Environ Ecol Stat (2009) 16:37–51
DOI 10.1007/s10651-007-0071-7

Comparing model averaging with other model selection strategies for benchmark dose estimation

Matthew W. Wheeler · A. John Bailer

Received: 1 March 2005 / Revised: 2 January 2006 / Published online: 25 March 2008

Abstract Model averaging (MA) has been proposed as a method of accommodating model uncertainty when estimating risk. Although the use of MA is inherently appealing, little is known about its performance using general modeling conditions. We investigate the use of MA for estimating excess risk using a Monte Carlo simulation. Dichotomous response data are simulated under various assumed underlying dose–response curves, and nine dose–response models (from the USEPA Benchmark dose model suite) are fit to obtain both model specific and MA risk estimates. The benchmark dose estimates (BMDs) from the MA method, as well as estimates from other commonly selected models, e.g., best fitting model or the model resulting in the smallest BMD, are compared to the true benchmark dose value to better understand both bias and coverage behavior in the estimation procedure. The MA method has a small bias when estimating the BMD that is similar to the bias of BMD estimates derived from the assumed model. Further, when a broader range of models are included in the family of models considered in the MA process, the lower bound estimate provided coverage close to the nominal level, which is superior to the other strategies considered. This approach provides an alternative method for risk managers to estimate risk while incorporating model uncertainty.

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

M. W. Wheeler (✉) · A. J. Bailer
Risk Evaluation Branch, National Institute for Occupational Safety and Health, MS C-15,
4676 Columbia Parkway, Cincinnati, OH 45226, USA
e-mail: aez0@cdc.gov

A. J. Bailer
Center for Environmental Toxicology and Statistics, Department of Mathematics and Statistics,
Miami University, Oxford, OH 45056, USA

Keywords Bayesian model averaging · Model uncertainty · Risk estimation

1 Introduction

Risk assessors are frequently interested in estimating the excess risk for populations exposed at some specified level of an occupational or environmental hazard. Alternatively, they may be interested in the dose associated with a specified excess risk. These endpoints, which are often estimated through regression modeling, are dependent on the model form used. As an example, risk-related endpoints are often estimated using animal toxicity studies where some dichotomous outcome (death, tumor response, etc.) is modeled as a function of the dose considered, and excess risk is determined from this dose–response model. In this setting there often exist multiple dose–response models that describe the data well, and the risk assessor has no a priori reason, biological or otherwise, to prefer a given model to all other models considered. The problem is further compounded by the observation that the risk estimates, specifically the lower bound estimates, derived from comparably-fitting models may significantly differ. Since the true model is unknown, model uncertainty exists in the risk estimation process. As there have been few techniques developed to address this uncertainty, risk assessors frequently ignore model uncertainty resulting in risk estimates that are based on the chosen model. In effect, this ignores all other risk estimates. We contend that this model uncertainty accounts for an important portion of the uncertainty in risk estimation and should be taken into account.

One method that can be used to account for this uncertainty in risk estimation is model averaging (MA) (Buckland et al. 1997). This technique, which estimates risk from a weighted average of all models considered, was used by Kang et al. (2000) to estimate risk across different microbiological dose response models. Bailer et al. (2005a), Bailer et al. (2005b) applied a similar technique, Bayesian model averaging (BMA), to animal toxicity studies where no a priori model was assumed. Both averaging methods calculate excess risk from a weighted average of risk estimated from each model considered. This central estimate, which incorporates model uncertainty, can then be used as an alternative to estimates derived from one specific model.

Although the idea of averaging risk across all models is inherently appealing, little is known about the general performance of MA in comparison to other model selection heuristics commonly employed in risk estimation. These heuristics, which include picking the "best" model or selecting the model that gives the greatest risk estimate or lowest benchmark dose estimate, can be compared to MA through a simulation experiment. We investigate the properties of a MA benchmark dose (BMD) estimate in comparison to the other BMD estimation strategies mentioned above by applying these methods to an animal toxicity experiment in which a dichotomous response is measured through a computer simulation experiment. Bias as well as the coverage is investigated to determine the characteristics of MA estimators.

