Published in final edited form as:

Regul Toxicol Pharmacol. 2007 July ; 48(2): 135–147.

On use of the multistage dose-response model for assessing laboratory animal carcinogenicity

Daniella Nitcheva^a, Walter W. Piegorsch^b, and R. Webster West^c

^a Department of Epidemiology and Biostatistics, Norman J. Arnold School of Public Health, University of South Carolina, Columbia, SC 29208, USA

^b Department of Mathematics and BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA

^c Department of Statistics, 3143 TAMU, Texas A&M University, College Station, TX 77843, USA

Abstract

We explore how well a statistical multistage model describes dose-response patterns in laboratory animal carcinogenicity experiments from a large database of quantal response data. The data are collected from the U.S. EPA's publicly available IRIS data warehouse and examined statistically to determine how often higher-order values in the multistage predictor yield significant improvements in explanatory power over lower-order values. Our results suggest that the addition of a second-order parameter to the model only improves the fit about 20% of the time, while adding even higher-order terms apparently does not contribute to the fit at all, at least with the study designs we captured in the IRIS database. Also included is an examination of statistical tests for assessing significance of higher-order terms in a multistage dose-response model. It is noted that bootstrap testing methodology appears to offer greater stability for performing the hypothesis tests than a more-common, but possibly unstable, "Wald" test.

Keywords

Bootstrap hypothesis test; Cancer; Dose-response modeling; Multistage model

1. Introduction

Among statistical dose-response models used in laboratory carcinogenicity testing, the Armitage-Doll multistage model (Armitage and Doll, 1954) is perhaps one of the most popular when data occur in the form of quantal outcomes. To be sure, many alternative dose-response models exist for this problem, including the well-known logistic, probit, gamma multi-hit, and Weibull forms (Krewski and van Ryzin, 1981; Edler *et al.*, 2005). While each of these models can be useful in select settings, the multistage approach enjoys a particular combination of potential biological plausibility and mathematical simplicity that is difficult to reproduce. In the former case, the multistage assumption posits that in organisms exposed to a carcinogen, a normal cell targeted by the toxin endures multiple, say, $k \geq 1$, insults ("stages") before becoming malignant and that the cancer is then induced by the first cell to enter malignancy (Edler and Kopp-Schneider, 1998). We refer readers to Edler *et al.* (2005), Moolgavkar (2004), Ritter *et al.* (2003), and the references therein for more theoretical background on the

* Corresponding author. Email: nitcheva@gwm.sc.edu, Telephone: 1-803-777-7353, fax: 1-803-777-2524

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

model. As greater scientific study has centered attention on the multistage paradigm, the model's form has evolved in both complexity and biological validity, and now includes multiple birth-death processes and elaborate stochastic pathways/stages for tumor onset and progression (Sherman and Portier, 1998; Moolgavkar, 1999). While debate continues on the full applicability such a model can achieve for explaining the intricacies of mammalian cancer induction, its use in cancer dose-response modeling remains quite common (Bailer and Smith, 1994; Luebeck and Moolgavkar, 1996; Sand *et al.*, 2002).

The mathematical simplicity of the multistage model also enjoys a certain popularity. Suppose we denote by $\pi(d)$ the probability that an exposed subject will express the cancer after exposure at a continuous dose $d \geq 0$. For a k -stage model, the multistage assumptions can then lead to a simple dose-response function of the form

$$\pi(d) = 1 - \exp \left\{ -\beta_0 - \beta_1 d - \beta_1 d^2 - \dots - \beta_k d^k \right\}, \quad (1)$$

where $\beta_j \geq 0, j = 0, \dots, k$. In this general form, the multistage model does not place a limit on the number of stages k , but in practice k is usually pre-specified to be no larger than the number of non-zero doses tested, or in some cases the number of non-zero doses tested minus 1. For practical use, one often encounters dose-response models whose core predictor function is linear or quadratic in dose (Gaylor and Slikker, 1990; Chen and Gaylor, 1992; Bailer and Oris, 1993; Kodell and West, 1993), which in (1) would correspond to $k = 1$ or $k = 2$.

Unavailable in the literature, however, is any formal assessment of how well the chosen value of k , and hence the multistage predictor, fits quantal data from carcinogenicity studies. A recent study by Chiu *et al.* (1999) offsets this trend: those authors found that a form of staged, initiator-promoter model produced a better fit than a popular approximation of (1) known as the linearized multistage (LMS) model (Crump *et al.*, 1976) when applied to a set of six carcinogenicity data sets. Other recent studies have illustrated similar shortcomings with the LMS model (Lovell and Thomas, 1996; Nitcheva *et al.*, 2005), and as such we do not consider it here. For a study of the LMS approach using cancer bioassay data from the large U.S. EPA IRIS database—which we also employ below—see Subramaniam *et al.* (2006).

The results Chiu *et al.* achieved applied to a more complicated mathematical characterization for $\pi(d)$ than we express in (1), but their model did have strong similarities to the general multistage form. Perhaps more importantly, it represented an extension beyond the simple LMS approach and its close cousin, the $k = 1$ model with $\pi(d) = 1 - \exp\{-\beta_0 - \beta_1 d\}$. In a similar vein, we have found that if an incorrect specification of k is made when employing the multistage form in (1), statistical inferences on the underlying dose-response function can be negatively affected (Piegorisch *et al.*, 2006). This encouraged us to ask the associated statistical question, how often in practice does use of higher-order multistage terms significantly improve the fit of (1) to quantal response data? Towards this end, we turned to one of the world's largest public-domain data sources, the Integrated Risk Information System (IRIS) Database for risk assessment developed by the U.S. Environmental Protection Agency (EPA); see U.S. EPA (1997). The IRIS database contains information on an extensive number of chemical substances. From it, we identified over 90 individual quantal data sets available for study and attempted to ascertain how successfully various values of $k \geq 1$ from (1) fit different dose-response patterns. The following sections detail our results.

2. Materials and methods

2.1. Statistical Model

As mentioned in §1, our interest centers on quantal response data from dose-response studies of animal carcinogenicity (Haseman, 1990; Ahn and Kodell, 1998). We assume that each individual study generates data as counts, Y_i , of animals or other subjects that exhibit a specific tumor, out of N_i total subjects all exposed for the same duration to the hazardous agent. The index i is associated with differing doses or levels, d_i , of the carcinogenic stimulus, $i = 1, \dots, n$. Under standard quantal response assumptions, Y_i possesses a binomial distribution, where $\pi(d)$ is the underlying risk that an individual animal will develop the cancer. The expression in Equation (1) for $\pi(d)$ presents $k + 1$ unknown parameters, $\beta_0, \dots, \beta_k$ that must be estimated, and about which inferences are to be drawn.

