Published in final edited form as:

Environmetrics. 2013 May 1; 24(3): 143–157. doi:10.1002/env.2201.

Information-theoretic model-averaged benchmark dose analysis in environmental risk assessment

Walter W. Piegorsch^{a,*}, Lingling An^{a,b,*}, Alissa A. Wickens^a, R. Webster West^c, Edsel A. Peña^d, and Wensong Wu^e

^aInterdisciplinary Program in Statistics, University of Arizona, Tucson, AZ, USA

^bDepartment of Agricultural and Biosystems Engineering, University of Arizona, Tucson, AZ, USA

^cDepartment of Statistics, North Carolina State University, Raleigh, NC, USA

^dDepartment of Statistics, University of South Carolina, Columbia, SC, USA

^eDepartment of Mathematics and Statistics, Florida International University, Miami, FL, USA

Abstract

An important objective in environmental risk assessment is estimation of minimum exposure levels, called Benchmark Doses (BMDs), that induce a pre-specified Benchmark Response (BMR) in a dose-response experiment. In such settings, representations of the risk are traditionally based on a specified parametric model. It is a well-known concern, however, that existing parametric estimation techniques are sensitive to the form employed for modeling the dose response. If the chosen parametric model is in fact misspecified, this can lead to inaccurate low-dose inferences. Indeed, avoiding the impact of model selection was one early motivating issue behind development of the BMD technology. Here, we apply a frequentist model averaging approach for estimating benchmark doses, based on information-theoretic weights. We explore how the strategy can be used to build one-sided lower confidence limits on the BMD, and we study the confidence limits' small-sample properties via a simulation study. An example from environmental carcinogenicity testing illustrates the calculations. It is seen that application of this information-theoretic, model averaging methodology to benchmark analysis can improve environmental health planning and risk regulation when dealing with low-level exposures to hazardous agents.

Keywords

Akaike information criterion (AIC); dose-response modeling; frequentist model averaging; model uncertainty; multi-model inference

1. Introduction and Background

1.1. Benchmark analysis

An important concern in quantitative risk assessment is the characterization and estimation of adverse effects after exposures to hazardous environmental stimuli. Analysts employ statistical methods to assess the risks of such exposures, where a primary objective is quantitative characterization of the severity and likelihood of damage to humans or to the environment caused by the hazardous agents (Stern, 2008). Risk is quantified via a function, $R(x)$, which is typically the probability of exhibiting the adverse effect in a subject exposed

Correspondence to: W.W. Piegorsch, BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA. piegorsch@math.arizona.edu.

*Walter W. Piegorsch and Lingling An contributed equally to the production of this manuscript.

to a particular dose or exposure level, x , of the hazardous agent. A major effort in risk analysis involves dose-response modeling, i.e., modeling of the risk function $R(x)$. Commonly, the observations are in the form of proportions associated with each x . This is the *quantal response* setting, and it is often encountered in toxicity analysis, carcinogenicity testing, and many other environmental/ecological risk studies (Piegorsch, 2012).

When conducting risk/safety studies that generate dose-response data, a popular statistical technique is *Benchmark Analysis*. As introduced by Crump (1984), the strategy manipulates components of $R(x)$ to yield a *benchmark dose* (BMD) of the agent at which a specified *benchmark risk* or *benchmark response* (BMR) is attained. If the exposure is measured as a concentration, one refers to the exposure point as a *benchmark concentration* (BMC). The BMD or BMC is used to arrive at a level of acceptable human or ecological exposure to the agent or to otherwise establish low-exposure guidelines, often after application of *uncertainty factors* to account for cross-species extrapolations or other ambiguities in the risk estimation process (Piegorsch and Bailer, 2005, §4.4.1). Risk assessors increasingly employ benchmark quantities as the basis for setting exposure limits or other so-called ‘points of departure’ (PODs) when assessing hazardous environmental stimuli (Kodell, 2005). Indeed, both the United States and the Organisation for Economic Co-operation and Development (OECD) provide guidance on BMDs in carcinogen risk assessment (OECD, 2008; U.S. EPA, 2005), and use of BMDs or BMCs is growing for risk management with a variety of toxicological endpoints (European Union, 2003; OECD, 2006; U.S. General Accounting Office, 2001). One critical enhancement is the use of $100(1 - \alpha)\%$ lower confidence limits on the BMD—called benchmark dose (lower) limits or simply BMDLs (Crump, 1995)—to account for statistical variability in the point estimator, BMD, calculated from the dose-response data.

In many environmental and public health scenarios, $R(x)$ is further refined into a form of excess risk, such as the *extra risk* $R_E(x) = \{R(x) - R(0)\} / \{1 - R(0)\}$. This adjusts for background or spontaneous effects out of the control of the risk regulator (Piegorsch and Bailer, 2005, §4.2). The BMD is then determined by setting $R_E(x) = \text{BMR}$ and solving for x . When applied to data, this is a form of inverse nonlinear regression, similar to estimation of an ‘effective dose’ such as the familiar median effective dose, ED_{50} , in toxicity testing (Piegorsch and Bailer, 2005, §4.1). The BMR is specified in advance of any data acquisition, usually in the range $0.01 \leq \text{BMR} \leq 0.10$ (U.S. EPA, 2012). [Indeed, Crump (1984) originally suggested this range to help reflect situations where the BMD is less-sensitive to the choice for $R(x)$.] Where needed for clarity, the notation adds a subscript for the BMR level at which each quantity is calculated: $\text{BMD}_{100\text{BMR}}$, $\text{BMDL}_{100\text{BMR}}$, etc.

1.2. Model averaging

A fundamental concern in benchmark analysis is potential sensitivity to/uncertainty with specification of the dose-response function, $R(x)$. Among the wide variety of parametric forms that have been tendered for $R(x)$, many operate well at (higher) doses near the range of the observed quantal outcomes; however, these different models can produce wildly different BMDs at very small levels of risk (Faustman and Bartell, 1997; Kang et al., 2000). Needed is a modern statistical methodology that can produce reliable inferences on benchmark exposure levels to the hazard under study, but that can avoid estimation biases and instabilities resulting from uncertain model specification. Towards this end, recent research has appeared that employs model averaging techniques to mitigate possible model uncertainty. Model averaging is perhaps most popular when applied from a Bayesian perspective (Hoeting et al., 1999), and a number of articles have applied some form of Bayesian model averaging (BMA) to the problem of BMD estimation (Bailer et al., 2005; Dankovic et al., 2007; Morales et al., 2006). Alternatively, the frequentist model averaging (FMA) perspective (Liang et al., 2011; Wang et al., 2009) also has garnered attention, and a

selection of works have applied FMA to BMD estimation (Faes et al., 2007; Moon et al., 2005; Wheeler and Bailer, 2007; Wheeler and Bailer, 2009). Herein we expand upon this latter avenue of research: we exploit a model averaging technique proposed by Buckland et al. (1997) and Burnham and Anderson (2002, §4.3) to produce model-robust BMDLs for use in characterizing and managing adverse effects of environmental agents. We focus on applications with quantal-response data; Section 2 reviews the basic modeling and (frequentist) estimation framework under this paradigm. Section 3 describes the Buckland et al. FMA estimators for the BMD and the associated BMDLs. Section 4 reports on a Monte Carlo study of the BMDL's operating characteristics, and Section 5 illustrates the calculations with an example from environmental carcinogenicity testing. Section 6 ends with a short discussion.

2. Benchmark Analysis with Quantal Data

2.1. Quantal response analysis

The benchmark method is often employed with data in the form of proportions, Y_i/N_i , and it is this 'quantal' sampling paradigm upon which we focus. The proportion numerators are taken as independent binomial variates $Y_i \sim \text{Bin}(N_i, R(x_i))$ at each exposure or dose index i , where the proportion denominator N_i is the number of subjects tested, and $R(x_i)$ is used to model the unknown probability that an individual subject will respond at dose $x_i > 0$; $i = 1, \dots, m$.

Typically, $R(x)$ is assigned a parametric specification, e.g., the ubiquitous logistic dose-response model $R(x) = 1/(1 + \exp\{-\theta_0 - \theta_1 x\})$, or the similar probit model $R(x) = (\theta_0 + \theta_1 x)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function (c.d.f.). The unknown θ -parameters are estimated from the data; maximum likelihood is a favored approach (Piegorsch and Bailer, 2005, §A.4.3). The maximum likelihood estimate (MLE) for the BMD is found by setting the estimated extra risk function, $R_E(x)$, equal to the chosen BMR and solving for x . We denote this as BMD or also $\text{BMD}_{100\text{BMR}}$ if further specificity is required.

The corresponding BMDL is built from the statistical properties of the BMD; e.g., a one-sided Wald lower limit employs the asymptotic properties of the MLE (Casella and Berger, 2002, §10.4) to yield

$$\text{BMDL}_{100\text{BMR}} = \hat{\text{BMD}}_{100\text{BMR}} - z_{\alpha} \text{se}[\hat{\text{BMD}}_{100\text{BMR}}], \quad (2.1)$$

where $z = -1$ ($1 - \alpha$) and $\text{se}[\text{BMD}_{100\text{BMR}}]$ is the large-sample standard error of the MLE. Other possibilities for the confidence limit include profile likelihood constructs (Crump and Howe, 1985; Moerbeek et al., 2004) or appeal to the bootstrap (see below).

2.2. Dose-response modeling and estimation

As noted above, a wide variety of possible forms is available for modeling the dose-response function, $R(x)$. In environmental toxicology and carcinogenicity testing, popular models generally correspond to functions available from the US EPA's BMDS software program for performing BMD calculations (Davis et al., 2012). Table 1 provides a selection of eight such dose-response models that we most often see in practice, along with their BMDs for a given BMR (0,1). These are taken from a collection of dose-response models employed by Wheeler and Bailer (2007) in their earlier study of FMA estimators for the BMD. (Wheeler and Bailer did not include the log-logistic model, M6, in all their FMA calculations, although they did present it as a possible generating model. They also presented a possible data generating model based on the gamma distribution c.d.f., but did not use it in their FMA calculations. In a similar vein, we do not consider the gamma model here.) Notice that

certain models impose constraints on selected parameters; these are listed in Table 1 to correspond with the most common constraints seen in the environmental toxicology literature.