2 Methods

2.1 Dose–response models

A number of dose–response models exist that can be used to fit dichotomous experimental data; however, for the purpose of this study we chose nine models which are available through the USEPA Benchmark Dose software (USEPA 2001). The models, as well as their parameter bounds, are described as follows:

$$\text{logistic:} \quad \pi(d_i) = \frac{1}{1 + \exp\left[-(\alpha + \beta d_i)\right]} \tag{1}$$

$$\text{log-logistic:} \quad \pi(d_i) = \gamma + \frac{(1-\gamma)}{1 + \exp\left[-(\alpha + \beta \ln(d_i))\right]}, \quad 0 \le \gamma < 1, \ \beta \ge 1 \tag{2}$$

$$\text{gamma:} \quad \pi(d_i) = \gamma + (1-\gamma)\frac{1}{\Gamma(\alpha)} \int_0^{\beta d_i} t^{\alpha-1} e^{-t} dt, \quad 0 \le \gamma < 1, \ \alpha \ge 1, \ \beta \ge 0 \tag{3}$$

$$\text{multistage (degree } = 2) \quad \pi(d_i) = \gamma + (1-\gamma)\left[1 - \exp(-\theta_1 d_i - \theta_2 d_i^2)\right], \quad 0 \le \gamma < 1, \ \theta_1 \ge 0, \ \theta_2 \ge 0 \tag{4}$$

$$\text{probit:} \quad \pi(d_i) = \Phi(a + \beta d_i) \tag{5}$$

$$\text{log-probit:} \quad \pi(d_i) = \gamma + (1-\gamma)\Phi\left[a + \beta \ln d_i\right], \quad 0 \le \gamma < 1, \ \beta \ge 1 \tag{6}$$

$$\text{quantal-linear:} \quad \pi(d_i) = \gamma + (1-\gamma)\left[1 - \exp(-\beta d_i)\right], \quad 0 \le \gamma < 1 \tag{7}$$

$$\text{quantal-quadratic} \quad \pi(d_i) = \gamma + (1-\gamma)\left[1 - \exp(-\beta d_i^2)\right], \quad 0 \le \gamma < 1 \tag{8}$$

$$\text{weibull} \quad \pi(d_i) = \gamma + (1-\gamma)\left[1 - \exp(-\beta d_i^{\alpha})\right], \quad 0 \le \gamma < 1, \ \alpha \ge 1, \ \beta \ge 0 \tag{9}$$

where π_i represents the probability of tumor response in the ith group given the dose d_i, $\Gamma(\alpha)$ = gamma function evaluated at α, $\Phi(x)$ is the cumulative distribution function

of the standard normal density at x (i.e., the integral of a $N(0, 1)$ density from $-\infty$ to x), and $\pi_i = \gamma$ when $d_i = 0$ for models (2) and (6). Further all bounds, which are described above, reflect the default specified bounds used by the USEPA software.

2.2 Excess risk definition

From these models, excess risk can be characterized through the use of the benchmark dose (Crump 1984). The BMD is defined as the dose that increases risk over the background response by some specified level relative to the control response. This level is known as the benchmark response (BMR) and is commonly set at values of 1%, 5% and 10%. Given the BMR and a dose–response model, the benchmark dose is the dose that satisfies the following equation

$$\text{BMR} = \frac{\pi(\text{BMD}) - \pi(0)}{1 - \pi(0)}, \tag{10}$$

where $\pi(d)$ represents the probability of response, given the dose d, described by the models (1–9). Though other formulations exist for excess risk characterization, including the added risk $= \pi(\text{BMD}) - \pi(0)$, we use the extra risk formula (2.10) exclusively for the remainder of the paper. The BMD is typically estimated using maximum likelihood methods, and the corresponding $100(1 - \alpha)\%$ lower bound on the benchmark dose (BMDL) can be estimated using likelihood profiling.

As mentioned previously, estimates of the BMD and the BMDL derived from solving equation (10) are model specific and different models often yield different estimates. Since there is often no reason to prefer one model over another, the BMD, in practice, is estimated using all reasonable models and the reported dose estimate is selected from this collection of estimates.

2.3 "Best" and smallest dose estimates

One such criterion a researcher may use is to pick the "best" fitting model. Where "best" can be described by any number of metrics including, but not limited to, penalized likelihood based summaries, or results of a goodness of fit test, such as the Pearson X^2 goodness-of-fit test. Given a specific metric the researcher then chooses the model and corresponding risk estimates that coincides with the "best" model for that specific parameter. We define "best" based upon the Pearson X^2 goodness-of-fit test, where the "best" model is chosen to have the largest P-value derived from this test. For example, consider a set of competing models that include the logistic and quantal linear forms, having goodness-of-fit P-values of 0.23 and 0.40, respectively, under this heuristic one would use the BMD/BMDL derived from the quantal linear model.

While the best fit provides a model that gives the closest match between the observed experimental data and the predicted responses, other model selection methods may be more desirable to the risk assessor. One might seek a risk estimate that attempts to guarantee the true BMD is larger than the selected BMD estimate. Given models which fit the data, one could choose the model that provides exposure estimates which

yield the lowest estimated benchmark dose (or highest estimated risk) amongst all competing models. For the purposes of this study, we pick the smallest BMD and its associated lower limit from all models considered.