2.2. Statistical Parameter Estimation

To fit the multistage model in (1) to quantal response data, we appeal to the method of maximum likelihood (Piegorisch and Bailer, 1997, Sec. 2.3). Since the model constrains the unknown β_k parameters to be non-negative, the maximum likelihood estimates (MLEs), denoted as $\hat{\beta}_k$, must be derived via constrained optimization. We also require the study design to have at least $k + 1$ dose levels, where we view any control group as a “zero-dose” level and set $d = 0$. Thus for a given k we only consider data sets with at least this many levels.

With each different model we applied, we also tested for goodness-of-fit: if a data set exhibited a significantly poor fit we excluded it from consideration for that model. For the goodness-of-fit statistic, we followed Anderson *et al.* (1983) and employed a generalized χ^2 statistic with degrees of freedom set to equal the number of dose levels minus the number of nonzero multistage parameters. As in Anderson *et al.* (1983), we operated at the 1% significance level to test the goodness-of-fit.

2.3. Statistical Hypothesis Testing

To determine whether addition of a higher-order polynomial term in (1) can significantly improve the fit of the model, we employed standard statistical hypothesis tests. Given the multistage model in (1) for a fixed $k \geq 1$, we test the set of hypotheses

$$H_0 : \beta_k = 0 \text{ vs. } H_A : \beta_k > 0. \quad (2)$$

Significant departure ($P \leq 0.05$) from H_0 in the direction of H_A provides evidence that the k th order term in the model supplies an improvement in the model's ability to describe the data. Conversely, failure to reject H_0 suggests that the use of the higher-order term adds little or no significance to the quality of the fit. Notice that the alternative hypothesis in (2) is one-sided, since the non-negativity constraint on β_k removes consideration of any negative values.

For testing (2), a common approach involves appeal to likelihood theory (Lehmann and Casella, 1998), since the statistical features of the MLEs are well-known in large samples. From this, construction of a standard operand known as a *Wald statistic* is calculable, taking the form $z_{Wald} = \hat{\beta}_k / se(\hat{\beta}_k)$. Here $se(\hat{\beta}_k)$ is the standard error of the MLE, which can be derived from components of the likelihood equations known as the expected Fisher information elements (Piegorisch and Bailer, 1997, Sec. 2.3). At 5% significance, we reject H_0 in favor of H_A in (2) when z_{Wald} exceeds the upper 5% critical point from a standard normal reference distribution.

Use of the Wald approach has been shown to be valid with extremely large sample sizes if the true values of β_j are strictly positive, *for all j* (Guess and Crump, 1976). We are assessing in (2) whether $\beta_k = 0$, however, and so by testing on this boundary of the parameter space, we

bring into question the large-sample features of z_{Wald} . Since use of a proper statistical testing technique is critical to our study, this potential instability with the Wald test led us also to consider a computer-intensive alternative which has gained favor in the statistical literature. Known as *bootstrap resampling* (Piegorsch and Bailer, 1997, Sec. 2.5), the method avoids restrictive assumptions on the distribution of the parameters and can lead in some cases to more robust inferences. It appeals to repeated resampling from the original sample in order to construct a reliable estimate of the distributional properties of the MLEs. For interested readers, we give more details on the bootstrap method in Appendix A.

To our knowledge, no statistical guidance exists on whether use of a bootstrap approach can overcome the small-sample instability possible when employing inferences such as a Wald test for assessing (2). Thus as part of our study we conducted a Monte Carlo simulation experiment to evaluate the operational characteristics of the two testing approaches. We found that the Wald test was very unstable, and could lead to radical departures from nominal error rates, even in large samples. The bootstrap-based method was far more stable, and is our method of choice for testing (2) under a multistage model. Details of the simulations are given in Appendix B.

Using our bootstrap methodology, our goal is to determine how often high-order versions of (1) such as with $k = 2$ or $k = 3$ adequately describe the dose response with animal carcinogenicity data. To do so, we applied the bootstrap test to a series of data sets taken from the IRIS database (detailed next). With each data set, we started with a high-order value for k and tested H_{0k} : $\beta_k = 0$ vs. H_{Ak} : $\beta_k > 0$ via the bootstrap test. If H_{0k} was rejected, we stopped testing and labeled the dose-response as consistent with that level of k . If not, we continued testing the significance of β_k over decreasing values of k until a significant result was achieved or until no significant dose response was evidenced.

Note that we applied no correction for possible multiplicity if more than one bootstrap test was performed on a given data set, since the numbers of multiple inferences per data set depended themselves on the data. Although corrections for such a complex testing schema are possible, we chose to err in favor of simplicity regarding this issue of multiplicity adjustment. Readers should be aware, however, that without multiplicity correction(s), our results may overstate the true level of significance we observed among the IRIS dose-response patterns.

2.4. The IRIS data base

The U.S. EPA's Integrated Risk Information System (IRIS) warehouse is an online, public-domain electronic database (U.S. EPA, 1997), available at <http://www.epa.gov/iris/>. It contains broad summary compilations of toxicological and other health sciences data from a variety of original sources. Depending on the study being reported, the adverse health effects available on IRIS can be both cancerous and non-cancerous; we centered attention on only cancerous outcomes. For each chemical listed in the database, descriptive and quantitative information is given which often, but not always, includes data from laboratory or field experiments, along with corresponding references. In some cases, the references themselves lead the user to original, online data.

Our interest was in those chemicals for which one or more data sets contained sufficient information to test the hypotheses in (2) under a multistage model. Since we centered attention on cancerous outcomes only, we focused our search on quantal response data where the cancer incidence was reported as the proportion of animals exhibiting a tumor. To assess a dose response, we required that an experiment provide quantal data for at least 3 dose levels, including a control level if it was present. (If so, we set the control dose to $d = 0$.) We also required that the exposure durations across dose group be essentially constant. Note that for some long-duration/long-expression studies, it might be possible to incorporate animal survival

into the statistical model in order to adjust for any intercurrent mortality (Dinse, 1994). Such information was not typically available from IRIS, since the site displays mainly the summary quantal data. The survival information may be available from the original source(s) of the data, but for consistency we chose to work only with what was given on the IRIS site. Nonetheless, it is important to keep in mind that differential mortality within a study may at times adversely affect statistical inferences that do not adjust for it (Bailer and Portier, 1988), such as tests of goodness-of-fit of (1) or of H_0 in (2). The results we present below should be viewed with this caveat in mind.

For simplicity, we employed data only on rat or mouse tumorigenicity and only if the animals' exposure to the substances was oral (diet, drinking water, or gavage) or thru inhalation. We were able to identify 91 such IRIS data sets, representing 60 different chemical substances; Table 1 lists the IRIS data sets as we captured them.

Some attention has been raised over the extent to which data inclusion in IRIS reflects modern carcinogenicity experimentation (Mills, 2006), and it is not our intent to enter into this debate here. For our purposes, the data in Table 1 represent tumorigenicity results from a reasonably broad selection of sources, and over a variety of tumor types.