To fit any model in Table 1 we find MLEs of the unknown parameters, which we denote here as the generic parameter vector β ; e.g., with the log-probit model (M7), $\beta = [\beta_0 \ \beta_1 \ \beta_2]^T$. With this, we also expand our notation for the risk function to $R(x; \beta)$. The log-likelihood to be maximized is $\ell(\beta) = \sum_{i=1}^m \{Y_i \log[R(x_i; \beta)] + (N_i - Y_i) \log[1 - R(x_i; \beta)]\}$, up to a constant not dependent upon β . For models M3–M8, this requires constrained optimization techniques. In all cases, closed-form expressions for β are unavailable, unfortunately, and so calculation proceeds via computer iteration. We employ the R programming environment (R Development Core Team, 2012), 64-bit version 2.13.1 on a Windows[®] workstation, using either the standard R function *glm* for models M1 and M2, box-constrained optimization via the R function *optim* for models M3–M5 (Deutsch et al., 2010), or the R package *drc* (Ritz and Streibig, 2005) for models M6–M8.

In all cases the usual regularity conditions hold for the standardized MLEs to approach a Gaussian distribution in large samples (Casella and Berger, 2002, §10.1), although where constraints exist on the elements of β we require that the true values of those constrained parameters lie in the interior of the parameter space. Large-sample standard errors of the β s are built from the inverse of the Fisher information matrix. With these, we find $se[\text{BMD}_{100\text{BMR}}]$ for use in (2.1) via a multivariate Delta-method approximation (Piegorsch and Bailer, 2005, §A.6.2). Appendix 1 details the associated operations for each of the eight models in Table 1.

2.3. Model uncertainty

The models in Table 1 are some of the most popular forms chosen by environmental risk analysts for describing toxicological dose-response patterns; the number of applications in which they appear is larger than can be reasonably reviewed here. Indeed, one or more of these models often can be found to ‘fit’ a given environmental toxicity data set reasonably well, at least as indicated by typical goodness-of-fit tests. Whether the chosen form is actually correct is uncertain, however, and there is growing recognition that this level of model uncertainty is more extensive and pernicious than has previously been imagined with BMD estimation. For instance, we recently studied each of the eight models in Table 1, and found that model misspecification can drastically affect the coverage performance of the Wald-type BMDL in (2.1): for most misspecification pairings—e.g., selecting a logistic model, M1, when the data are actually generated from the similar probit model, M2—severe undercoverage from the nominal 95% coverage level often occurred, and could spiral down to 0% as sample size increased (West et al., 2012). [As expected, when the correct parametric model was employed to build the BMDL, large-sample coverage patterns were stable; cf. Wheeler and Bailer (2008).] We also translated this to the extra risk scale by calculating how far the actual value of $R_E(\text{BMDL})$ varied from the BMR. While some attained extra risks were close to the target level, we found that in extreme cases a misspecified $R_E(\text{BMDL})$ could reach almost eight times above the desired BMR.

Worse still, we found that when employing an information-theoretic (IT) quantity such as the Akaike Information Criterion (AIC; see Akaike, 1973) to first select a model from among the eight candidates, incorrect selection resulted over half the time for most models. The consequent BMDLs constructed under the often-incorrect, AIC-selected model again exhibited unpredictable coverage patterns. Moving to alternative IT measures such as BIC (Schwarz, 1978), ‘corrected’ AICc (Hurvich and Tsai, 1989), or KIC (Cavanaugh, 1999) did not improve matters.

Our message from these results was clear, if troubling: existing practice of either (i) unequivocally employing a specific dose-response model when there is any possibility of it being misspecified, or (ii) using IT-based selection to first choose the dose-response model, can lead to BMDLs that fail, sometimes utterly, to cover the true BMD. Further adjustment or correction for this model uncertainty *must* be employed when calculating BMDLs for use in environmental risk assessment. Towards this end, we describe in the next section an FMA approach based on IT quantities that overcomes the debilitating effects of model uncertainty on BMD estimation and inferences.

3. Frequentist MODEL-AVERAGED BMD ESTIMATION

3.1. IT-weighted model averaging

To address model uncertainty when constructing BMDs and BMDLs from quantal-response data, we consider a FMA technology introduced by Buckland et al. (1997). The strategy estimates an unknown parameter, say, τ , common to a class of competing-but-uncertain models $\mathcal{U}_Q = \{\mathcal{M}_1, \dots, \mathcal{M}_Q\}$ by first estimating τ under each q th model and then

synthesizing the point estimators into a single weighted average, $\bar{\tau} = \sum_{q=1}^Q w_q \hat{\tau}_q$. The weights w_q are chosen to represent the quality or adequacy of the q th candidate model in the uncertainty class $\mathcal{U}_Q$. Most popular are weights built from IT quantities such as the AIC:

$$w_q = \frac{\exp\{-\Delta_q/2\}}{\sum_{q=1}^Q \exp\{-\Delta_q/2\}}, \quad (3.1)$$

where $\Delta_q = \text{AIC}_q - \min\{\text{AIC}_1, \dots, \text{AIC}_Q\}$ and AIC_q is the AIC from the ML fit of the q th model $\mathcal{M}_q$. (We employ the ‘lower-is-better’ form of AIC: $-\Delta_q = -2 \log \ell_q + 2$ where ℓ_q is the maximized log-likelihood and Δ_q is the number of free parameters to be estimated, under model $\mathcal{M}_q$.) The quantities in (3.1) are known as “Akaike weights” (Burnham and Anderson, 2002, §2.9). The AIC estimates the expected, relative, lost Kullback-Leibler (K-L) information under model misspecification. Thus the w_q s in (3.1) quantify the IT-based ‘weight of evidence’ for the K-L quality of each $\mathcal{M}_q$, assuming that the correct model is included in $\mathcal{U}_Q$ (Burnham and Anderson, 2002, §2.9; Faes et al., 2007). If desired, one can modify (3.1) to employ alternative IT quantities such as BIC, KIC, AICc, etc., instead of AIC.

IT-weighted estimation has seen growing acceptance in a variety of estimation settings (Candolo et al., 2003; Fletcher and Dillingham, 2011; Lukacs et al., 2010), including selected applications in risk assessment (Kang et al., 2000; Moon et al., 2005; Namata et al., 2008). This prompts our exploration of it for addressing BMD model uncertainty.

3.2. IT-weighted BMDs and BMDLs

Applied to our setting, an IT estimator for the BMD using weights from (3.1) takes the form

$$\overline{\text{BMD}} = \sum_{q=1}^Q w_q \hat{\text{BMD}}_q, \quad (3.2)$$

where BMD_q is the MLE computed under $\mathcal{M}_q$. [For simplicity, we suppress the BMR subscript throughout this and the next section, although it is understood that all these operations hold for some fixed BMR $\in (0,1)$.] The FMA construction in (3.2) is similar to BMD estimation strategies described by Kang et al. (2000) and Bailer et al. (2005). For purposes of inference, however, construction of a BMDL under this IT-weighted scheme remains an open issue. As a first approach, Kang et al. (2000) and Bailer et al. (2005) averaged the separate BMDL $_q$ s from each individual model fit. While reasonable, it is

unclear if such an averaged quantity possesses valid $100(1 - \alpha)\%$ coverage. Wheeler and Bailer (2007, 2008) applied a slightly different tactic and constructed FMA estimates of the entire dose-response curve, producing in effect $R(x)$ and $R_E(x)$. The latter function was then inverted to estimate BMD for a fixed BMR. Bootstrap operations were employed to find the BMDL. Moon et al. (2013) applied a related FMA strategy for bootstrapped BMDLs in microbial risk assessment. One crucial aim of our approach is to provide an easily-calculated BMDL to enable widespread application. Thus, while advances in computer technology have facilitated implementation of the bootstrap to levels unimagined only a few years ago, the approach nonetheless adds a level of complexity that exceeds our goals here. Thus we will refrain from calls to the bootstrap. We do acknowledge its value, however, and similar to Wheeler and Bailer (2007, 2008) and Moon et al. (2013) we are exploring ways to employ the bootstrap for more complex benchmark calculations. We hope to report on these in a future manuscript.

As a simple, easy-to-calculate alternative, we apply the Wald limit in (2.1), for which we require the asymptotic standard error $se[\overline{\text{BMD}}]$. The individual BMD_q s are likely to be highly correlated, however, since they all employ the same set of data. Coupled with the data-dependency of the w_q s, determination of $se[\overline{\text{BMD}}]$ can hence become unwieldy. Following Buckland *et al.* (1997), we adopt a conservative strategy and assume that the various correlations approach one; this leads to

$$se[\overline{\text{BMD}}] \approx \sum_{q=1}^Q w_q \{var[\hat{\text{BMD}}_q | \mathcal{M}_q] + (\hat{\text{BMD}}_q - \overline{\text{BMD}})^2\}^{1/2} \quad (3.3)$$

where $var[\hat{\text{BMD}}_q | \mathcal{M}_q]$ is the estimated variance of BMD_q assuming $\mathcal{M}_q$ is correct (cf.