2.4 Model averaging

The BMD/BMDL from "best" fit as well as the lowest-dose-model-selection strategy yield excess risk estimates; however, they still ignore the model uncertainty. Model averaging seeks to take into account this uncertainty by incorporating all models from some specified family of models into the estimation process through a weighted average of the models considered. This technique has been applied in a general modeling context by Raftery (1995), who suggested the use of the posterior model probabilities as weights; these weights were derived from a Bayesian analysis of all models considered. As a full Bayesian analysis is often difficult, Buckland et al. (1997) proposed simpler methods, where weights are based upon the penalized likelihood functions formed from the AIC (Akaike 1978) and BIC (Schwartz 1978). These criteria are defined as $-2 \log L + 2p$ for the AIC, and $-2 \log (L) + p \log (n)$ for the BIC. Here p represents the number of parameters in the model, L is the value of the likelihood at its' maximum, and n represents the sample size. Using Buckland et al.'s procedure, when the BIC is used, the procedure is identical to the BMA employed by Bailer et al. (2005a) to estimate excess risk, and if the AIC is used the procedure is identical to Kang et al.'s approach. We thus describe the Buckland et al. MA procedure in application to risk analysis, using both the AIC and the BIC. For a description of BMA in which priors are fully specified we refer the reader to the description by Hoeting et al. (1999).

Model averaging seeks to estimate the BMD using an average of model-specific benchmark doses. This weighted average is based upon summing weighted model-specific benchmark dose estimates, $\hat{\Delta}_i$, over all models considered, and is represented as $\hat{\Delta} = \sum_{i=1}^{k} \hat{\Delta}_i w_i$, where $\hat{\Delta}_i$ represents the estimated benchmark dose derived from the ith model and w_i represents the corresponding weight for the ith model. Equivalent estimates on the lower bound can be constructed using the BMDL in the above formula. Alternatives to this construction of BMDL would be a weighted average/mixture of dose–response curves. Razzaghi and Kodell (2000) discuss this mixture in the context of finite mixture modeling applied to risk estimation. Given the model M_j and a model space that includes k models, the weight is calculated according to the following formula

$$w_j = \frac{\exp(-I_j/2)}{\sum_{i=1}^{K} \exp(-I_i/2)},$$

where I_i = AIC (used in Kang et al. 2000) or I_i = BIC (used in Bailer et al. 2005a). The BIC based weights have the added interpretation of posterior model probabilities, or the probability the model is the true model, given the data; no such interpretation exists for the AIC based weights.

Though models (1–9) represent nominally different dose–response curves, some of the models considered are nested within other models considered. Further it is frequently the case that some of these nested model, due to the nature of the estimation procedure, degenerate into their simpler counterparts in the model space. For example, based upon the default behavior of the EPA software, the Weibull shape parameter α is bounded to be greater than or equal to 1, and if this value hits the bound the Weibull model degenerates into the quantal linear model. We address this problem using the Occam's razor, which was used by Raftery et al. (1997) to eliminate models whose nested counterparts have a higher posterior probability relative to the weighted average. Given the case of the Weibull model, the model would be removed from the MA process if its shape parameter reached the lower estimated bound.

3 Computer simulation

For the simulation study, we consider two experimental designs, one allowing four dose groups and the other allowing six dose groups. The dose levels were specified as $d = 0$, 0.25, 0.50, and 1.0 for the four group experiment, and $d = 0$, 0.0625, 0.125, 0.25, 0.5 and 1.0 for the six group experiment. For both designs, 50 animals were assigned to each dose group, implying a total sample of $n = 200$ for the four dose group design and $n = 300$ for the six dose group design. Within each design, the measured dichotomous outcome (death, tumor response, etc.) is recorded for each animal, and the probability of response π is estimated by the dose–response curves (1–9).

In order to represent the scope and variety of all possible dose–response curves data were simulated under a variety of assumed truths. Data were generated in a two-step process. First general response patterns were chosen to reflect the variety of dose–response patterns frequently observed. Six dose–response patterns were chosen to represent shallow, moderate and steep dose–response; as well as low and high background response rates (Table 1). These response patterns were then used as the basis for determining the underlying true model, which was used in the simulation. This was accomplished by estimating the models (1–9) parameters corresponding to these response patterns. This estimation procedure is described graphically in Fig. 1 for the case of the Weibull model. Given the six response patterns and nine dose–response models, a total of 54 different truth conditions were used in the simulation.