We should note that our inclusion of these 91 data sets is not intended to allude nor contest that the underlying mechanisms leading to cancer in these animals is multi-staged in nature. Neither is it the case that any specific value of k chosen for any chemical's dose response guarantees that the carcinogenic mechanism is specifically k -staged in that case. We do appreciate the biological applicability that the multistage assumptions may have for cancer induction in laboratory animals; however, our study's results are necessarily limited to the quality of the statistical fit of model (1) to the quantal data we collected from IRIS, and to whether such a model can be profitably employed when performing statistical analyses on similar data. The larger issue of the multistage model's biological plausibility extends beyond the scope of our presentation.

3. Results

To each of the 91 data sets we downloaded from the IRIS database, we fit a multistage model with order k as described in §2. The dose, d , was taken in units as listed in Table 1; whenever available, we used units of the administered dose. We allowed k to be as large as the number of non-zero dose levels minus 1. Among the 91 data sets we studied, only 4 were designed with more than four non-zero dose levels, leaving us with so few studies to explore the validity of $k \geq 4$ as to make the assessment untenable. Thus we began by studying $k = 3$.

To assess the value of setting $k = 3$ in (1), 14 of the 91 data sets provided sufficient data for statistical evaluation under our selection criteria. Among these 14 dose-response patterns, 12 passed our goodness-of-fit criterion for a multistage fit, but none were able to reject the null hypothesis after performing our bootstrap test of $H_0: \beta_3 = 0$. Although based on only a limited number of statistical evaluations, this result nonetheless calls into question the explanatory value of employing 3rd-order terms in (1) with these sorts of quantal response data.

With no data sets showing significant improvement in fit at $k = 3$, we moved to $k = 2$ and tested the null hypothesis of $H_0: \beta_2 = 0$. Including the 14 studies from the $k = 3$ assessment above, a total of 50 data sets provided at least three non-zero doses for dose-response evaluation. Of these, 44 exhibited an adequate multistage fit, among which only 10 (23%) exhibited a significant improvement by including a non-zero β_2 parameter.

Statistically-attuned readers may note that our restriction to at least three non-zero doses when fitting the $k = 2$ form is not a minimal statistical necessity; we employ it primarily due to its

traditional, accepted status for fitting the model in (1). From a purely statistical perspective, a $k = 2$ model can in fact be fit to data with as few as three distinguishable dose groups, total. If we took this approach with the IRIS data sets we collected, the $k = 2$ model could be fit to all 91 sets. In this case, a total of 68 exhibited an adequate multistage fit, among which 14 (21%) exhibited a significant improvement by including a non-zero β_2 parameter. In either case, we find that the $k = 2$ model does have the potential to provide a significant quality fit to animal carcinogenicity data, at least as seen in the IRIS database, but perhaps only about 20–25% of the time.

Moving to the $k = 1$ model, we found that 67 of our 91 data sets exhibited an adequate multistage fit. Of these, β_1 was significantly greater than zero in 64 (96%) of them. (The other 3 data sets did not exhibit a significantly increasing dose response.) Table 2 summarizes all of our results for testing (2). Our findings for $k = 1$ may not be altogether surprising, since the simple $k = 1$ model is known to be effective in identifying an increasing dose response; indeed, when testing $H_0: \beta_1 = 0$ with $k = 1$ the significance of the trend in dose response can drive much of the effect identified by the statistical test. Nonetheless, our results do confirm the ability of the $k = 1$ form to serve as a valid dose-response assessor with quantal carcinogenicity data.

4. Discussion

Recent research into the capability of multistage models to describe dose-response patterns in carcinogenicity data (Chiu *et al.*, 1999; Subramaniam *et al.*, 2006) prompted us to explore this issue from a statistical, dose-response perspective. We collected together a large base of quantal response carcinogenicity data from the U.S. EPA's publicly available IRIS data warehouse and from this identified 91 individual laboratory animal carcinogenicity data sets for statistical examination. Our goal was to identify whether and how often higher-order values of k in Equation (1), centering on orders as "high" as $k = 3$ and $k = 2$, yielded statistically significant improvements in explanatory power. Our results suggest that the addition of a third ($k = 2$) parameter to the model only improves the fit about 20–25% of the time. Adding an even higher-order term at $k = 3$ apparently does not contribute to the fit at all.

It is worthwhile to note that the non-negativity requirement we enforce on each β_j is connected to the base multistage assumptions that several mutations must occur in order for a cell to turn cancerous. From this, the coefficients are often interpreted as representing components of the mutation process. If, as a reviewer has noted, one considers additional factors such as cell death, there is justification for building cancer risk models with negative coefficients. That is, if the general form of the dose-response model has a negative second derivative, the slope can decrease as dose increases. For such a underlying effect, a $k = 2$ multistage fit will seem to be as good as one can get, since positive higher order terms will make the fit worse. Connected with this, one often sees in cancer risk assessment that dose-response data are sometimes pre-whitened to eliminate observations where the only significant elevation in tumor rates is observed at the highest dose level. When left unaltered, such dose-response patterns are thought to support unusual, incorrect, and even irrelevant risk analytic decisions, particularly if the high dose exceeds the so-called "maximum tolerated dose" (MTD) for the animals under study. In fact, a number of data sets in Table 1 do exhibit such a response, although the issue of whether an MTD was exceeded in those studies is beyond the scope of our investigation. (Goodness-of-fit P -values for these data sets varied widely, with no determinable pattern.) The concern of whether to perform high-dose pre-whitening has engendered some debate among risk assessors and while important to keep in mind, we chose not to integrate it into our selection and analysis process. For our purposes the larger issue of whether multistage models fit these various data sets remains pertinent whether or not such high-dose exclusions are undertaken. Indeed, how laboratory animal data can be used to evaluate the shape of dose-response curves such as (1) is an ongoing risk analytic and scientific concern; the matter connects to the question

of how “linear” the dose-response pattern appears at or below the observational ranges of the data (Hoel and Portier, 1994; French and Williams, 2001; Subramaniam *et al.*, 2006). To more fully study the shape of low-dose relationships would require the introduction of biologically-based mode-of-action models and data (Luebeck and Moolgavkar, 1996; Andersen and Dennison, 2001; Clewell, 2005), which extends beyond our scope here. It is an inviting avenue of future research, however.

Along similar lines, an important consideration for any dose-response model is how and if the design of the study affects statistical estimation and inference. Although we did not make this issue a major component of our study, we did notice that certain dose-response patterns were difficult to fit with the multistage model, particularly if large increases in response occurred between two widely spaced doses. This leads us to speculate that a design containing more doses could allow for high-order values of k to fit certain laboratory carcinogenicity data sets. Unfortunately, the designs in the IRIS data we recovered were usually limited to 3 or 4 dose levels, and hence were not rich enough to give substantive guidance on this question. We can, however, cross-classify the highest fitting value of k against the number of doses; see Table 3. Intriguingly, we find that IRIS data sets/chemicals with more dose levels fit $k = 2$ less often than those with only 3 or 4 levels, while the $k = 1$ model is fit most often by a variety of designs. (Recall that no data sets fit $k = 3$ well.) Clearly, the issue of optimal study design for multistage dose-response models is an open, and important, problem in laboratory animal testing; while past work on the issue has been instructive (Portier and Hoel, 1984; Portier and Edler, 1990; Kitsos, 1999), more needs to be done.