Appendix 1), and $(\hat{\text{BMD}}_q - \overline{\text{BMD}})^2$ is an estimate of the misspecification bias under each $\mathcal{M}_q$. [An adjustment for small samples is available that multiplies $var[\hat{\text{BMD}}_q | \mathcal{M}_q]$ by a squared ratio of t - and Normal-distribution critical points (Burnham and Anderson, 2002, §4.3.3), but this essentially collapses to 1 under our asymptotic construction.] The resulting $100(1 - \alpha)\%$ FMA BMDL is $\overline{\text{BMD}} - z_\alpha se[\overline{\text{BMD}}]$, i.e.,

$$\text{BMDL} = \sum_{q=1}^Q w_q \hat{\text{BMD}}_q - z_\alpha \sum_{q=1}^Q w_q \{var[\hat{\text{BMD}}_q | \mathcal{M}_q] + (\hat{\text{BMD}}_q - \overline{\text{BMD}})^2\}^{1/2}. \quad (3.4)$$

For the uncertainty class we employ the collection of models in Table 1. Notice therein that some of the forms nest within others: the two-stage model (M5) clearly reduces to the quantal-linear (M3) when its $\beta_2 = 0$, as does the Weibull model (M8) after appropriate reparameterization when its $\beta_1 = 1$. The two-stage model (M5) also reduces to the quantal-quadratic model (M4) when its $\beta_1 = 0$. (In all cases, these correspond to limiting boundary conditions on the larger model.) This could be problematic: suppose that for a given data set ML estimation reduces a larger model to one of its nested alternatives in the class, *and* that the larger model receives less IT support—which we interpret as having a lower w_q —from the data than the simpler, reduced form. Then, Wheeler and Bailer (2009), following Raftery *et al.* (1997), argue for application of “Occam’s razor” (Britannica Editors, 2002): expunge the weaker model from the averaging process, or equivalently, reset its weight equal to 0. In effect, one excises the larger model when a simpler, nested model explains the data at least as well. We follow their strategy here.

4. Performance evaluations of the IT-weighted, FMA BMDL

4.1. Simulation design

The relative simplicity and model-robustness of the BMDL in (3.4) give this FMA strategy an enticing potential, despite the admitted burden of fitting $Q = 8$ different dose-response models (or whatever size the analyst's uncertainty class is taken to be) for a single set of data. The BMDL's underlying theoretical validity appeals to the asymptotic normality of the MLEs, however, and whether these confidence limits operate nominally in small samples is an important question. To assess this, we studied the performance of (3.4) via a series of Monte Carlo evaluations. We set the BMR to the typical default level of $\text{BMR} = 0.10$ (U.S. EPA, 2012) and operated at 95% nominal coverage. The doses were taken at $m = 4$ levels: $x_1 = 0$, $x_2 = 1/4$, $x_3 = 1/2$, $x_4 = 1$, corresponding to a standard design in cancer risk experimentation (Portier, 1994). Equal numbers of subjects, $N_i = N$, were taken per dose group. We considered four different possibilities for the per-dose sample sizes: $N = 25, 50, 100$, or 1000 ; the latter approximates a 'large-sample' setting, while the former two are more commonly seen in, e.g., environmental carcinogenicity investigations (cf. §5, below). As throughout, all of our calculations were performed within the R programming environment (R Development Core Team, 2012).

For the true dose-response patterns we returned to all eight models in Table 1. Background risks at $x = 0$ were set between 1% and 30%, and the other risk levels were increased to produce a variety of (strictly) increasing forms, ending with high-dose risks at $x = 1$ between 10% and 90%. To set the parameters for each model, we fixed $R(x)$ at $x = 0$ and $x = 1$ and solved for two unknown parameters. This completed the specification for the two-parameter models (M1–M4). For the 3-parameter models (M5–M8), we additionally fixed $R(x)$ at $x = 1/2$, and then solved for the third unknown parameter. The actual specifications and resulting parameter configurations for the various models are given in Table 2.

At each of the $8 \text{ (models)} \times 6 \text{ (configurations)} \times 4 \text{ (sample-sizes)} = 192$ combinations, we simulated at least 2000 pseudo-binomial quantal-response data sets via R's *rbinom* function. Then for each data set, we found the MLEs from every model in Table 1 and calculated the consequent BMD_q and $\text{var}[\text{BMD}_q | \mathcal{M}_q]$ ($q = 1, \dots, 8$), at $\text{BMR} = 0.10$. These were combined into the IT-weighted FMA estimators in (3.2) and (3.3). From these we computed the IT-weighted FMA BMDL in (3.4). We then assessed whether this BMDL was below the true value of BMD for the $\mathcal{M}_q$ under which the data were generated. The resulting empirical coverage rate is the proportion of cases that covered the true BMD out of the 2000 simulated samples. Notice then that in terms of Monte Carlo sampling variability, the standard error of the empirical coverage is approximately $\sqrt{(0.05)(0.95)/2000} = 0.005$ and this never exceeds $\sqrt{(0.5)(0.5)/2000} = 0.011$.

We also explored what specific level of extra risk the IT-weighted BMDLs achieved, since this is the crucial underlying feature targeted by benchmark analysis. That is, given an IT-weighted BMDL from a simulated data set, what is the value of $R_E(\text{BMDL})$ under the true $\mathcal{M}_q$? Since the BMDL is a lower confidence bound, $R_E(\text{BMDL})$ should be slightly less than the target BMR. The fundamental goal of the risk-analytic process here is to closely achieve, and in particular *not* exceed, this BMR. Therefore, how well our IT-weighted BMDL performs in this fashion is critically important. For summary purposes, we captured both the median and the upper quartile of $R_E(\text{BMDL})$ among the 2000 simulated data sets for each parameter configuration.

4.2. Infinite BMDs

We came across an unusual artifact while conducting our Monte Carlo computations: for some very shallow observed dose-response patterns, calculation of BMD could sometimes break down. One obvious case is when the observed dose response is exactly horizontal, i.e., when $Y_1/N_1 = Y_2/N_2 = \dots = Y_m/N_m$. Every $\mathcal{M}_q$ in Table 1 then estimates the dose response as a constant at that observed rate. This produces $R_E(x) = 0$ for all x , from which there is no solution to the BMD-defining relationship $R_E(x) = \text{BMR}$. The data are essentially telling us that the observed dose response cannot attain the BMR, no matter how large x grows. In such instances we simply drove BMD_q to ∞ , or, equivalently, reported it as undefined; Wheeler and Bailer (2007) identified a similar phenomenon in their study of FMA-based BMDs. Note that simulations producing $Y_1 = Y_2 = \dots = Y_m = 0$ were excluded entirely from our analysis, viewing this as lack of any response information upon which the benchmark calculations could operate.

Computationally, when any ‘infinite’/undefined BMD_q was observed, we overrode the IT quantification for the weight and set the corresponding $w_q = 0$. In effect, we viewed such occurrences as indications that fitted model $\mathcal{M}_q$ could not provide any ‘information’ towards the FMA calculation of $\overline{\text{BMD}}$. If BMD_q was ‘infinite’/undefined for $q = 1, \dots, 8$, we defined $\overline{\text{BMD}}$ as ‘infinite’ and recorded this as a failure to cover the true BMD.

This issue of undefined or infinite BMDs was not uncommon with some of the very shallow dose-response configurations in Table 2, especially configurations A and C. Particularly at the smaller sample sizes, these shallow configurations produced flat, and sometimes even non-monotone response patterns in the simulated data, despite the fact that the true underlying $R(x)$ function was strictly increasing. This usually caused no difficulty. In the few instances where the simulated data led to a decreasing effect, however, the fits for models M1 and M2 produced decreasing estimated extra risks, which led to a negative BMD_q . (As per Table 1, we placed no constraints on the M1 or M2 model fits.) When this was observed we reported the BMD as undefined. For the more complex three-parameter models (M5–M8) this sort of instability was more pronounced, and either caused the iterative fit to fail or, for the models constrained to be non-decreasing (M3–M8 when $x > 0$), drove the fit to a flat, horizontal response. In these cases, we again reported BMD_q as undefined.

We note in passing that the popular BMDS software program for benchmark analysis available from the U.S. EPA (Davis et al., 2012) will not estimate a BMD for a flat or negatively trending dose response (Wheeler and Bailer, 2009). Given the complexities we observed when working with such aberrant data sets, this may not be an unreasonable strategy, although it does reduce the options available to the analyst when forced to complete the calculations directly.

4.3. Monte Carlo results

We summarize the empirical coverage results from our Monte Carlo study in a series of figures and tables. Complete descriptions are available in Wickens (2011). Recall that we fix $\text{BMR} = 0.10$ and operate at nominal 95% coverage. Figure 1 displays summary boxplots of the empirical coverages recorded across all model configurations, as a function of per-dose sample size N , when the AIC is used for the weights in (3.1). We find that the median coverages are generally quite stable, centering at conservative levels and moving closer to the nominal level as sample size increases. The high coverage rates are not altogether surprising, since as noted above (3.3) is a conservative approximation for the standard error (Burnham and Anderson, 2002, §4.3). The AIC-based coverage values also do not drop appreciably below nominal, and when this does occur it is almost always within Monte

Carlo sampling variability. Thus in general the AIC-based, FMA BMDL appears quite stable, if slightly conservative at $\alpha = 0.05$.

We saw similar results (not shown) when the AICc was used for the IT weights. This was expected: the number of parameters with our models is much smaller than the total sample sizes studied, and in such instances the AIC and AICc closely approximate each other.

Conversely, when the KIC was used for the IT weights in (3.1), the empirical coverages exhibited greater instabilities, dropping below the nominal 95% level far more often than the AIC- and AICc-based FMA BMDLs. In the extreme, coverage at $\alpha = 0.05$ reached as low as 88% (results not shown). We found this anti-conservative coverage even more prevalent when the BIC was used for the IT weights. Figure 2 displays summary boxplots of the empirical coverages for BIC (note the scale difference vs. Figure 1): while the median coverage rates remain above the nominal 95% level and roughly imitate the AIC's pattern, we see large skew towards under-coverage at all samples sizes, reaching as low as 60% at the highest per-dose sample size.