Table 1 This describes the six response patterns, for the doses 0, 0.5 and 1.0, respectively, that were used to generate the true underlying curvature for the simulation

These dose–response curves exhibited curvature ranging from linear to quadratic responses

Response pattern	Response curvature
1	(2%, 7%, 20%)
2	(2%, 14%, 34%)
3	(2%, 25%, 50%)
4	(7%, 12%, 25%)
5	(7%, 21%, 41%)
6	(7%, 32%, 55%)

Response Pattern 2			
Dose	0	0.5	1
Response	2%	14%	34%

Weibull Fit →

True Weibull Model		
γ	α	β
0.02	1.60	0.40

Fig. 1 Graphical description of the true model generation technique that was used to determine all true underlying curves used in the simulation

Given true dose–response curves (π_d), defined by the procedure described above, responses were assumed to be distributed binomially with response probability π_d related to the underlying true model at dose d. For each simulation condition, 2,000 samples were taken and the benchmark dose (BMD) and its lower limit (BMDL), using the models (1–9), were estimated for BMRs of 1% and 10%. From these nine fits a single risk estimate and its corresponding 95% lower bound were constructed using the "best" method, the smallest-BMD method, and MA using both the BIC as well as the AIC. To investigate the effect of the true model on risk estimation these values were computed by both including and excluding the true model from the family of models that served as the basis for the MA computation. The estimated values were then compared against the true value allowing estimates of bias as well as coverage.

Data were simulated using the *R* statistical programming language (R Development Core Team 2005), and the dose–response models were fit using a modified version of the USEPA benchmark dose software. Since the USEPA benchmark dose software will not estimate BMDs for flat and negative dose–response curves, a Cochran–Armitage trend test was performed before models (1-9) were fit; if this test detected a significant trend, corresponding to a test with α set to 0.2, all models were fit to the data, otherwise the condition was noted and risk was not estimated, which resulted in <2,000 simulation conditions for some of the response patterns. While this does introduce some bias into the simulation, it reflects standard experimental practice; risk is seldom estimated from extremely shallow or negative dose–response curves (since a BMD estimate is infinite if no dose–response is present). Of the true dose–response curves investigated, those generated from response pattern four were the most affected having approximately 2.5% or about 50 of the simulations removed for each condition.

4 Results

All model-averaging results reported describe a dose estimate computed excluding the true model from the family of models considered in the averaging process; this reflects an assumption that the true model is typically not included in any risk estimation experiment. Figures 2 and 3 display the estimated bias of BMA in comparison to the "best" model risk estimate for the 4 dose group experimental condition at BMRs of both 1% and 10%. Results from the 6 dose group condition give similar results and are not shown. Each point represents the estimated absolute bias of the BMA risk estimate (plotted on the x-axis), compared to the "best" model's estimated bias (plotted on the y-axis). Further, points which lie above the line $y = x$ imply the "best" estimate is more biased for that simulation condition, and points below the line imply the BMA estimate is more biased for that simulation condition. Figures 2 and 3 suggest that the two methods are similar in their bias when estimating extra risk for BMRs of 1% and

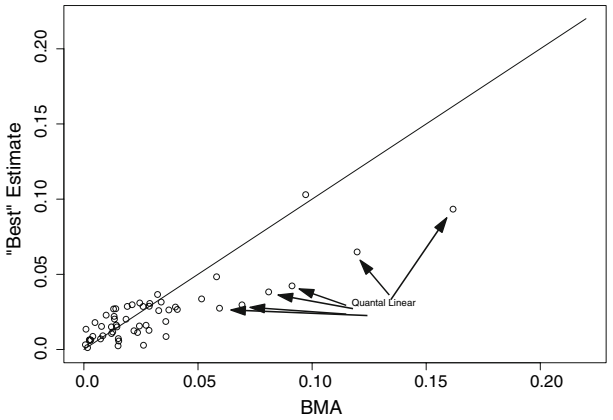

Fig. 2 Estimated absolute bias of the BMA is graphed in comparison to the estimated absolute bias of the "best" model. Each point represents estimated bias for one of the 54 simulation conditions for a BMR = 10%. The line represents equal bias between the two methods. Here points below the line imply the BMA is more biased and points above the line imply more bias for the "best" model estimate

10%, with most points lying at or around the line $y = x$. The notable exception occurs when the true underlying model is quantal linear in these six cases the MA procedure is noticeably more biased than the "best" estimate. When one compares the bias of MA to that of the conservative model strategy (figure not shown), MA consistently exhibits lower bias relative to conservative strategy for all underlying true model forms, except that of the quantal linear case; as with the "best" estimate, the conservative strategy provides less overall bias in comparison to MA. If one were to look at the bias of MA with weights derived from the BIC and the AIC the estimated bias is virtually identical. Finally, the bias of BMD estimates from the true model are comparable to the bias from the MA BMD estimates.