Acknowledgements

Portions of this research were conducted while all the authors were with the University of South Carolina. Thanks are due Drs. Ralph L. Kodell, Edsel A. Peña, and Brooke E. Buckley for their input during the research effort, and three anonymous reviewers for their helpful suggestions on the manuscript. This work was funded under grant #R01-CA76031 from the U.S. National Cancer Institute, STAR grant #RD-83241901 from the U.S. Environmental Protection Agency, and as part of the research arm of the U.S. Department of Homeland Security’s Center of Excellence for the Study of Terrorism and Responses to Terrorism (START). Its contents are solely the responsibility of the authors and do not necessarily reflect the official views of these various agencies.

References

- Ahn, H.; Kodell, RL. Analysis of long-term carcinogenicity studies. In: Chow, S-C.; Liu, J-P., editors. *In Design and Analysis of Animal Studies in Pharmaceutical Development*. New York: Marcel Dekker; 1998. p. 259-289.
- Andersen ME, Dennison JE. Mode of action and tissue dosimetry in current and future risk assessments. *The Science of the Total Environment* 2001;274:3–14. [PubMed: 11453304]
- Anderson E. The Carcinogen Assessment Group of the U.S. EPA. Quantitative approaches in use to assess cancer risk. *Risk Analysis* 1983;3:277–295.
- Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer* 1954;8:1–12. [PubMed: 13172380]
- Bailer AJ, Oris JT. Modeling reproductive toxicity in *Ceriodaphnia* tests. *Environmental Toxicology and Chemistry* 1993;12:787–791.
- Bailer AJ, Portier CJ. An illustration of dangers of ignoring survival differences in carcinogenicity data. *Journal of Applied Toxicology* 1988;8:185–189. [PubMed: 3171079]
- Bailer AJ, Smith RJ. Estimating upper confidence limits for extra risk in quantal multistage models. *Risk Analysis* 1994;14:1001–1010. [PubMed: 7846307]
- Chen JJ, Gaylor DW. Dose-response modeling of quantitative response for risk assessment. *Communications in Statistics. Theory and Methods* 1992;21:2367–2381.
- Chiu WA, Hassenzuhl DM, Kammen DM. A comparison of regulatory implications of traditional and exact two-stage dose-response models. *Risk Analysis* 1999;19:15–22.

- Clewell HJ. Use of mode of action in risk assessment: Past, present, and future. *Regulatory Toxicology and Pharmacology* 2005;42:3–14. [PubMed: 15896438]
- Crump KS, Hoel DG, Langley CH, Peto R. Fundamental carcinogenic processes and their implications for low dose risk assessment. *Cancer Research* 1976;36:2973–2979. [PubMed: 975067]
- Dinse GE. A comparison of tumor incidence analyses applicable in single-sacrifice animal experiments. *Statistics in Medicine* 1994;13:689–708. [PubMed: 8023043]
- Edler L, Kopp-Schneider A. Statistical models for low dose exposure. *Mutation Research* 1998;405:227–236. [PubMed: 9748596]
- Edler, L.; Kopp-Schneider, A.; Heinzl, H. Dose-response modeling. In: Edler, L.; Kitsos, CP., editors. *Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment*. Chichester: John Wiley & Sons; 2005. p. 5-23.
- French JL, Williams PL. Characterizing dose-response relationships in multiple cancer bioassays. *Risk Analysis* 2001;21:91–102. [PubMed: 11332555]
- Gaylor DW, Slikker WL. Risk assessment for neurotoxic effects. *NeuroToxicology* 1990;11:211–218. [PubMed: 1700346]
- Guess HA, Crump KS. Low-dose extrapolation of data from animal carcinogenicity experiments — Analysis of a new statistical technique. *Mathematical Biosciences* 1976;32:15–36.
- Haseman JK. Use of statistical decision rules for evaluating laboratory animal carcinogenicity studies. *Fundamental and Applied Toxicology* 1990;14:637–648. [PubMed: 2193843]
- Hoel DG, Portier CJ. Nonlinearity of dose-response functions for carcinogenicity. *Environmental Health Perspectives* 1994;102(Suppl 1):109–113. [PubMed: 8187698]
- Kitsos CP. Optimal designs for estimating the percentiles of the risk in multistage models of carcinogenesis. *Biometrical Journal* 1999;41:33–43.
- Kodell, RL.; Park, CN. Linear extrapolation in cancer risk assessment. In: Olin, S.; Farland, W.; Park, C.; Rhomborg, L.; Scheuplein, R.; Starr, T.; Wilson, J., editors. *Low-Dose Extrapolation of Cancer Risks: Issues and Perspectives*. Washington, DC: ILSI Press; 1995. p. 87-104.
- Kodell RL, West RW. Upper confidence intervals on excess risk for quantitative responses. *Risk Analysis* 1993;13:177–182. [PubMed: 8502790]
- Krewski, D.; van Ryzin, J. Dose response models for quantal response toxicity data. In: Csörgö, M.; Dawson, DA.; Rao, JNK.; Saleh, AKME., editors. *Statistics and Related Topics*. Amsterdam: North-Holland; 1981. p. 201-231.
- Lehmann, EL.; Casella, G. *Theory of Point Estimation*. 2. New York: Springer-Verlag; 1998.
- Lovell DP, Thomas G. Quantitative risk assessment and the limitations of the linearized multistage model. *Human & Experimental Toxicology* 1996;15:87–110. [PubMed: 8645508]
- Luebeck, EG.; Moolgavkar, SH. Biologically based cancer modeling. *Toxicology and Risk Assessment*. In: Fan, AM.; Chang, LW., editors. *Principles, Methods, and Applications*. New York: Marcel Dekker; 1996. p. 533-555.
- Mills A. IRIS from the inside. *Risk Analysis* 2006;26:1409–1410. [PubMed: 17184384]
- Moolgavkar, SH. Stochastic models for estimation and prediction of cancer risk. In: Barnett, V.; Stein, A.; Turkman, KF., editors. *Statistics for the Environment 4: Statistical Aspects of Health and the Environment*. Chichester: John Wiley & Sons; 1999. p. 237-257.
- Moolgavkar SH. Commentary: Fifty years of the multistage model: remarks on a landmark paper. *International Journal of Epidemiology* 2004;33:1182–1183. [PubMed: 15319398]
- Nitcheva DK, Piegorsch WW, West RW, Kodell RL. Multiplicity-adjusted inferences in risk assessment: Benchmark analysis with quantal response data. *Biometrics* 2005;61:277–286. [PubMed: 15737104]
- Piegorsch, WW.; Bailer, AJ. *Statistics for Environmental Biology and Toxicology*. Boca Raton, FL: Chapman & Hall/CRC; 1997.
- Piegorsch WW, Nitcheva DK, West RW. Excess risk estimation under multistage model misspecification. *Journal of Statistical Computation and Simulation* 2006;76:423–430.
- Portier CJ, Edler L. Two-stage models of carcinogenesis, classification of agents, and design of experiments. *Fundamental and Applied Toxicology* 1990;14:444–460. [PubMed: 2340975]
- Portier CJ, Hoel DG. Design of animal carcinogenicity studies for goodness-of-fit of multistage models. *Fundamental and Applied Toxicology* 1984;4:949–959. [PubMed: 6519375]

