $v_{01}$ ,  $v_{02}$  based on summary statistics. In this case, we let

$$m_{01} = \log\{-\log(1 - \hat{P}_0)/T\}, \tag{2}$$

where  $\hat{P}_0 = (X_0 + 0.5)/(N_0 + 0.5)$ ,  $X_0$  is the number of vehicle control animals with papillomas in previous studies,  $N_0$  is the total number of vehicle control animals in the historical database, and  $T$  is the study duration ( $T = 26$  for 26 week studies). We include 0.5 as a correction factor for low incidence. Under the assumption that  $\beta_1$  is constant from study to study, we can choose the prior variance by letting

$$v_{01} = \frac{\hat{P}_0(1 - \hat{P}_0)}{N_0 m_{01}^2 T^4 \exp(2m_{01})}, \tag{3}$$

where this expression is derived using the delta method (Morgan, 1992). To allow for a reasonable degree of study-to-study variability, one can multiply  $v_{01}$  by a factor of 10. We choose  $m_{02}$  by solving the following equation:

$$\hat{M}_0 = \sum_{k=1}^T \{\exp(m_{01})S_{k-1} + \exp(m_{02})(1 - S_{k-1})\}, \tag{4}$$

where  $\hat{M}_0 = (Z_0 + 0.5)/(N_0 + 0.5)$ ,  $Z_0$  is the total number of papillomas in the vehicle control mice in the historical studies, and  $S_{k-1} = \exp\{-\sum_{h=1}^{k-1} \exp(m_{01})\}$ . Under the assumption that  $\beta_2$  is constant from study to study, we can choose the prior variance by letting

$$v_{02} = \exp(2\mu_{02}) \left\{ \sum_{k=1}^T (1 - S_{k-1}) \right\}^2 \hat{V}_0, \tag{5}$$

where  $\hat{V}_0$  is the estimated between animal variability in the maximum papilloma burden for vehicle control animals in the historical studies. To allow for a reasonable degree of study-to-study variability, one can multiply  $v_{02}$  by a factor of 10.

## APPENDIX B

### Choosing Initial Values for the Maximum Likelihood Analysis

To fit Model 1 using either the Bayesian or the maximum likelihood approach, it is necessary to choose initial values for the parameters. While the Bayesian approach is not sensitive to the initial values, current maximum likelihood programs may fail to converge if the initial values are unreasonable. To choose initial values for  $\beta_1$ ,  $\beta_2$ ,  $\gamma_1$ , and  $\gamma_2$ , we can modify the procedure described in Appendix A for choosing the prior parameters  $m_{01}$  and  $m_{02}$ :

1. For each dose group, choose  $m_{01}$  and  $m_{02}$ , as described in the final paragraph of Appendix A, using the data from the current study instead of the historical controls.

2. Let the initial values for  $\beta_1$  and  $\gamma_1$  be the least squares estimates of the intercept and slope, respectively, for the simple linear model with  $m_{01}$  (selected in step 1) as the dependent variable and one half the log dose as the independent variable.

3. Let the initial values for  $\beta_2$  and  $\gamma_2$  be the least squares estimates of the intercept and slope, respectively, for the simple linear model with  $m_{02}$  (selected in step 1) as the dependent variable and the log dose as the independent variable.

In addition, we set the initial value for  $1/\sigma^2$  to 0.1.

#### ACKNOWLEDGMENTS

We would like to thank Jun Zhai and Richard Morris for providing expert programming. Thanks go also to Judson Spalding, Gregg Dinse, and Skip Eastin for their useful comments. The TCDD study was performed under contract with the National Toxicology Program by the staff at Battelle Memorial Laboratories, Columbus, OH, USA under supervision of Jerry D. Johnson.