Coverage of the estimates, that is Pr(true BMD > estimated BMDL), was also investigated. The BMA estimate and the "best" model estimate, for the 95% lower bounds, is compared in Figs. 4 and 5 for BMRs of 10% and 1%, respectively. These figures suggest that under many conditions MA provides coverage closer to the nominal value, and in some circumstances provide coverage at the nominal level. Further in situations where the "best" model does provide coverage closer to that of the nominal 95% level it never reaches the expected level, often exhibiting coverage probabilities of 85% or lower.

One can also compare MA to the smallest-BMD estimation method. In this situation, the smallest-BMD estimates coverage (figure not shown) provides near 100% coverage regardless of the underlying true model form. Figure 6 illustrates the differences between MA using the AIC and BIC. As with bias, the two methods perform similarly under most circumstances, The differences are most pronounced when both MA techniques show poor coverage. In this case, the AIC based weighting procedure consistently outperforms BMA. Finally Fig. 7 illustrates the difference between MA when the true model is included in the computation and when it is not. In most cases there is no real difference between coverage, and, as with the case of AIC

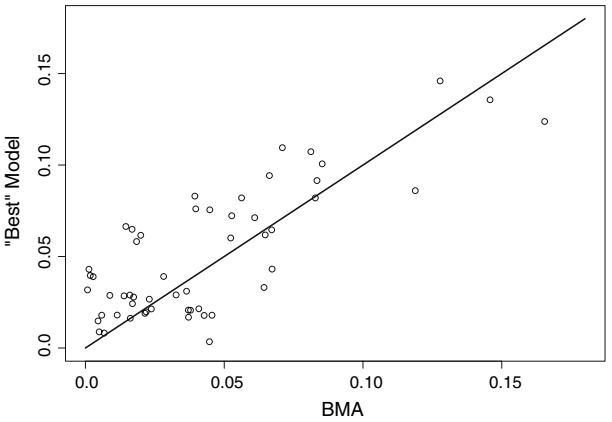

Fig. 3 Estimated absolute bias of the BMA is graphed in comparison to the estimated absolute bias of the "best" model. Each point represents estimated bias for one of the 54 simulation conditions for a BMR = 1%. The line represents equal bias between the two methods. Here points below the line imply the BMA is more biased and points above the line imply more bias for the "best" model estimate

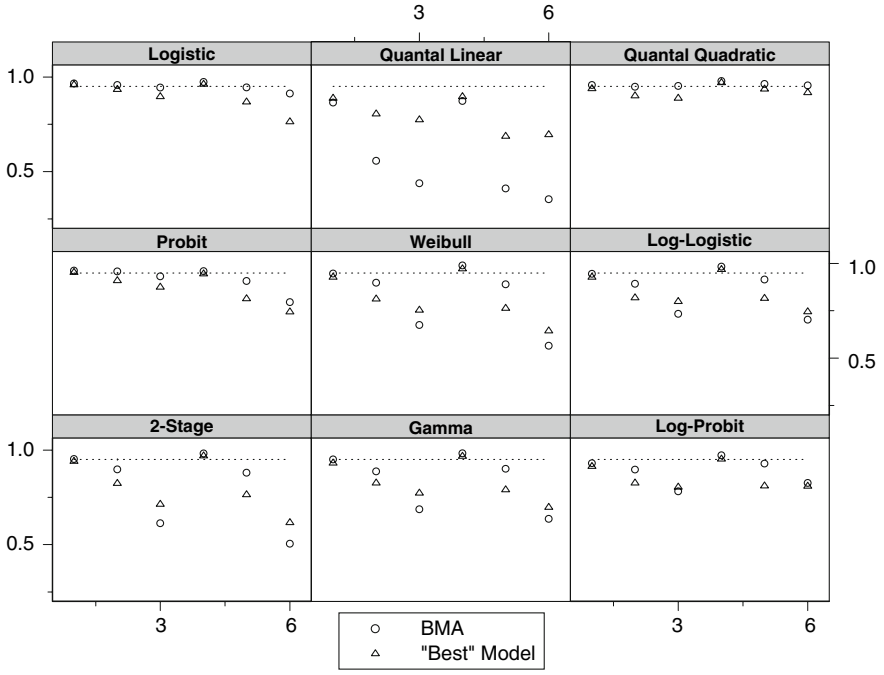

Fig. 4 Observed coverage for BMA as well as the "Best" model under all assumed truths with BMR = 10% and a 95% confidence level. Given the true underlying dose–response model, represented by the header, the points display each method's estimated coverage of the BMD for response patterns 1–6 from Table 1