- Ritter G, Wilson R, Pompei F, Burmistrov D. The multistage model of cancer development: some implications. *Toxicology and Industrial Health* 2003;19:125–145. [PubMed: 15747774]
- Sand S, Filipsson AF, Victorin K. Evaluation of the benchmark dose method for dichotomous data: Model dependence and model selection. *Regulatory Toxicology and Pharmacology* 2002;36:184–197. [PubMed: 12460753]
- Sherman, C.; Portier, CJ. Multistage carcinogenesis models. In: Armitage, P.; Colton, T., editors. *Encyclopedia of Biostatistics*. 4. New York: John Wiley & Sons; 1998. p. 2808-2814.
- Subramaniam RP, White P, Coglian VJ. Comparison of cancer slope factors using different statistical approaches. *Risk Analysis* 2006;26:825–830. [PubMed: 16834636]
- U.S. EPA. Integrated Risk Information System (IRIS). National Center for Environmental Assessment, U.S. Environmental Protection Agency; Cincinnati, OH: 1997. www.epa.gov/iris/www.epa.gov/iris

Appendix A: Semi-parametric bootstrap testing

An alternative to large-sample hypothesis testing exists by employing the computer and applying *bootstrap resampling* (Piegorsch and Bailer, 1997, Sec. 2.5). The method appeals to repeated resampling from the original sample in order to construct a reliable estimate of the distributional properties of point estimators such as the MLE. To do so, suppose we denote by $\{Y_{1j}^*, \dots, Y_{nj}^*\}$ for $j = 1, \dots, J$, a sequence of J independent bootstrap resamples taken randomly with replacement from the observed data. Y_{ij}^* represents a computer-generated, pseudo-binomial random variable with parameters N_i and Y_i/N_i , where Y_i/N_i is the observed proportion of subjects at dose d_i that “respond” and N_i is the number of animals in the i th group. If the observed proportion is zero or one, it is replaced with the estimated risk, $\hat{\pi}(d_i)$, from the fitted model. Thus one might call this a form of “semi-parametric” bootstrapping.

For each such generated bootstrap resample, we fit the model in (1) and compute the constrained (non-negative) MLE of β_k . The resulting collection of J bootstrapped ML estimates represents an approximate, small-sample sampling distribution of the MLE; the larger the value taken for J , the more accurate the approximate bootstrap distribution. (We recommend $J = 2000$ and we use this value in all our calculations.) If the null hypothesis, $H_0: \beta_k = 0$, is true then we expect to find a large number of exact 0 estimates among these bootstrap estimates. If it is false, there should be a smaller number of estimates that are exactly 0 with many more estimates greater than 0. Therefore, we reject the null, when this fraction of estimates equal to 0 is very small (less than 0.05). Finally, we compute a 95% bootstrap lower confidence limit for β_k using the percentile method: this is simply the 5th percentile of the estimated bootstrap distribution. If 0 is the lower 95% confidence limit for β_k , then we fail to reject the null, whereas if the 95% lower confidence limit is greater than 0, we reject the null in favor of (2). Inverting this lower confidence limit to form a hypothesis test, the bootstrap P -value for testing H_0 against the one-sided alternative in (2) is the proportion of bootstrap estimates equal to 0. We reject H_0 if this bootstrap P -value is less than or equal to 0.05 which corresponds to a 95% lower confidence limit on β_k that is greater than 0.

Appendix B: Monte Carlo simulations to compare Wald tests with bootstrap tests

To examine the small-sample properties of the Wald and bootstrap tests of (2) under a multistage dose-response model, we generated a series of binomial data sets via Monte Carlo methods. Eight multistage dose-response models with $k = 1$ were employed in our simulation study. The models were based on parameterizations used in previous studies of low-dose risk estimation described by Bailer and Smith (1994) and Kodell and Park (1995): the background risk, $\pi(0)$, used in the models ranged between 1 and 30%, and the corresponding risk at the highest dose ranged between 2.8 and 57%. Table B.1 summarizes the model parameterizations employed in the simulations.

The dose-response design in our Monte Carlo simulations consisted of four dose levels, $d = 0, 0.25, 0.5, 1.0$ with equal numbers of subjects, $N_i = N$ per dose group. We considered five different values of N , ranging over $N = 25, 50, 100, 300, 500$.

For each parameterization selected for study, we generated 2000 binomial data sets with probability of success $\pi(d)$, where $\pi(d)$ is defined by the two-parameter multistage model given in Equation (1) with β -parameters as specified in Table B.1. Our goal was to assess the false positive rate associated with rejecting the null hypotheses in (2) at $k = 2$. As such, for each Monte Carlo-generated data set we fit the $k = 2$ multistage model and calculated the Wald test statistic z_{Wald} and its corresponding large-sample P -value, and also a P -value based on bootstrap resampling as described in Appendix A. The level of significance was set to $\alpha = 0.05$. We counted the number of times (out of the 2000 simulated data sets) either test rejected the null hypothesis in (2), i.e., how often its $P \leq 0.05$. These proportions are estimates of the false positive error rate of each test. Notice that with 2000 simulated data sets, the standard error of the estimated error rate at $\alpha = 0.05$ is $\sqrt{(0.05)(0.95)/2000} \approx 0.005$, and it never exceeds $\sqrt{(0.5)(0.5)/2000} \approx 0.011$.

Our results gave empirical false positive rates for the Wald test that were highly unstable, and even in large samples could exceed the nominal 5% level; see Table B.2. The instabilities often appeared at small samples sizes when the underlying probabilities of response were relatively small, but this pattern reversed itself when the underlying probabilities of response were relatively large. As suggested in §2.3, it is likely that these gross instabilities occur because of the null parameter value's existence on the border of the model's parameter space. That is, when testing whether a specific β_j is zero we are operating under that null assumption, so in Table B.2 we are operating with $\beta_2 = 0$. But, recall from §2.3 that to validate the asymptotic features of the MLEs we require that all the β -parameters be strictly positive. Thus when $\beta_2 = 0$ we are in effect violating a necessary operating assumption; the disturbing instability evidenced in Table B.2 is a likely consequence of this violation.