#### REFERENCES

- Best, N. G., Spiegelhalter, D. J., Thomas, A., and Brayne, C. E. G. (1996). Bayesian analysis of realistically complex models. *Journal of the Royal Statistical Society: Series A* **159**, 323–342.
- Boorman, G. A., Anderson, L. E., Morris, J. E., Sasser, L. B., Mann, P. C., Grumbein, S. L., Hailey, J. R., McNally, A., Sills, R. C., and Haseman, J. K. (1999). Effect of 26-week magnetic field exposure in a DMBA initiation-promotion mammary gland model in Sprague-Dawley rats. *Carcinogenesis* **20**, 899–904.
- Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.* **88**, 9–25.
- Burnett, R. T., Ross, W. H., and Krewski, D. (1995). Nonlinear mixed regression models. *Environmetrics* **6**, 85–99.
- Carlin, B. P., and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.
- Contrera, J. F., and DeGeorge, J. J. (1998). *In vivo* transgenic bioassays and assessment of the carcinogenic potential of pharmaceuticals. *Environ. Health Perspect.* **106**(Suppl. 1), 71–80.
- Dempster, A. P. (1974). The direct use of likelihood for significance testing. In *Proceedings of Conference on Foundational Questions in Statistical Inference*, O. Barndorff-Nielsen, P. Blaesild, and G. Schou (eds). Department of Theoretical Statistics, University of Aarhus.
- Dempster, A. P., Selwyn, M. R., and Weeks, B. J. (1983). Combining historical and randomized controls for assessing trends in proportions. *J. Am. Statist. Ass.* **78**, 221–227.
- Dewanji, A., Goddard, M. J., Krewski, D., and Moolgavkar, S. H. (1999). Two stage model for carcinogenesis: number and size distributions of premalignant clones in longitudinal studies. *Math. Biosci.* **155**, 1–12.
- Dunson, D. B. (2000). Models for papilloma multiplicity and regression: applications to transgenic mouse studies. *Appl. Statist.* **49**, 19–30.
- Dunson, D. B., and Haseman, J. K. (1999). Modeling tumor onset and multiplicity using transition models with latent variables. *Biometrics* **55**, 965–970.
- Eastin, W. C. (1998). The U.S. National Toxicology Program evaluation of transgenic mice as predictive models for identifying carcinogens. *Environ. Health Persp.* **106**, 81–84.
- Eastin, W. C., Haseman, J. K., Mahler, J. F., and Bucher, J. R. (1998). The National Toxicology Program evaluation of genetically altered mice as predictive models for identifying carcinogens. *Toxicol. Pathol.* **26**, 461–473.
- Fung, K. Y., Lin, X., and Krewski, D. (1998). Use of generalized linear mixed models in analyzing mutant frequency data from the transgenic mouse assay. *Environ. Mol. Mutagen.* **31**, 48–54.
- Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1996). *Bayesian Data Analysis*. Chapman & Hall, London.
- Gilks, W. R., Wang, C. C., Yvonnet, B., and Coursaget, P. (1993). Random-effects models, for longitudinal data using gibbs sampling. *Biometrics* **49**, 441–453.
- Haseman, J. K., Huff, J., and Boorman, G. A. (1984). Use of historical control data in carcinogenicity studies in rodents. *Toxicol. Pathol.* **12**, 126–135.
- Haseman, J. K., Young, E., Eustis, S. L., and Hailey, J. R. (1997). Body weight-tumor incidence correlations in long-term rodent carcinogenicity studies. *Toxicol. Pathol.* **25**, 256–263.
- Ibrahim, J. G., Ryan, L. M., and Chen, M. H. (1998). Using historical controls to adjust for covariates in trend tests for binary data. *J. Am. Statist. Ass.* **93**, 1282–1293.
- Kaplan, E. L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Statist. Ass.* **53**, 457–481.
- Karim, M. R., and Zeger, S. L. (1992). Generalized linear models with random effects: salamander mating revisited. *Biometrics* **48**, 631–644.
- Kokoska, S. M., Hardin, J. M., Grubbs, C. J., and Hsu, C. (1993). The statistical analysis of cancer inhibition/promotion experiments. *Anticancer Res.* **13**, 1357–1363.
- Laird, N. M. (1988). Missing data in longitudinal studies. *Stat. Med.* **7**, 305–315.
- Leder, A., Kuo, A., Cardiff, R. D., Sinn, E., and Leder, P. (1990). v-Ha-ras transgene abrogates the initiation step in mouse skin tumorigenesis: effects of phorbol esters and retinoic acid. *Proc. Natl. Acad. Sci. USA* **87**, 9178–9182.
- Mahler, J. F., Flagler, N. D., Malarkey, D. E., Mann, P. C., Haseman, J. K., and Eastin, W. (1998). Spontaneous and chemically induced proliferative lesions in Tg.AC transgenic and p53-heterozygous mice. *Toxicol. Pathol.* **26**, 501–511.
- Malakoff, D. (1999). Bayes offers a ‘new’ way to make sense of numbers. *Science* **286**, 1460–1464.
- Morgan, B. J. T. (1992). *Analysis of Quantal Response Data*. Chapman & Hall, London.
- Piepho, H. P. (1999). Analysing disease incidence data from designed experiments by generalized linear mixed models. *Plant Pathology* **48**, 668–674.
- Pinheiro, J. C., and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J. Comput. Graph. Statist.* **4**, 12–35.
- Shampine, L. F., Allen, R. C., and Pruess, S. (1997). *Fundamentals of numerical computing*. John Wiley and Sons, New York.
- 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. *Toxicol. Sci.* **49**, 241–254.
- Spalding, J. W., Momma, J., Elwell, M. R., and Tennant, R. W. (1993). Chemically induced skin carcinogenesis in a transgenic mouse line (Tg.AC) carrying a v-Ha-ras gene. *Carcinogenesis* **14**, 1335–1341.