It is worth noting that in selected cases Wheeler and Bailer (2009, §6) also reported stronger performance of AIC- vs. BIC-based weighting schemes when evaluating coverage of their FMA BMDLs. As a referee has remarked, however, the substandard performance we find with BIC- and KIC-based weights seems puzzling, since we would generally expect model instabilities to settle down as N increases, irrespective of the particular IT weight. Indeed, the ability of each model to fit dose-response data based on $N = 1000$ observations per dose was quite good: no cases of 'infinite' or negative BMD_q values were observed at this sample size. We did find, however, that the BIC- and, to a lesser extent, the KIC-based coverages could fall victim to extreme under-coverage in select cases when $N = 1000$. This was generally with configuration B and occasionally C, and true across all but models M1 and M2. As best we can determine, the under-coverage appears to be driven by extreme positive bias in the estimator: at least for these more-shallow configurations, the BIC-based weights appear to favor models M1, M2 and M4 (especially the latter), even when the correct model is a more-complex form such as M7 or M8. For example, at $N = 1000$ under model M8 with configuration B the empirical coverage using the BIC-based weights was 74.75%, one of the worst we recorded. (With KIC-based weights at this configuration it was 92.65%, far better but still significantly different from the nominal 95% level. With AIC-based weights it was 94.80%.) The corresponding average empirical BIC-based weights among our 2000 simulated samples were $w_{2,\text{BIC}} = 0.332$ and $w_{4,\text{BIC}} = 0.407$ for models M2 and M4, resp., but only $w_{8,\text{BIC}} = 0.030$ for (correct) model M8. Conversely, the corresponding average AIC-based weights were $w_{2,\text{AIC}} = 0.158$, $w_{4,\text{AIC}} = 0.173$, and $w_{8,\text{AIC}} = 0.151$. This may be related to the BIC's well-known propensity to favor models with fewer parameters (Claeskens and Hjort, 2008, §4.1). Adding to the problem, the simpler models M2 and M4 are slightly more convex than model M8 closer to $x = 0$, hence they tend to underestimate small extra risks. This results in overestimation of the BMD. Thus the BIC-based weights drive the FMA estimate to less-flexible, positively biased models, leading to increased positive bias in the point estimate. Since we employ these to produce a one-sided lower limit, the positive bias will act to decrease the coverage probability, especially at larger sample sizes where the standard errors are tighter. The KIC-based weights appear to operate similarly, but to a somewhat reduced degree. (Detailed results are available from the authors.) For BIC and to a lesser extent KIC, we view these instabilities as too severe to recommend use of these measures with (3.1), at least for the class of models we consider here.

For a more-detailed examination of these various effects, we give in Table 3 the specific empirical coverages achieved under each model configuration in Table 1, at the popular per-

dose sample size of $N = 50$. As expected, we see a general trend towards conservative coverage, with many entries in the 98–99% range. Also as expected, coverages based on AIC and AICc weights for (3.1) are essentially identical, although those based on KIC and BIC weights show greater instability. In particular, and corroborating Figure 2, BIC-based coverage weakens into the 88–92% range in a number of instances. This latter effect is isolated primarily to configurations D and F, but exists across a variety of models (M3, M5–M6, and M8). Slightly less unstable are the KIC-based weights, but these nonetheless drop significantly below the nominal 95% level a number of times, at essentially the same settings for which the BIC-based results exhibit sub-nominal coverage. More generally, we see greater variation in coverage performance across different configurations than across different generating models, e.g., the more-shallow configurations (A, B, and C) yield more-conservative coverage. One exception is that to a limited extent, data from the ‘multi-stage’ forms (M3–M5 and their cousin M8) can yield lower coverage than the probit/logit-based forms.

The coverage results in Table 3 compare favorably to analogous Monte Carlo evaluations of bootstrapped FMA BMDLs given by Wheeler and Bailer (2008). Similar to our study, Wheeler and Bailer (i) model averaged using (3.1) with AIC-based weights, (ii) employed the same four-dose design, (iii) operated with $N = 50$ observations/dose level, (iv) set the nominal coverage rate to 95%, and (v) operated at $\text{BMR} = 0.10$. Their Table 4 provided empirical BMDL coverages using a suite of $Q = 7$ models that included all but model M6 (log-logistic) in our Table 1, and under a series of parameter configurations comparable to many of those in our Table 2. They reported many FMA bootstrap-based empirical coverage rates near the nominal 95% level, but also found coverages that rose to near or above 98–99% along with many that dropped below nominal, some at or below 80%. Thus their results shared many similarities to the patterns we find in our Table 3, although we did not experience cases where our FMA empirical coverage sank as low as 80% (likely due to the known conservatism built into our standard error approximation).

We also conducted Monte Carlo coverage evaluations of (3.4) when $\text{BMR} = 0.01$. While less common in practice, this smaller BMR may be employed when sufficient data are available to support inferences at extreme low doses (Kodell, 2009; U.S. EPA, 2012). Our results (not shown) were generally similar to those seen above; in particular, the AIC-weighted and AICc-weighted BMDLs rates again exhibited generally conservative coverage, although greater skew towards undercoverage was evidenced. In fact, the AIC’s and AICc’s patterns of coverage were more similar to that of the KIC-weighted BMDLs, whose own overall coverage exhibited somewhat greater stability than at $\text{BMR} = 0.10$. The BIC remained the most unstable of the four IT-weighting schemes. In general, at $\text{BMR} = 0.01$ we encountered a greater number of empirical coverage rates that dropped significantly below the nominal 95% level, especially with configurations B, C, and, occasionally, D. This is consistent with established benchmarking experience: when response rates are very small at low doses, information in the data may be insufficient to perform effective inferences if the BMR is set closer to zero. We therefore urge caution in practice when employing these methods at $\text{BMR} = 0.01$ with very shallow or low-response patterns.

Figure 3 presents summary boxplots of the median extra risks, $R_E(\text{BMDL})$, achieved by the IT-weighted BMDLs under the true dose-response model and using only the AIC in (3.1). Once again, $\text{BMR} = 0.10$ and $\alpha = 0.05$. Figure 4 give the corresponding boxplots of the upper quartiles. The patterns are predictable: the somewhat-conservative nature of the AIC-based BMDL seen in Figure 1 translates to true extra risks below the BMR (of course, since these are lower confidence limits even exact coverage would translate to true extra risks below the BMR). The conservatism decreases at a fairly rapid rate as N increases, and in no

cases do we encounter median or upper-quartile extra risks above the required level of benchmark response.

In similar summary boxplots for the AICc, KIC, and BIC (not shown), we observed results that corresponded to the patterns of empirical coverage discussed above: extra risks for AICc were essentially similar to AIC, while KIC and BIC extra risks were somewhat less conservative (i.e., closer to the BMR), and in some cases *exceeded* the BMR. Here again, we view this as a highly unstable characteristic, and cannot recommend the BIC, and to a lesser extent the KIC, for use in the IT-weights with FMA benchmark calculations, at least under the models in the uncertainty class studied here.

5. Example: Liver carcinogenesis in laboratory animals

From an environmental toxicity study of the organochlorine fungicide/fumigant pentachlorophenol (PCP), Piegorsch and Bailer (2005, Ex. 4.13) describe data on carcinogenicity of the compound in female B6C3F₁ mice. At human-equivalent exposures of $x_1 = 0$, $x_2 = 1.3$, $x_3 = 2.7$, and $x_4 = 8.7$ mg/kg/day, the mice exhibited the following proportions of hemangiosarcomas/hepatocellular tumors:

$Y_1/N_1 = \frac{1}{34}$, $Y_2/N_2 = \frac{6}{49}$, $Y_3/N_3 = \frac{9}{46}$, and $Y_4/N_4 = \frac{42}{49}$. In analyzing these data for BMD estimation, Piegorsch and Bailer applied a two-stage model (M5) and at BMR = 0.10 reported $BMD_{10} = 1.69$ mg/kg/day. They also discussed other models for performing benchmark analyses with quantal response data—including, in effect, our quantal-linear (M3) and Weibull (M8) models—but did not pursue the issue of model adequacy or model selection with the PCP data. In effect, their analysis is incomplete, since potential model uncertainty when calculating BMD_{10} was left unaddressed. We pick up on that thread here, in order to illustrate the calculations of our IT-weighted BMD and BMDL from §3 and to improve upon their earlier analysis. Based on our Monte Carlo evaluations in §4.3, we consider only the AIC for use in (3.1).

Following Piegorsch and Bailer, we set BMR = 0.10. We include their original two-stage model (M5), the quantal-linear (M3) and Weibull (M8) models they mentioned, and the remaining five models from Table 1 to build our uncertainty class, $\mathcal{U}_8$. Fitting all eight models to the carcinogenicity data produces the results given in Table 4. All models yield valid point estimates, with the individual BMD_q values at BMR = 0.10 ranging between 0.7114 mg/kg/day for the quantal-linear model (M3, which exhibits the highest/worst AIC) and 2.3132 mg/kg/day for the log-probit model (M7, which, interestingly, exhibits the second-highest AIC). Notice also that our calculations for model M5 give essentially the same benchmark dose estimate as reported in Piegorsch and Bailer (2005, Ex. 4.13): $BMD_{M5} = 1.7006$ mg/kg/day from Table 4 vs. $BMD_{M5} = 1.69$ mg/kg/day from Piegorsch and Bailer. (We attribute the slight difference to rounding errors in their original calculation.)

Table 4 also lists the AIC-based weights from (3.1) and intermediate calculations to produce the FMA BMD estimate in (3.2): we find $\overline{BMD}_{10} = 1.9862$ mg/kg/day. As would be expected, this FMA value sits roughly intermediate to the individual-model BMD_q values, but because of its construction it provides a more objective estimate of the actual benchmark dose in the presence of model uncertainty.

To complete the analysis, we find the standard error from (3.3) to be $se[\overline{BMD}_{10}] = 0.4041$. This gives a 95% Wald lower limit as $BMDL_{10} = 1.9862 - (1.645)(0.4041) = 1.3215$ mg/kg/day. This value represents a model-robust benchmark for conducting further risk-analytic calculations on PCP carcinogenicity based on these data. Had the analyst committed to only

one particular model such as the two-stage (M5) used by Piegorsch and Bailer, the corresponding 95% Wald BMDL₁₀ of $1.7006 - (1.645)(1.1144)$ would in fact be negative and of little practical use! (This is, of course, as much a limitation of the Wald-type limit as it is of model M5; alternatives such as profile-likelihood limits—mentioned above—or parameter transformations to ensure positive dose estimates could be applied to avoid BMDLs below zero.) Alternatively, had the analyst chosen the model based on the minimum AIC, which here is the probit (M2), the 95% BMDL_{M2} would be $1.9266 - (1.645)(0.2414) = 1.5295$ mg/kg/day. While slightly larger than our FMA BMDL, this limit (i) is unadjusted for its data-based model selection and (ii) if based on an incorrect model may likely suffer from non-trivial under-coverage/overestimated extra risk, as we identified in West et al. (2012). Corroborating reports by many who have come before, we find that FMA adjustment frees risk assessors from the selection biases, model inadequacies, and inferential uncertainties one encounters when committing to only a single parametric model to perform the benchmark analysis.