Contrastingly, and as we had expected, empirical false positive rates achieved under a bootstrap-based test were much more stable. For small underlying probabilities of response, the error rates were generally too low, but did stabilize somewhat as N grew. For larger underlying probabilities of response, the error rates were generally close to the nominal 5% level. These results lead us to warn against use of the Wald-type test in this setting, and to recommend instead the bootstrap-based methodology.

Table 1
Chemical substances from the U.S. EPA IRIS database and recovered dose-response data used in our study.

IRIS Chemical Name Tumor type; test animal	CAS No.	Dose (units given at first listed dose)	Tumor Incidence	IRIS web link
Acephate Hepatic adenomas/carcinomas; female CD-1 mice	CASRN 30560-19-1 0 ppm 1/62	50 3/61	250 0/62	http://www.epa.gov/iris/subst/0354.htm 1000 15/61
Acetaldehyde Nasal squamous cell carcinoma or adenocarcinoma; male SPF Wistar rats	CASRN 75-07-0 0 ppm 1/94	750 20/95	1500 49/95	http://www.epa.gov/iris/subst/0290.htm 1540 47/92
Acrylamide CNS/mammary and thyroid glands/uterus/oral cavity tumors; female F344/N rats	CASRN 79-06-1 0 mg/kg/day 13/60	0.01 18/60	0.1 14/60	http://www.epa.gov/iris/subst/0286.htm 0.5 21/60
Acrylonitrile Brain and spinal cord astrocytomas, Zymbal gland carcinomas and stomach papillomas/carcinomas; male Spartan Sprague-Dawley rats	CASRN 107-13-1 0 mg/kg/day 6/100	0.09 36/98	7.98 36/98	http://www.epa.gov/iris/subst/0206.htm 0.81 7/100
Brain and spinal cord astrocytomas, Zymbal gland carcinomas, stomach papillomas/carcinomas; male F344/N rats	0 mg/kg/day 5/200	0.11 4/100	0.25 5/100	2.49 20/100
Brain and spinal cord astrocytomas, Zymbal gland carcinomas, stomach papillomas/carcinomas; male Sprague-Dawley rats	0 mg/kg/day 4/80	3.42 18/47	8.53 36/48	21.18 45/48
Aldrin Hepatic carcinoma; male B6C3F ₁ mice	CASRN 309-00-2 0 ppm 3/20	4 16/49	8 25/45	http://www.epa.gov/iris/subst/0130.htm 2000 31/90
Aniline Spleen, combined fibrosarcoma, stromal sarcoma, capsular sarcoma and hemangiosarcoma; male CD-F rats	CASRN 62-53-3 0 ppm 0/64	200 0/90	600 1/90	http://www.epa.gov/iris/subst/0350.htm 20 30
Benz[a]pyrene Forestomach, squamous cell papillomas and carcinomas; CFW mice	CASRN 50-32-8 0 ppm 0/289 45	1 0/25 50	10 0/24 100	http://www.epa.gov/iris/subst/0136.htm 20 1/23 250 66/73
Benzotrithloride Lung adenocarcinoma; female ICR mice	CASRN 98-07-7 0 mg/kg/day 0/35	0.131 1/37	0.52 9/38	http://www.epa.gov/iris/subst/0388.htm 2.1 16/40
Benzyl chloride Thyroid, C-cell adenoma/carcinoma; female F344/N rats	CASRN 100-44-7 0 mg/kg, 3 times/week 4/52	15 8/51	15 8/51	http://www.epa.gov/iris/subst/0393.htm 30 14/52
Bis(chloromethyl)ether Respiratory tract tumors; male Sprague-Dawley rats	CASRN 542-88-1 0 (# 0.1 ppm exposures) 0/240 80	10 1/41	10 1/41	http://www.epa.gov/iris/subst/0375.htm 20 3/46
Bromate Testicular mesothelioma; male F344/N rats	CASRN 15541-45-4 15/34 12/20	100 12/20	60 4/18	http://www.epa.gov/iris/subst/1002.htm

Regul Toxicol Pharmacol. Author manuscript; available in PMC 2008 July 1.

IRIS Chemical Name Tumor type; test animal	CAS No. Dose (units given at first listed dose)	Tumor Incidence	IRIS web link	
p,p'-Dichlorodiphenyldichloroethylene Hepatocellular carcinomas; female B6C3F ₁ mice	18/50 CASRN 72-55-9 0 ppm 0/19	11/33 148 19/47	30/35 http://www.epa.gov/iris/subst/0328.htm	11/11
Hepatocellular carcinomas; female B6C3F ₁ mice	0 ppm 0/19	148 7/41		
Hepatomas; female Syrian golden hamsters	0 ppm 0/31	500 7/30		
Hepatomas; male Syrian golden hamsters	0 ppm 0/42	1000 8/39		
1,2-Dichloroethane Hemangiosarcomas; male Osborne-Mendel rats	CASRN 107-06-2 0 mg/kg/day 0/40	500 4/39	http://www.epa.gov/iris/subst/0149.htm	
Dichloromethane Hepatocellular adenomas/carcinomas; female B6C3F ₁ mice	CASRN 75-09-2 0 mg/kg/day 3/50	47 9/48		
Hepatocellular cancer and neoplastic nodules; male B6C3F ₁ mice	0 mg/kg/day 24/125	95 7/27	http://www.epa.gov/iris/subst/0070.htm	
1,3-Dichloropropene Urinary bladder carcinoma; female B6C3F ₁ mice	CASRN 542-75-6 0 mg/kg/day 3/50	1582 16/48		
Hepatocellular adenomas/carcinomas; male F344/N rats	0 mg/kg/event 0/50	50 8/50		
Hepatocellular adenoma/carcinoma; male F344/N rats	0 mg/kg/event 1/49	25 6/48		
Forestomach papillomas/squamous cell carcinomas; female B6C3F ₁ mice	CASRN 62-73-7 0 ppm 5/49	25 1/50	25 10/49 http://www.epa.gov/iris/subst/0151.htm	
Pancreatic acinar adenoma; male F344/N rats	0 ppm 11/50	280 19/50		
Leukemias; male F344/N rats	0 ppm 16/50	160 30/50		
2,4-Dinitrotoluene mixture Hepatocellular carcinomas/neoplastic nodules and mammary gland adenomas/fibroadenomas/carcinomas; female Sprague-Dawley rats	CASRN N/A 0 ppm 11/23	160 30/50	http://www.epa.gov/iris/subst/0397.htm	
1,2-Diphenylhydrazine Hepatocellular carcinomas/neoplastic nodules; male F344 rats	CASRN 122-66-7 0 % diet 6/95	100 34/35	http://www.epa.gov/iris/subst/0049.htm	