- Tennant, R. W., French, J. E., and Spalding, J. W. (1995). Identifying chemical carcinogens and assessing potential risk in short-term bioassays using transgenic mouse models. *Environ. Health Perspect.* **103**, 942–950.
- Tennant, R. W., Spalding, J. W., and French, J. E. (1996). Evaluation of transgenic mouse bioassays for identifying carcinogens and noncarcinogens. *Mutat. Res.* **365**, 119–127.
- Tennant, R. W., Tice, R. R., and Spalding, J. W. (1998). The transgenic Tg.AC mouse model for identification of chemical carcinogens. *Toxicol. Lett.* **102–103**, 465–471.
- Turturro, A., Duffy, P. H., and Hart, R. W. (1993). Modulation of toxicity by diet and dietary macronutrient restriction. *Mut. Res.* **295**, 151–164.
- van Birgelen, A. P. J. M., Johnson, J. D., Fuciarelli, A. F., Toft II, J. D., Mahler, J., and Bucher, J. R. (1999). Dose- and time-response of TCDD in Tg.AC mice after dermal and oral exposure. *Organohalogen Compounds* **42**, 235–239.
- Wolfinger, R. (1993). The GLIMMIX SAS Macro. Cary, NC: SAS Institute Inc.
- Yamamoto, S., Mitsumori, K., Kodama, Y., Matsunuma, N., Manabe, S., Okamiya, H., Suzuki, H., Fukuda, T., Sakamaki, Y., Sunaga, M., Nomura, G., Hioki, K., Wakana, S., Nomura, T., and Hayashi, Y. (1996). Rapid induction of more malignant tumors by various genotoxic carcinogens in transgenic mice harboring human prototype c-Ha-ras gene than in control nontransgenic mice. *Carcinogenesis* **17**, 2455–2461.
- Yamamoto, S., Urano, K., Koizumi, H., Wakana, S., Hioki, K., Mitsumori, K., Kurokawa, Y., Hayashi, Y., and Nomura, T. (1998). Validation of transgenic mice carrying the human prototype c-Ha-ras gene as a bioassay model for rapid carcinogenicity testing. *Environ. Health Perspect.* **106**, 57–69.
- Zeger, S. L., and Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *J. Am. Statist. Ass.* **86**, 79–86.