6. DISCUSSION

Herein, we consider frequentist model averaging (FMA) for estimating benchmark doses (BMDs) in quantitative risk analysis. Placing emphasis on environmental toxicity assessment, our approach estimates the BMD within an uncertainty class, $\mathcal{U}_Q$, of possible dose-response models, and account for the model uncertainty via FMA calculations. The final estimator is an information-theoretic (IT) weighted average of estimated BMD values across $\mathcal{U}_Q$. A closed-form lower confidence limit (BMDL) using this weighted estimator is derived from a conservative, large-sample approximation for the standard error of the point estimator (Buckland et al., 1997; Burnham and Anderson, 2002), building upon previous FMA explorations for benchmark inference (Faes et al., 2007; Moon et al., 2005; Wheeler and Bailer, 2007; 2009).

Using Monte Carlo evaluations, we find that the BMDL exhibits stable, if sometimes-conservative, coverage when based on the Akaike Information Criterion (AIC) for the IT-based weights, but that other, competitor information measures such as the BIC or the KIC can lead to less stable coverage characteristics. Risk analysts can apply our results to construct model-robust inferences for the BMD that avoid concerns over parametric model uncertainty, expanding past the collection of parametric models seen in practice. This extended operability leads to improved risk analytic decision-making in environmental toxicology testing and other adverse-event risk assessments.

Of course, some caveats and qualifications are in order. We have implicitly assumed that the true dose-response model is one of the $\mathcal{M}_Q$ s included in our uncertainty class, and our Monte Carlo evaluations operated under this assumption. While this is consistent with the IT framework in multi-model inference (Grueber et al., 2011), any FMA estimate will account for model inadequacy only as far as it can incorporate information about the response contained within the uncertainty class from which it is built. Indeed, selection of $\mathcal{U}_Q$ is, perhaps ironically, another source of uncertainty in the FMA process. In response to this concern, one could simply increase the size of the class to capture as many candidate models as possible; however, use of too large an uncertainty class can induce inclusion of spurious models/weightings/inferences, which then decreases precision of, in our case, the $\overline{\text{BMD}}$ and of its FMA BMDL. A wide variety of suggestions—too many to recount here—has been proffered for how to reduce $\mathcal{U}_Q$ in an *objective* fashion, besides, e.g., simple application of Occam's razor (see §3.2 above), to increase FMA precision. How these might be applied to advance benchmark analysis remains open for study; cf. Noble et al. (2009) or Wheeler and Bailer (2009). One could also eschew the modeling process altogether and default to fully

nonparametric estimation for the dose response; however, this has its own set of caveats and qualifications (Piegorsch et al., 2012).

It is also of interest to investigate how our FMA BMDL operates under different design configurations. Our Monte Carlo evaluations in §4 employed a geometric, four-dose design, arguably the quintessential standard in cancer and laboratory-animal toxicology testing. We may gain greater information about the pattern of dose response, however, and therefore about the BMD, if we increase the number of doses and/or change the dose spacings. Can increasing m to, say, 10 doses improve the small-sample operating characteristics of the BMDL if resource constraints force the N 's down to perhaps only 10–20 subjects/dose? Experimental design for dose-response studies with focus on the BMD is an emerging area in the statistical literature (Muri et al., 2009; Öberg, 2010; Sand et al., 2008), and how to optimally design/allocate experimental resources for BMD estimation and inferences under model uncertainty is yet another open question.

Acknowledgments

Thanks are due Prof. Rabi Bhattacharya, Dr. Hui Xiong, the Associate Editor, and two anonymous referees for their helpful suggestions. This research was supported by grant #RD-83241902 from the U.S. Environmental Protection Agency, by grant #R21-ES016791 from the U.S. National Institute of Environmental Health Sciences, and by grant #DMS-1106435 from the U.S. National Science Foundation. Its contents are solely the responsibility of the authors and do not necessarily reflect the official views of these agencies.

References

- Akaike, H. Information theory and an extension of the maximum likelihood principle. In: Petrov, BN.; Csaki, B., editors. *Proceedings of the Second International Symposium on Information Theory*. Budapest: Akademiai Kiado; 1973. p. 267–281.
- Bailer AJ, Noble RB, Wheeler MW. Model uncertainty and risk estimation for experimental studies of quantal responses. *Risk Analysis*. 2005; 25(2):291–299. [PubMed: 15876205]
- Britannica Editors. *The New Encyclopædia Britannica*. Chicago, IL: Encyclopædia Britannica, Inc.; 2002. Ockham's Razor; p. 867–868.
- Buckland ST, Burnham KP, Augustin NH. Model selection: An integral part of inference. *Biometrics*. 1997; 53(2):603–618.
- Buckley BE, Piegorsch WW. Simultaneous confidence bands for Abbott-adjusted quantal response models. *Statistical Methodology*. 2008; 5(3):209–219. [PubMed: 19412325]
- Burnham, KP.; Anderson, DA. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. 2nd edn.. New York: Springer-Verlag; 2002.
- Candolo C, Davison AC, Demetrio CGB. A note on model uncertainty in linear regression. *Journal of the Royal Statistical Society, series D (The Statistician)*. 2003; 52(2):165–177.
- Casella, G.; Berger, RL. *Statistical Inference*. 2nd edn.. Pacific Grove, CA: Duxbury; 2002.
- Cavanaugh JE. A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics and Probability Letters*. 1999; 42(4):333–343.
- Claeskens, G.; Hjort, NL. *Model Selection and Model Averaging*. New York: Cambridge University Press; 2008.
- Crump KS. A new method for determining allowable daily intake. *Fundam. Appl. Toxicol.* 1984; 4(5): 854–871. [PubMed: 6510615]
- Crump KS. Calculation of benchmark doses from continuous data. *Risk Analysis*. 1995; 15(1):79–89.
- Crump, KS.; Howe, R. A review of methods for calculating confidence limits in low dose extrapolation. In: Clayson, DB.; Krewski, D.; Munro, I., editors. *Toxicological Risk Assessment, Volume I: Biological and Statistical Criteria*. Boca Raton, FL: CRC Press; 1985. p. 187–203.
- Dankovic DA, Kuempel E, Wheeler M. An approach to risk assessment for TiO₂. *Inhalation Toxicology*. 2007; 19(Suppl 1):205–212. [PubMed: 17886069]

- Davis JA, Gift JS, Zhao QJ. Introduction to benchmark dose methods and U.S. EPA's Benchmark Dose Software (BMDs) version 2.1.1. *Toxicology and Applied Pharmacology*. 2012; 254(2):181–191. [PubMed: 21034758]
- Deutsch RC, Grego JM, Habing BT, Piegorsch WW. Maximum likelihood estimation with binary-data regression models: small-sample and large-sample features. *Advances and Applications in Statistics*. 2010; 14(2):101–116. [PubMed: 20640219]
- European Union. Technical Report #EUR 20418 EN/1-4. Ispra, Italy: European Chemicals Bureau (ECB); 2003. Technical Guidance Document (TGD) on Risk Assessment of Chemical Substances following European Regulations and Directives, Parts I–IV.
- Faes C, Aerts M, Geys H, Molenberghs G. Model averaging using fractional polynomials to estimate a safe level of exposure. *Risk Analysis*. 2007; 27(1):111–123. [PubMed: 17362404]
- Faustman EM, Bartell SM. Review of noncancer risk assessment: Applications of benchmark dose methods. *Human and Ecological Risk Assessment*. 1997; 3(5):893–920.
- Fletcher D, Dillingham PW. Model-averaged confidence intervals for factorial experiments. *Computational Statistics and Data Analysis*. 2011; 55(11):3041–3048.
- Grueber CE, Nakagawa S, Laws RJ, Jamieson IG. Multimodel inference in ecology and evolution: challenges and solutions. *Journal of Evolutionary Biology*. 2011; 24(4):699–711. [PubMed: 21272107]
- Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial. *Statistical Science*. 1999; 14(4):382–401. (corr. vol. 15(383), pp. 193–195).
- Hurvich CM, Tsai C-L. Regression and time series model selection in small samples. *Biometrika*. 1989; 76(2):297–307.
- Kang S-H, Kodell RL, Chen JJ. Incorporating model uncertainties along with data uncertainties in microbial risk assessment. *Regulatory Toxicology and Pharmacology*. 2000; 32(1):68–72. [PubMed: 11029270]
- Kodell RL. Managing uncertainty in health risk assessment. *Intl. J. Risk Assessment Manage*. 2005; 5(2/3/4):193–205.
- Kodell RL. Replace the NOAEL and LOAEL with the BMDL₀₁ and BMDL₁₀. *Environmental and Ecological Statistics*. 2009; 16(1):3–12.
- Liang H, Zou G, Wan ATK, Zhang X. Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association*. 2011; 106(495):1053–1066.
- Lukacs PM, Burnham KP, Anderson DR. Model selection bias and Freedman's paradox. *Annals of the Institute of Statistical Mathematics*. 2010; 62(1):117–125.
- Moerbeek M, Piersma AH, Slob W. A comparison of three methods for calculating confidence intervals for the benchmark dose. *Risk Analysis*. 2004; 24(1):31–40. [PubMed: 15027998]
- Moon H, Kim H-J, Chen JJ, Kodell RL. Model averaging using the Kullback information criterion in estimating effective doses for microbial infection and illness. *Risk Analysis*. 2005; 25(5):1147–1159. [PubMed: 16297221]
- Moon H, Kim SB, Chen JJ, George NI, Kodell RL. Model uncertainty and model averaging in the estimation of infectious doses for microbial pathogens. *Risk Analysis*. 2013 (in press).
- Morales KH, Ibrahim JG, Chen C-J, Ryan LM. Bayesian model averaging with applications to benchmark dose estimation for arsenic in drinking water. *Journal of the American Statistical Association*. 2006; 101(473):9–17.
- Muri SD, Schlatter JR, Brüschweiler BJ. The benchmark dose approach in food risk assessment: Is it applicable and worthwhile? *Food and Chemical Toxicology*. 2009; 47(12):2906–2925. [PubMed: 19682530]
- Namata H, Aerts M, Faes C, Teunis P. Model averaging in microbial risk assessment using fractional polynomials. *Risk Analysis*. 2008; 28(4):891–905. [PubMed: 18564995]
- Noble RB, Bailer AJ, Park R. Model-averaged benchmark concentration estimates for continuous response data arising from epidemiological studies. *Risk Analysis*. 2009; 29(4):558–564. [PubMed: 19144062]
- Öberg M. Benchmark dose approaches in chemical health risk assessment in relation to number and distress of laboratory animals. *Regulatory Toxicology and Pharmacology*. 2010; 58(3):451–454. [PubMed: 20800084]