IRIS Chemical Name Tumor type; test animal	CAS No. Dose (units given at first listed dose)	Tumor Incidence	IRIS web link
Epichlorohydrin Forestomach papillomas/carcinomas; male Wistar rats	CASRN 106-89-8 0 ppm 0/10	375 0/9	http://www.epa.gov/iris/subst/0050.htm
Folpet Digestive tract adenomas/adenocarcinomas; male CD1 mice	CASRN 133-07-3 0 ppm 1/104	1000 2/80	http://www.epa.gov/iris/subst/0242.htm
Digestive tract adenomas/adenocarcinomas; female CD1 mice	0 ppm 0/104	5000 8/80	12000 41/80
Fomesafen Hepatic adenomas/carcinomas; male CD-1 mice	CASRN 72178-02-0 0 ppm 30/127	1000 1/80	12000 41/80
Formaldehyde Squamous cell carcinomas; male F344 rats	CASRN 50-00-0 0 ppm 0/156	1 26/63	http://www.epa.gov/iris/subst/0348.htm
Furmecycloz Hepatic nodules/carcinomas; female Sprague-Dawley OF-A rats	CASRN 60568-05-0 0 ppm 4/45	2 0/159	1000 62/65
Heptachlor Hepatocellular carcinomas; male B6C3F ₁ mice	CASRN 76-44-8 0 ppm 5/19	500 10/65	http://www.epa.gov/iris/subst/0362.htm
Heptachlor epoxide Hepatocellular carcinomas; female CD1 mice	CASRN 1024-57-3 0 ppm 6/76	6.1 11/46	http://www.epa.gov/iris/subst/0243.htm
Hepatocellular carcinomas; male CD1 mice	0 ppm 0/62	1 1/70	http://www.epa.gov/iris/subst/0160.htm
Hexachlorobenzene Hepatocellular carcinomas; female Sprague-Dawley rats	CASRN 118-74-1 0 ppm 0/52	1 2/68	10 30/57
Hexachlorobutadiene Renal tubular adenomas/adenocarcinomas; male Sprague-Dawley rats	CASRN 87-68-3 0 mg/kg/day 1/90	75 36/56	10 52/80
alpha-Hexachlorocyclohexane Hepatic nodules/hepatocellular carcinomas; male DD mice	CASRN 319-84-6 0 mg/kg/day 0/20	0.2 0/40	http://www.epa.gov/iris/subst/0058.htm
Technical Hexachlorocyclohexane Hepatic nodules/hepatocellular carcinomas; male Swiss mice	CASRN 608-73-1 0 mg/kg/day 2/22	13.0 0/20	20 9/39
Hexachlorodibenzo-p-dioxin, mixture Hepatic adenomas/carcinomas; male B6C3F ₁ mice	CASRN 19408-74-3 0 ug/kg/wk 15/73	16.25 8/20	65.0 20/20
		32.5 12/12	http://www.epa.gov/iris/subst/0165.htm
		1.25 14/50	http://www.epa.gov/iris/subst/0166.htm
		2.5 14/49	5 24/48

IRIS Chemical Name Tumor type; test animal	CAS No. Dose (units given at first listed dose)	Tumor Incidence	IRIS web link
Hepatic neoplastic nodules/hepatocellular carcinomas; female Osborne-Mendel rats	0 ug/kg/wk 2/75	1.25 5/50	5 18/50 http://www.epa.gov/iris/subst/0167.htm
Hexachloroethane Hepatocellular carcinomas; male B6C3F ₁ mice	CASRN 67-72-1 0 mg/kg/day 3/20	421 15/50	
Hexahydro-1,3,5-trinitro-1,3,5-triazine Hepatocellular adenomas/carcinomas; female B6C3F ₁ mice	CASRN 121-82-4 0 mg/kg/day 1/65	841 31/49	http://www.epa.gov/iris/subst/0313.htm
Hydrazine/Hydrazine sulfate Hepatomas; male CBA/Cb/Se mice	CASRN 302-01-2 0 mg/kg/day 0 ppm 3/30	1.5 5/62	35.0 12/64 http://www.epa.gov/iris/subst/0352.htm
Isophorone Preputial gland carcinoma; male F344/N rats	CASRN 78-59-1 0 mg/kg/day 0/49	0.28 7/25	0.56 12/25 http://www.epa.gov/iris/subst/0063.htm
4,4'-Methylene bis(N,N'-dimethyl)aniline Thyroid follicular cell adenomas/carcinomas; female F344 rats	CASRN 101-61-1 0 mg/kg/day 0 ppm 0/20	357 5/44	http://www.epa.gov/iris/subst/0386.htm
2-Methylnaphthalene Lung adenomas/adenocarcinomas; female B6C3F ₁ mice	CASRN 91-57-6 0 mg/kg/day 5/50	10.64 4/46	http://www.epa.gov/iris/subst/1006.htm
Lung adenomas/adenocarcinomas; male B6C3F ₁ mice	0 mg/kg/day 2/49	50.3 4/49	107.6 6/48
N-Nitrosodiethanolamine Hepatocellular carcinoma/cholangiocellular carcinoma/adenoma/neoplastic nodules; female F344 rats	CASRN 1116-54-7 0 mg/kg/day 1/20	54.3 10/49	113.8 6/49
N-Nitrosodiphenylamine Transitional cell carcinomas (bladder); female F344 rats	CASRN 86-30-6 0 ppm 0/18	0.88 10/39	2.01 14/20
Pentachlorophenol Hepatocellular tumors/hemangiosarcomas; female B6C3F ₁ mice	CASRN 87-86-5 0 mg/kg/day (technical grade) 5/31	50 0/48	200 40/49
Hepatocellular tumors/hemangiosarcomas; female B6C3F ₁ mice	0 mg/kg/day (Dowside EC-7) 1/34		
Polychlorinated biphenyls Hepatocellular adenomas/carcinomas/cholangiocarcinomas; female Sprague-Dawley rats	CASRN 1336-36-3 0 ppm (Aroclor 1260) 1/85	25 10/49	114 42/49
	0 ppm (Aroclor 1254)	25	