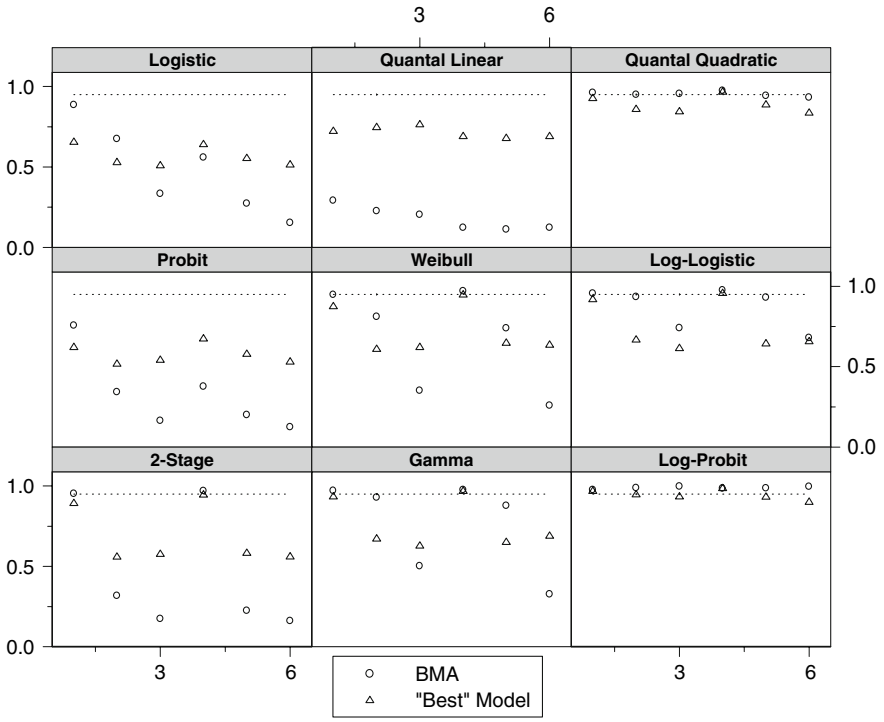

Fig. 5 Observed coverage for BMA as well as the "Best" model under all assumed truths with BMR = 1% and a 95% confidence level. Given the true underlying dose–response model, represented by the header, the points display each method's estimated coverage of the BMD for response patterns 1–6 from Table 1

weighting and BIC weighting, the largest difference occurs when the MA procedure poorly describes the true underlying risk.

5 Model averaging performance

The above results, in terms of bias and coverage, imply that MA gives mixed results for BMD risk estimation. There are certain situations where the bias is minimal and the lower bound estimate provides coverage very near the nominal 95% level. For other cases, however, specifically the quantal linear model, MA performs poorly. Analyzing the model fits give insight into why this pattern is observed, and further provides direction on how the conditions of low coverage may be avoided.

Figure 8 describes one such situation where MA calculates accurate dose estimates for BMRs of 1% and 10%. In this situation, one where the true underlying Gamma parameterization was used under response pattern 1, risk was adequately described with minimal bias using the MA procedure; further estimated coverage was 97.2% and 95.0% for BMRs of 1% and 10%, respectively. This figure graphically depicts the true underlying gamma model in relation to the average fit across all 2,000 simulations