- OECD. Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application. Series on Testing and Assessment Report #54. Paris: Environment Directorate, Organisation For Economic Co-Operation and Development; 2006.
- OECD. Draft Guidance Document on the performance of chronic toxicity and carcinogenicity studies, supporting TG 451, 452 and 453. Paris: Organisation For Economic Co-Operation and Development; 2008.
- Piegorsch, WW. Quantal response data. In: El-Shaarawi, AH.; Piegorsch, WW., editors. Encyclopedia of Environmetrics. Vol. 4. Chichester: John Wiley & Sons; 2012. p. 2065-2067.
- Piegorsch, WW.; Bailer, AJ. Analyzing Environmental Data. Chichester: John Wiley & Sons; 2005.
- Piegorsch WW, Xiong H, Bhattacharya RN, Lin L. Nonparametric estimation of benchmark doses in quantitative risk analysis. *Environmetrics*. 2012; 23(8)
- Portier CJ. Biostatistical issues in the design and analysis of animal carcinogenicity experiments. *Environmental Health Perspectives*. 1994; 102(Suppl. 1):5–8. [PubMed: 8187725]
- R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.
- Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*. 1997; 92(437):179–191.
- Ritz C, Streibig JC. Bioassay analysis using R. *Journal of Statistical Software*. 2005; 12(5) Art. No. 5.
- Sand S, Victorin K, Falk Filipsson A. The current state of knowledge on the use of the benchmark dose concept in risk assessment. *Journal of Applied Toxicology*. 2008; 28(4):405–421. [PubMed: 17879232]
- Schwarz G. Estimating the dimension of a model. *Annals of Statistics*. 1978; 6(2):461–464.
- Stern, AH. Environmental health risk assessment. In: Melnick, EL.; Everitt, BS., editors. Encyclopedia of Quantitative Risk Analysis and Assessment. Vol. 2. Chichester: John Wiley & Sons; 2008. p. 580-589.
- U.S. EPA. Technical Report #EPA/630/P-03/001F. Washington, DC: U.S. Environmental Protection Agency; 2005. Guidelines for Carcinogen Risk Assessment.
- U.S. EPA. Technical Report #EPA/100/R-12/001. Washington, DC: U.S. Environmental Protection Agency; 2012. Benchmark Dose Technical Guidance Document.
- U.S. General Accounting Office. Chemical Risk Assessment. Selected Federal Agencies' Procedures, Assumptions, and Policies. Report to Congressional Requesters #GAO-01-810; U.S. General Accounting Office; Washington, DC. 2001.
- Wang H, Zhang X, Zou G. Frequentist model averaging estimation: a review. *Journal of Systems Science & Complexity*. 2009; 22(4):732–748.
- West RW, Piegorsch WW, Peña EA, An L, Wu W, Wickens AA, Xiong H, Chen W. The impact of model uncertainty on benchmark dose estimation. *Environmetrics*. 2012; 23(8)
- Wheeler MW, Bailer AJ. Properties of model-averaged BMDLs: A study of model averaging in dichotomous response risk estimation. *Risk Analysis*. 2007; 27(3):659–670. [PubMed: 17640214]
- Wheeler MW, Bailer AJ. Model averaging software for dichotomous dose response risk estimation. *Journal of Statistical Software*. 2008; 26(5) Art. No. 5.
- Wheeler MW, Bailer AJ. Comparing model averaging with other model selection strategies for benchmark dose estimation. *Environmental and Ecological Statistics*. 2009; 16(1):37–51.
- Wickens, AA. Unpublished M.S. Thesis. Tucson, AZ: Program in Statistics, University of Arizona; 2011. Model Adequacy in Benchmark Risk Analysis For Cancer Risk Assessment.

Appendix 1

Calculating BMDs and BMDLs under the dose-response models in Table 1

For quantal response data $Y_i \sim \text{indep. Bin}(N_i, R[x_i])$, $i = 1, \dots, m$, we estimate the BMD and BMDL by appealing to likelihood-based operations. For each of the eight models in Table 1, the basic approach is the same, although the technical details change as the form of dose response changes. This appendix briefly summarizes the model-specific details.

Model M1

Under a simple linear-logistic model, $R(x; \beta) = \frac{1}{1 + \exp\{-\beta_0 - \beta_1 x\}}$, the extra risk is $R_E(x) = \{R(x; \beta) - R(0; \beta)\} / \{1 - R(0; \beta)\} = \frac{1 - \exp\{-\beta_1 x\}}{1 + \exp\{-\beta_0 - \beta_1 x\}}$. The MLE for the unknown parameter vector $\beta = [\beta_0 \ \beta_1]^T$ is found by maximizing the log-likelihood function

$$\ell(\beta) = \sum_{i=1}^m Y_i \log\{R(x_i; \beta)\} + \sum_{i=1}^m (N_i - Y_i) \log\{1 - R(x_i; \beta)\},$$

up to a constant not dependent upon β . Unfortunately, a closed-form solution is unavailable and hence calculation proceeds via computer iteration. Throughout, we employ the R programming environment (R Development Core Team, 2012). Given the MLE $\hat{\beta} = [\hat{b}_0 \ \hat{b}_1]^T$, the MLE for the extra risk function is simply $\hat{R}_E(x) = \frac{1 - \exp\{-\hat{b}_1 x\}}{1 + \exp\{-\hat{b}_0 - \hat{b}_1 x\}}$. Set this equal to the pre-specified BMR and solve for x to estimate the BMD:

$$\hat{\text{BMD}}_{100\text{BMR}} = \frac{1}{\hat{b}_1} \log \left\{ \frac{1 + e^{-\hat{b}_0} \text{BMR}}{1 - \text{BMR}} \right\} \quad (\text{A.1})$$

as per Table 1.

To compute a BMDL, a number of formulations are possible. Herein we employ the Wald $100(1 - \alpha)\%$ lower limit in (2.1). Needed is the standard error $se[\hat{\text{BMD}}_{100\text{BMR}}]$, which we find from the large-sample variance-covariance components of $\hat{\beta}$ using the well-known Delta method (Piegorsch and Bailer, 2005, §A.6.2). This approximation produces

$$\text{var}[\hat{\text{BMD}}_{100\text{BMR}}] \approx \frac{1}{\hat{b}_1^2} \left\{ \frac{e^{-\hat{b}_0} \text{BMR}}{1 + e^{-\hat{b}_0} \text{BMR}} \right\}^2 \text{var}[\hat{b}_0] + \frac{2\hat{\text{BMD}}_{100\text{BMR}}}{\hat{b}_1^2} \left\{ \frac{e^{-\hat{b}_0} \text{BMR}}{1 + e^{-\hat{b}_0} \text{BMR}} \right\} \text{cov}[\hat{b}_0, \hat{b}_1] + \frac{\hat{\text{BMD}}_{100\text{BMR}}^2}{\hat{b}_1^2} \text{var}[\hat{b}_1],$$

where $\text{var}[\hat{b}_j]$ are the estimated variances of b_j ($j = 0, 1$) and $\text{cov}[\hat{b}_0, \hat{b}_1]$ is their estimated covariance. These latter quantities are taken from the inverse of the Fisher information matrix, and are available, e.g., from the R function *glm*. Take

$$se[\hat{\text{BMD}}_{100\text{BMR}}] = \sqrt{\text{var}[\hat{\text{BMD}}_{100\text{BMR}}]}.$$

Model M2

Under a simple linear probit model, $R(x; \beta) = \Phi(\beta_0 + \beta_1 x)$, the extra risk is $R_E(x) = \{\Phi(\beta_0 + \beta_1 x) - \Phi(\beta_0)\} / \{1 - \Phi(\beta_0)\}$. As with model M1, the MLE for the unknown parameter vector $\beta = [\beta_0 \ \beta_1]^T$ is found by maximizing $\ell(\beta)$ via computer iteration. Given the MLE $\hat{\beta} = [\hat{b}_0 \ \hat{b}_1]^T$, the MLE for the extra risk is $\hat{R}_E(x) = \{\Phi(\hat{b}_0 + \hat{b}_1 x) - \Phi(\hat{b}_0)\} / \{1 - \Phi(\hat{b}_0)\}$. Set this equal to the pre-specified BMR and solve for x to estimate the BMD:

$$\hat{\text{BMD}}_{100\text{BMR}} = \frac{Q_{\text{BMR}} - \hat{b}_0}{\hat{b}_1} \quad (\text{A.2})$$

as per Table 1. Here $Q_{\text{BMR}} = \Phi^{-1}\{\text{BMR}[1 - \Phi(\hat{b}_0)] + \Phi(\hat{b}_0)\}$.