IRIS Chemical Name Tumor type; test animal	CAS No. Dose (units given at first listed dose)	Tumor Incidence	IRIS web link
Prochloraz	1/85	19/45	28/49
Hepatocellular adenomas/carcinomas; male CD-1 mice	0 ppm (Aroclor 1242)	50	100
	1/85	11/49	15/45
	0 ppm (Aroclor 1016)	50	100
	1/85	1/48	7/45
	CASRN 67747-09-5		http://www.epa.gov/iris/subst/0378.htm
Hepatocellular adenomas/carcinomas; female CD-1 mice	0 mg/kg/day	3/9	65
	37/92	21/48	44/49
Hepatocellular adenomas/carcinomas; female CD-1 mice	0 mg/kg/day	3/9	65
	5/73	6/39	39/44
Propylene oxide	CASRN 75-56-9	16/25	http://www.epa.gov/iris/subst/0403.htm
Forestomach squamous cell carcinomas; female Sprague-Dawley rats	0 mg/kg/day	11/42	
	0/100	10/28	
Quinoline	CASRN 91-22-5	19/50	http://www.epa.gov/iris/subst/1004.htm
Hepatic hemangioendotheliomas/hemangiosarcomas; male Sprague-Dawley rats	0 mg/kg/day	25	125
	0/6	6/11	18/19
1,1,1,2-Tetrachloroethane	CASRN 630-20-6	50	http://www.epa.gov/iris/subst/0265.htm
Hepatocellular adenomas/carcinomas; female B6C3F ₁ mice	0 mg/kg/day	500	
	5/49	30/48	
1,1,2,2-Tetrachloroethane	CASRN 79-34-5		http://www.epa.gov/iris/subst/0193.htm
Hepatocellular carcinomas; female B6C3F ₁ mice	0 mg/kg/day	174	
	0/20	43/47	
Toxaphene	CASRN 8001-35-2		http://www.epa.gov/iris/subst/0346.htm
Hepatocellular carcinomas and neoplastic nodules; male B6C3F ₁ mice	0 mg/kg/day	2/6	6/5
	10/53	12/53	18/51
Tributyltin oxide	CASRN 56-35-9		http://www.epa.gov/iris/subst/0349.htm
Pituitary tumors; female Wistar rats	0 ppm	5/0	50
	22/50	22/50	35/50
Pituitary tumors; male Wistar rats	0 ppm	5/0	50
	34/50	29/50	43/50
Pheochromocytomas; female Wistar rats	3/50	3/50	34/50
Pheochromocytomas; male Wistar rats	0 ppm	5/0	50
	16/50	14/50	33/50
Parathyroid adenomas; female Wistar rats	0 ppm	5/0	50
	0/64	1/40	1/44
Parathyroid adenomas; male Wistar rats	0 ppm	5/0	50
	0/39	1/51	6/43
1,1,2-Trichloroethane	CASRN 79-00-5		http://www.epa.gov/iris/subst/0198.htm
Hepatocellular carcinomas; male B6C3F ₁ mice	0 mg/kg/day	139	
	2/20	18/49	
2,4,6-Trichlorophenol	CASRN 88-06-2	279	http://www.epa.gov/iris/subst/0122.htm
		37/49	

IRIS Chemical Name type; test animal	CAS No. Dose (units given at first listed dose)	Tumor Incidence	IRIS web link
Leukemia; male F344 rats	0 mg/kg/day 3/20	258 23/50	544 29/50
Trifluralin	CASRN 1582-09-8		
Renal pelvic carcinomas, urinary bladder papillomas, thyroid adenomas/carcinomas (combined); male F344 rats	0 ppm 5/60	813 5/60	3250 9/60
2,4,6-Trinitrotoluene	CASRN 118-96-7		
Urinary bladder/transitional cell papillomas/transitional squamous cell carcinomas; female F344 rats	0 ppm 0/54	0.4 0/54	2.0 0/55
Vinyl chloride	CASRN 75-01-4		
Hepatic angiosarcomas/neoplastic nodules/hepatocellular carcinomas; female Wistar rats	0 mg/kg/day 2/57	1.7 28/58	5.0 49/59
			14.1 56/57

Table 2

Summary of results for testing the hypotheses $H_0: \beta_k = 0$ vs. $H_A: \beta_k > 0$ from (2) at available levels of $k \geq 1$ as applied to the 91 IRIS data sets from Table 1.

k	Data sets identified for testing [*]	Data sets where the multistage fit was adequate at that k	Data sets where H_0 was rejected
3	14	12	0 (0%)
2	52	44	10 (23%)
2	91 ^{**}	68	14 (21%)
1	91	67	64 (96%)

* Requires design to produce proportion data from $k - 1$ separate, non-zero dose groups.

** Including data sets with only 3 dose groups; see text.

Table 3

Cross-classification between number of doses and highest fitting polynomial order, k , among the 74 data sets that exhibited adequate fit to any multistage model (1) in the IRIS database.

		# doses				
		3	4	5	6	7
k	0	2	1	0	0	0
	1	28	19	7	2	1
	2	4	8	2	0	0

Table B.1

$k = 1$ multistage model parameterizations used in the Monte Carlo evaluations. $\pi(0)$ is the response probability at the control level, and $\pi(1)$ is the response probability at the highest dose.

<i>Model Code</i>	β_0	β_1	$\pi(0)$	$\pi(1)$
A	0.01005	0.01833	0.01	0.028
B	0.01005	0.02150	0.01	0.031
C	0.01005	0.02500	0.01	0.034
D	0.05129	0.01528	0.05	0.064
E	0.10536	0.05630	0.10	0.149
F	0.10536	0.08231	0.10	0.171
G	0.10536	0.22000	0.10	0.278
H	0.35667	0.48648	0.30	0.570

Wald test empirical false positive error rates for testing the hypotheses in (2) at $k = 2$ under a $k = 1$ multistage model. (Model codes correspond to the parameterizations in Table B.1.) Nominal false positive level is $\alpha = 0.05$.

Table B.2

<i>Model Code</i>	<i>N = 25</i>	<i>N = 50</i>	<i>N = 100</i>	<i>N = 300</i>	<i>N = 500</i>
A	0.0000	0.0055	0.0395	0.1425	0.1800
B	0.0000	0.0050	0.0560	0.1420	0.1630
C	0.0000	0.0115	0.0775	0.1565	0.1445
D	0.0045	0.0270	0.0365	0.0670	0.0930
E	0.0345	0.0610	0.0930	0.1395	0.1920
F	0.0565	0.0925	0.1375	0.1700	0.1505
G	0.1215	0.1560	0.1445	0.0560	0.0510
H	0.1560	0.1600	0.1020	0.0520	0.0425

Table B.3

Bootstrap empirical false positive error rates for testing the hypotheses in (2) at $k = 2$ under a $k = 1$ multistage model. (Model codes correspond to the parameterizations in Table B.1.) Nominal false positive level is $\alpha = 0.05$.

<i>Model Code</i>	<i>N = 25</i>	<i>N = 50</i>	<i>N = 100</i>	<i>N = 300</i>	<i>N = 500</i>
A	0.0000	0.0000	0.0030	0.0375	0.0470
B	0.0000	0.0005	0.0075	0.0350	0.0445
C	0.0000	0.0005	0.0060	0.0350	0.0460
D	0.0010	0.0045	0.0065	0.0105	0.0150
E	0.0125	0.0145	0.0175	0.0360	0.0505
F	0.0130	0.0210	0.0360	0.0490	0.0440
G	0.0275	0.0525	0.0545	0.0555	0.0510
H	0.0515	0.0525	0.0490	0.0520	0.0435