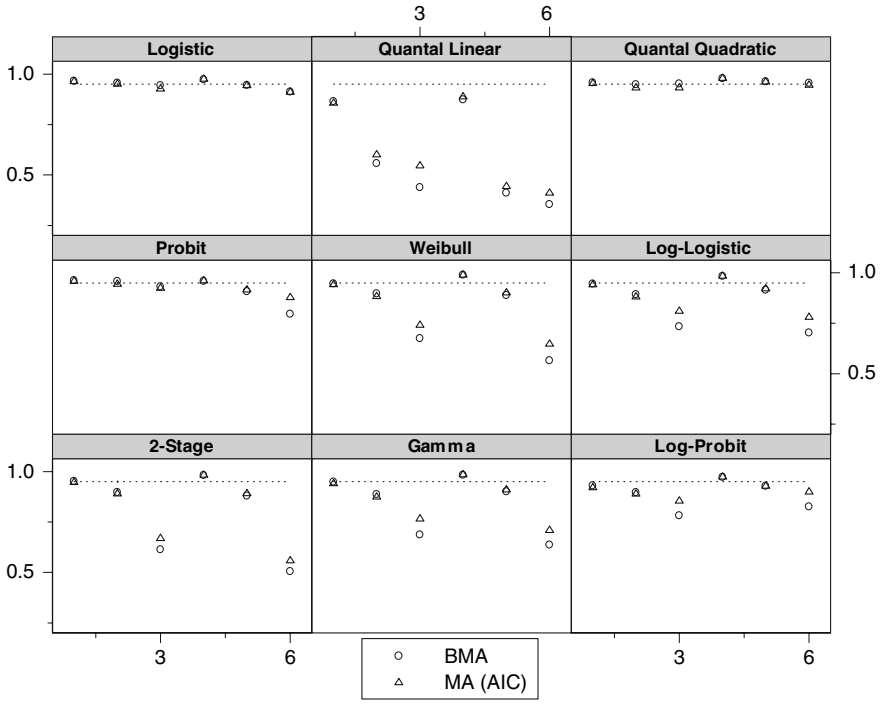

Fig. 6 Observed coverage for BMA as well as MA using the AIC under all assumed truths with BMR = 10% and a 95% confidence level. Given the true underlying dose–response models, represented by the header, the points display each method's estimated coverage of the BMD for response patterns 1–6 from Table 1

for models (1–9). In this case, the gamma model is reasonably well described by the other models, and the true dose–response model is bounded both above and below the by the models considered. This is in contrast to Fig. 9 which describes the same information for the sixth response pattern of the quantal linear model. In this situation the true dose–response is well described by some models in the model space; however, other models describe the dose–response relationship poorly. The models that describe the dose–response relationship poorly also consistently underestimate the true risk by providing BMD estimates that are greater than the true value. This effects MA by shifting BMD estimates away from the true value.

We also note that the simulations did not check the adequacy of fit of the models to the data. Many of the fits, which were used in the average, described the data poorly. Thus situations, like those described by Fig. 9, can be further exacerbated as all models, regardless of their ability to adequately describe the data, and are included in the weighted average. Though one may argue such models receive small weights, and thus affect the average minimally, it is important to note that small weights applied to risk estimates that are drastically different from truth can substantially affect the averaged values. This effect can be more substantial if all models, which poorly describe the data, provide risk estimates in the same direction.

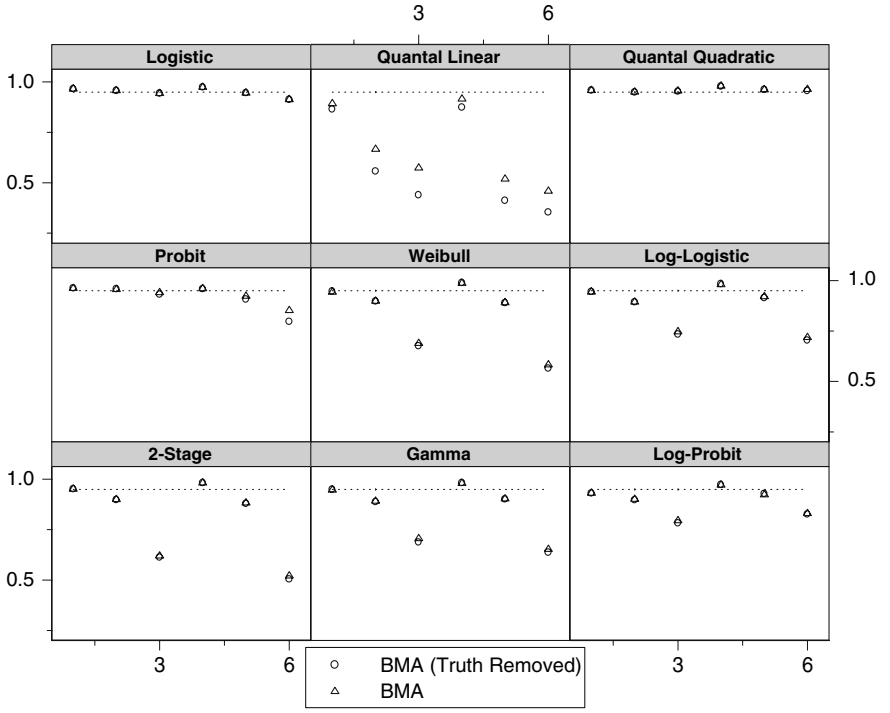

Fig. 7 Observed coverage for BMA, where truth was removed from the estimation, in comparison to the BMA, where the true model is used in the estimation, under all assumed truths with BMR = 10% and a 95% confidence level. Given the true underlying dose–response models, represented by the header, the points display each method's estimated coverage of the BMD for response patterns 1–6 from Table 1