Similar to model M1, to compute the Wald BMDL via (2.1) we find

$$se[\hat{\text{BMD}}_{100\text{BMR}}] = \sqrt{\text{var}[\hat{\text{BMD}}_{100\text{BMR}}]} \text{ via a Delta-method approximation:}$$

$$\begin{aligned} \text{var}[\hat{\text{BMD}}_{100\text{BMR}}] &\approx \frac{I}{b_1^2} [(1 - \text{BMR}) \exp\{(Q_{\text{BMR}}^2 - b_0^2)/2\} - I]^2 \text{var}[b_0] \\ &\quad - \frac{2\hat{\text{BMD}}_{100\text{BMR}}}{b_1^2} [(1 - \text{BMR}) \exp\{(Q_{\text{BMR}}^2 - b_0^2)/2\} - 1] \text{cov}[b_0, b_1] \\ &\quad + \frac{\hat{\text{BMD}}_{100\text{BMR}}^2}{b_1^2} \text{var}[b_1]. \end{aligned}$$

Model M3

Under a quantal-linear (also known as a ‘one-stage’) model, $R(x) = 1 - \exp\{-b_0 - b_1 x\}$, the extra risk is $R_E(x) = 1 - \exp\{-b_1 x\}$. The MLE for the unknown parameter vector $\beta = [b_0 \ b_1]^T$ is found by maximizing $\ell(\beta)$ via constrained optimization to account for the boundary conditions $b_j \geq 0$, $j = 0, 1$. Given the constrained MLE $\hat{\beta} = [\hat{b}_0 \ \hat{b}_1]^T$, the MLE for the extra risk is $R_E(x) = 1 - \exp\{-\hat{b}_1 x\}$. Set this equal to the pre-specified BMR and solve for x to estimate the BMD:

$$\hat{\text{BMD}}_{100\text{BMR}} = \frac{-\ln(1 - \text{BMR})}{\hat{b}_1}. \quad (\text{A.3})$$

To compute the Wald BMDL via (2.1), we find $se[\hat{\text{BMD}}_{100\text{BMR}}] = \sqrt{\text{var}[\hat{\text{BMD}}_{100\text{BMR}}]}$ via a Delta-method approximation:

$$\text{var}[\hat{\text{BMD}}_{100\text{BMR}}] \approx \left\{ \frac{\hat{\text{BMD}}_{100\text{BMR}}}{\hat{b}_1} \right\}^2 \text{var}[\hat{b}_1].$$

Model M4

Under a quantal-quadratic model, $R(x) = 1 - \exp\{-b_0 - b_1 x^2\}$, the extra risk is $R_E(x) = 1 - \exp\{-b_1 x^2\}$. The MLE for $\beta = [b_0 \ b_1]^T$ is found by maximizing $\ell(\beta)$ via constrained optimization to account for the boundary conditions $b_j \geq 0$, $j = 0, 1$. Given the constrained MLE $\hat{\beta} = [\hat{b}_0 \ \hat{b}_1]^T$, the MLE for the extra risk is $R_E(x) = 1 - \exp\{-\hat{b}_1 x^2\}$. Set this equal to the pre-specified BMR and solve for x to estimate the BMD:

$$\hat{\text{BMD}}_{100\text{BMR}} = \sqrt{\frac{-\ln(1 - \text{BMR})}{\hat{b}_1}}. \quad (\text{A.4})$$

To compute the Wald BMDL via (2.1), we find $se[\hat{\text{BMD}}_{100\text{BMR}}] = \sqrt{\text{var}[\hat{\text{BMD}}_{100\text{BMR}}]}$ via a Delta-method approximation:

$$\text{var}[\hat{\text{BMD}}_{100\text{BMR}}] \approx \left\{ \frac{\hat{\text{BMD}}_{100\text{BMR}}}{2\hat{b}_1} \right\}^2 \text{var}[\hat{b}_1].$$

Model M5

Under a two-stage (multi-stage) model, $R(x) = 1 - \exp\{-\beta_0 - \beta_1 x - \beta_2 x^2\}$, the extra risk is $R_E(x) = 1 - \exp\{-\beta_1 x - \beta_2 x^2\}$. The MLE for $\beta = [\beta_0 \ \beta_1 \ \beta_2]^T$ is found by maximizing () via constrained optimization to account for the boundary conditions $\beta_j \geq 0 \ j$. Given the constrained MLE $\hat{\beta} = [\hat{b}_0 \ \hat{b}_1 \ \hat{b}_2]^T$, the MLE for the extra risk is $R_E(x) = 1 - \exp\{-\hat{b}_1 x - \hat{b}_2 x^2\}$. Set this equal to the pre-specified BMR and solve for x to estimate the BMD. The result is a quadratic equation in x , both of whose roots can be shown to be real. To determine which root to use, Piegorsch and Bailer (2005, Ex. 4.12) employ Descartes' Rule of Signs and find there is exactly one real positive root to take for BMD:

$$\hat{BMD}_{100BMR} = \frac{-\hat{b}_1 + \sqrt{\hat{b}_1^2 + 4\hat{b}_2 T_{BMR}}}{2\hat{b}_2}, \quad (A.5)$$

where $T_{BMR} = -\log(1 - BMR)$.

To compute the Wald BMDL via (2.1), we find $se[\hat{BMD}_{100BMR}] = \sqrt{var[\hat{BMD}_{100BMR}]}$ via a Delta-method approximation:

$$\begin{aligned} var[\hat{BMD}_{100BMR}] &\approx \frac{1}{4\hat{b}_2^2} \left\{ 1 - \frac{\hat{b}_1}{\sqrt{\hat{b}_1^2 + 4\hat{b}_2 T_{BMR}}} \right\}^2 var[\hat{b}_1] - \frac{1}{2\hat{b}_2^3} \left\{ 1 - \frac{\hat{b}_1}{\sqrt{\hat{b}_1^2 + 4\hat{b}_2 T_{BMR}}} \right\} \left\{ \hat{b}_1 - \frac{\hat{b}_1^2 + 2\hat{b}_2 T_{BMR}}{\sqrt{\hat{b}_1^2 + 4\hat{b}_2 T_{BMR}}} \right\} cov[\hat{b}_1, \hat{b}_2] \\ &+ \frac{1}{4\hat{b}_2^4} \left\{ \hat{b}_1 - \frac{\hat{b}_1^2 + 2\hat{b}_2 T_{BMR}}{\sqrt{\hat{b}_1^2 + 4\hat{b}_2 T_{BMR}}} \right\}^2 var[\hat{b}_2]. \end{aligned}$$

Model M6

Under a log-logistic model, $R(x) = \frac{\gamma_0 + \frac{1 - \gamma_0}{1 + \exp\{-\beta_0 - \beta_1 \log(x)\}}}{1 + \exp\{-\beta_0 - \beta_1 \log(x)\}}$, the extra risk is $R_E(x) = \{1 + \exp\{-\beta_0 - \beta_1 \log(x)\}\}^{-1}$. The MLE for $\beta = [\beta_0 \ \beta_1]^T$ is found by maximizing () via constrained optimization to account for the boundary conditions $\beta_0 \geq 0$ and $\beta_1 \geq 0$. The MLE for the extra risk is $R_E(x) = \{1 + \exp\{-\hat{b}_0 - \hat{b}_1 \log(x)\}\}^{-1}$. Set this equal to the pre-specified BMR and solve for x to estimate the BMD:

$$\hat{BMD}_{100BMR} = \exp\left\{\frac{U_{BMR} - \hat{b}_0}{\hat{b}_1}\right\}, \quad (A.6)$$

where $U_{BMR} = \log\{BMR/(1 - BMR)\} = \text{logit}(BMR)$.

To compute the Wald BMDL via (2.1), we find $se[\hat{BMD}_{100BMR}] = \sqrt{var[\hat{BMD}_{100BMR}]}$ via a Delta-method approximation:

$$var[\hat{BMD}_{100BMR}] \approx \frac{\hat{BMD}_{100BMR}^2}{\hat{b}_1^2} var[\hat{b}_0] + \frac{2\hat{BMD}_{100BMR}^2 \ln(\hat{BMD}_{100BMR})}{\hat{b}_1^2} cov[\hat{b}_0, \hat{b}_1] + \frac{\hat{BMD}_{100BMR}^2 \ln^2(\hat{BMD}_{100BMR})}{\hat{b}_1^2} var[\hat{b}_1]. \quad (A.7)$$

Model M7

Under a log-probit model, $R(x; \cdot) = \phi_0 + (1 - \phi_0) \Phi(\phi_0 + \phi_1 \log[x])$, the extra risk is $R_E(x) = \Phi(\phi_0 + \phi_1 \log[x])$. This has the same structure as model M6—called an *Abbott-adjusted model* (Buckley and Piegorsch, 2008)—with the logistic c.d.f. in M6 replaced by the standard normal c.d.f. in M7. Thus the equation for BMD_{100BMR} is similar to (A.6) where now $BMD_{100BMR} = \exp\{(V_{BMR} - b_0)/b_1\}$, with $V_{BMR} = \Phi^{-1}(BMR)$. Further, the equation for $var[BMD_{100BMR}]$ has exactly the same form as that in (A.7), now using the MLEs found under model M7.

Model M8

Under a Weibull model, $R(x; \cdot) = \phi_0 + (1 - \phi_0)[1 - \exp\{-e^{\phi_0 x^{\phi_1}}\}] = \phi_0 + (1 - \phi_0)[1 - \exp\{-\exp(\phi_0 + \phi_1 \log[x])\}]$, we again recover an Abbott-adjusted dose-response function, now with extra risk $R_E(x) = 1 - \exp\{-\exp(\phi_0 + \phi_1 \log[x])\}$ (Buckley and Piegorsch, 2008). Thus the expression for BMD_{100BMR} is once again similar to (A.6): $BMD_{100BMR} = \exp\{(W_{BMR} - b_0)/b_1\}$, with $W_{BMR} = \log\{-\log(1 - BMR)\}$. Further, the equation for $var[BMD_{100BMR}]$ has exactly the same form as that in (A.7), now using the MLEs found under model M8.

Figure 1. Summary boxplots of Frequentist Model Averaged (FMA) BMDL empirical coverage across all model configurations, as a function of per-dose sample size N , when the AIC is used for the IT weights in (3.1). $BMR = 0.10$ and $\alpha = 0.05$.

Figure 2. Summary boxplots of Frequentist Model Averaged (FMA) BMDL empirical coverage across all model configurations, as a function of per-dose sample size N , when the BIC is used for the IT weights in (3.1). $BMR = 0.10$ and $\alpha = 0.05$.

Figure 3. Summary boxplots of median extra risks achieved by the Frequentist Model Averaged (FMA) BMDL, $R_E(\text{BMDL})$, as a function of per-dose sample size N . Results only for the AIC weights in (3.1). $\text{BMR} = 0.10$ and $\alpha = 0.05$.

Figure 4. Summary boxplots of upper-quartile extra risks achieved by the Frequentist Model Averaged (FMA) BMDL, $R_E(\text{BMDL})$, as a function of per-dose sample size N . Results only for the AIC weights in (3.1). $\text{BMR} = 0.10$ and $\alpha = 0.05$.

Table 1

Common quantal dose-response models in environmental toxicology and carcinogenic risk assessment

Model	Code	R(x)	BMD	Restrictions/Notes
Logistic	M1	$\frac{1}{1+\exp\{-\beta_0 - \beta_1 x\}}$	$\frac{1}{\beta_1} \log \left\{ \frac{1+e^{-\beta_0} \text{BMR}}{1 - \text{BMR}} \right\}$	—
Probit	M2	$\Phi(\beta_0 + \beta_1 x)$	$\frac{Q_{\text{BMR}} - \beta_0}{\beta_1}$	$Q_{\text{BMR}} = \Phi^{-1}\{\text{BMR}[\Phi(\beta_0) + (1 - \Phi(\beta_0))]\}$
Quantal-linear	M3	$1 - \exp\{-\beta_0 - \beta_1 x\}$	$\frac{-\ln(1 - \text{BMR})}{\beta_1}$	$\beta_0 \geq 0, \beta_1 \geq 0$
Quantal-quadratic	M4	$\beta_0 + (1 - \beta_0)(1 - \exp\{-\beta_1 x^2\})$	$\sqrt{\frac{-\ln(1 - \text{BMR})}{\beta_1}}$	$\beta_0 \geq 0, \beta_1 \geq 0$
Two-stage	M5	$1 - \exp\{-\beta_0 - \beta_1 x - \beta_2 x^2\}$	$\frac{-\beta_1 + \sqrt{\beta_1^2 + 4\beta_2 T_{\text{BMR}}}}{2\beta_2}$	$T_{\text{BMR}} = -\log(1 - \text{BMR})$ $\beta_j \geq 0, j = 0, 1, 2$
Log-logistic	M6	$\gamma_0 + \frac{1 - \gamma_0}{1 + \exp\{-\beta_0 - \beta_1 \log[x]\}}$	$\exp \left\{ \frac{\text{logit}(\text{BMR}) - \beta_0}{\beta_1} \right\}$	$\beta_0 \geq 0, \beta_1 \geq 0$
Log-probit	M7	$\beta_0 + (1 - \beta_0) \Phi(\beta_1 \log[x])$	$\exp \left\{ \frac{\Phi^{-1}(\text{BMR}) - \beta_0}{\beta_1} \right\}$	$\beta_0 \geq 0, \beta_1 \geq 0$
Weibull	M8	$\beta_0 + (1 - \beta_0)[1 - \exp\{-e^{\beta_1 x}\}]$	$\exp \left\{ \frac{W_{\text{BMR}} - \beta_0}{\beta_1} \right\}$	$W_{\text{BMR}} = \log\{-\log(1 - \text{BMR})\}$ $\beta_0 \geq 0, \beta_1 \geq 0$

Notes:

Abbreviations: BMD, Benchmark Dose; BMR, Benchmark Response.

The quantal linear (M3) model is also referred to as the ‘one-stage’ model or as the ‘complementary-log’ model, and may equivalently appear as $\beta_0 + (1 - \beta_0)(1 - \exp\{-\beta_1 x\})$, where $\beta_0 = 1 - \exp\{-\beta_0\}$.

The two-stage model (M5) and the quantal-linear/one-stage model (M3) are special cases of the more general ‘Multi-stage’ model in carcinogenicity testing (Piegorsch and Bailer, 2005, §4.2.1).

Table 2

Models and configurations for the Monte Carlo evaluations. Model codes are from Table 1

Model code	Parameters	Configuration					
		A	B	C	D	E	F
M1	0	-4.5951	-4.5951	-2.1972	-2.9444	-0.8473	-2.1972
	1	2.3979	3.2088	1.3499	2.9444	1.9459	4.3944
M2	0	-2.3263	-2.3263	-1.2816	-1.6449	-0.5244	-1.2816
	1	1.0448	1.4847	0.7572	1.6449	1.1989	2.5631
M3	0	0.0101	0.0101	0.1054	0.0513	0.3567	0.1054
	1	0.0953	0.2131	0.2513	0.6419	1.0296	2.1972
M4	0	0.0100	0.0100	0.1000	0.0500	0.3000	0.1000
	1	0.0953	0.2131	0.2513	0.6419	1.0296	2.1972

Model code	Parameters	Configuration					
		A	B	C	D	E	F
M5	0	0.0101	0.0101	0.1054	0.0513	0.3567	0.1054
	1	0.0278	0.0370	0.0726	0.5797	0.4796	0.1539
	2	0.0675	0.1761	0.1788	0.0622	0.5501	2.0433
M6	0	0.0100	0.0100	0.1000	0.0500	0.3000	0.1000
	0	-2.3026	-1.4376	-1.2528	-0.1054	0.5878	2.0794
M7	1	1.6781	1.8802	1.7603	1.3333	1.9735	3.3219
	0	0.0100	0.0100	0.1000	0.0500	0.3000	0.1000

	0	-1.3352	-0.8708	-0.7647	-0.0660	0.3661	1.2206
	1	0.7808	0.9794	0.9456	0.8189	1.2261	1.9626
M8	0	0.0100	0.0100	0.1000	0.0500	0.3000	0.1000
	0	-2.3506	-1.5460	-1.3811	-0.4434	0.0292	0.7872
	1	1.6310	1.7691	1.6341	1.0716	1.4483	1.9023

Table 3

Frequentist Model Averaged (FMA) BMDL empirical coverage results from the Monte Carlo evaluations at per-dose sample size $N = 50$. BMR = 0.10 and $\alpha = 0.05$

Model code	IT weight	Configuration (from Table 2)					
		A	B	C	D	E	F
M1	AIC	0.9945	0.9990	0.9990	0.9990	0.9995	0.9800
	AICc	0.9945	0.9990	0.9990	0.9990	0.9995	0.9805
	KIC	0.9945	0.9990	0.9990	0.9990	1.0000	0.9845
M2	BIC	0.9945	0.9990	0.9990	0.9990	1.0000	0.9855
	AIC	0.9960	0.9990	0.9985	0.9955	0.9995	0.9680
	AICc	0.9960	0.9990	0.9985	0.9955	0.9995	0.9700
M3	KIC	0.9960	0.9990	0.9985	0.9960	1.0000	0.9720
	BIC	0.9960	0.9990	0.9985	0.9970	1.0000	0.9685
	AIC	0.9965	0.9710	1.0000	0.9390	0.9965	0.9530
M4	AICc	0.9965	0.9710	1.0000	0.9365	0.9970	0.9510
	KIC	0.9965	0.9710	1.0000	0.9210	0.9980	0.9345
	BIC	0.9965	0.9710	1.0000	0.8870	0.9980	0.9015
M5	AIC	0.9950	0.9990	0.9995	0.9985	1.0000	0.9845
	AICc	0.9950	0.9990	0.9995	0.9985	1.0000	0.9855
	KIC	0.9950	0.9990	0.9995	0.9985	1.0000	0.9905
M6	BIC	0.9950	0.9990	0.9995	0.9985	1.0000	0.9950
	AIC	0.9970	0.9990	0.9990	0.9375	1.0000	0.9845
	AICc	0.9970	0.9990	0.9990	0.9365	1.0000	0.9860
M7	KIC	0.9970	0.9990	0.9990	0.9230	1.0000	0.9905
	BIC	0.9970	0.9990	0.9990	0.8855	1.0000	0.9955
	AIC	0.9970	0.9995	0.9990	0.9535	1.0000	0.9970
M8	AICc	0.9970	0.9995	0.9990	0.9525	1.0000	0.9970
	KIC	0.9970	0.9995	0.9990	0.9465	1.0000	0.9980
	BIC	0.9970	0.9995	0.9990	0.9270	1.0000	0.9995
M9	AIC	0.9985	1.0000	0.9985	0.9680	1.0000	0.9990
	AICc	0.9985	1.0000	0.9985	0.9680	1.0000	0.9990

Configuration (from Table 2)

	A	B	C	D	E	F
M8	AICc	0.9985	1.0000	0.9985	0.9665	1.0000
	KIC	0.9985	1.0000	0.9985	0.9620	1.0000
	BIC	0.9985	1.0000	0.9985	0.9470	1.0000
	AIC	0.9970	0.9995	0.9990	0.9430	1.0000
	AICc	0.9970	0.9995	0.9990	0.9425	1.0000
	KIC	0.9970	0.9995	0.9990	0.9295	1.0000
	BIC	0.9970	0.9995	0.9990	0.8945	1.0000

Notes:

Abbreviations: BMDL, Benchmark Dose lower Limit; BMR, Benchmark Response.

Model codes are from Table 1. IT weights for use in (3.1) are discussed in §2.3 and §3.1.

Multi-model benchmark analysis of carcinogenicity data in §5. Models, $\mathcal{M}_q$ from Table 1. Benchmark response (BMR) set to 0.10

Table 4

Model, $\mathcal{M}_q$	BMD_q	$\text{var}[\text{BMD}_q \mathcal{M}_q]$	AIC_q	w_q from (3.1)	$w_q \text{BMD}_q$
Logistic, M1	2.0814	0.0757	17.6599	0.2464	0.5129
Probit, M2	1.9266	0.0583	17.4719	0.2707	0.5216
Q-linear, M3	0.7114	0.0105	28.2104	0.0013	0.0009
Q-quadratic, M4	2.0460	0.0323	17.9837	0.2096	0.4289
Two-stage, M5	1.7006	1.2418	19.6303	0.0920	0.1565
Log-logistic, M6	2.1303	0.4140	20.6561	0.0551	0.1174
Log-probit, M7	2.3132	0.4593	21.1819	0.0424	0.0980
Weibull, M8	1.8197	0.3354	19.8494	0.0825	0.1501
					$\overline{\text{BMD}}_{10}=1.9862$