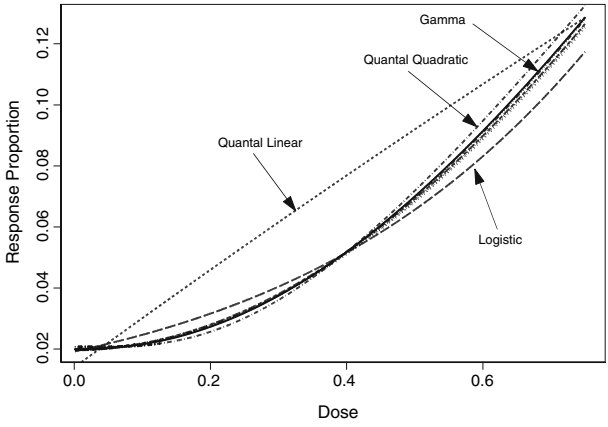

Fig. 8 True underlying gamma model in comparison to the average model fits of the other eight models considered. The figure represents a situation where the true model is well described by the other models considered. In this situation model averaging performs well, providing accurate BMD estimates in terms of bias as well as coverage

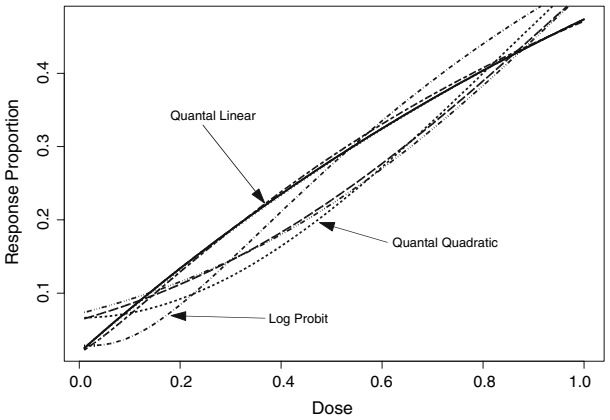

Fig. 9 True underlying quantal linear model in comparison to the average model fits of the other eight models considered. The figure represents a typical situation where the true model is not well described by the other models considered. In this situation model averaging performs poorly, providing poor BMD estimates in terms of bias as well as coverage

To illustrate, consider augmenting the model space with supra-linear dose–response models, e.g., models of the form $\pi(d_i) = \gamma + (1-\gamma)\left[1 - \exp(-\beta d_i^{\alpha})\right]$ (i.e., model 9) where the shape parameter α is fixed at the values 0.5, 0.6, 0.7, 0.8, and 0.9. Further consider an estimate based upon averaging model-specific estimates over this new model space. Finally, restrict the models averaged to those with a Pearson X^2 goodness-of-fit statistic with a P-value of 0.1 or greater, which is a common experimental practice. Table 2 reports the coverage for such a MA procedure in comparison to the coverage of the original simulation experiment for the underlying true quantal linear model. If these specific models are included in the model average, as well as models that fit the data poorly excluded (i.e., the observed statistic had a P-value ≤ 0.1), the coverage for the risk estimates reaches the nominal level. This is in stark contrast to the original simulation experiment where coverage rates of approximately 35.4% were observed for a nominally specified level of 95%.

Table 2 Table compares observed coverage for Bayesian model averaging to the true underlying risk, when model averaging is performed using two different model spaces

Response pattern	Coverage of Bayesian model averaging	
	Original simulation	Supra-linear models
1	0.864	0.961
2	0.557	0.967
3	0.439	0.959
4	0.873	0.979
5	0.411	0.939
6	0.354	0.944

6 Discussion

Model averaging provides the risk assessor a method for incorporating uncertainty into the analysis, but it does so with results dependent on the chosen model space. If the true dose–response model lies within the underlying model space, MA provides levels of bias and coverage superior to other approaches that are commonly used. It however does fail to perform adequately when the model space is chosen in a way that does not properly describe the true dose–response relationship. Consequently the inclusion of a wide variety of model curvature becomes extremely important. It also suggests that the USEPA Benchmark dose software, as it is currently implemented, is in itself inadequate to perform risk estimation using MA. This is primarily due to the default curvature available to the modeler. Further, as was demonstrated, MA does not supersede the proper use of model fit diagnostics. Residual plots, as well as other fit information, should be used to diagnose the reliability of an individual model to describe the data.

Also, in cases where the models adequately describe the data, neither the AIC or BIC based weights appeared to be superior when compared to the other. However in cases where the averaging procedure failed, the AIC based weights provided closer to nominal coverage. As there is no clear answer as to which weighting procedure is superior, we feel that dose estimates using either criterion for MA could be used.

Although this study investigated MA in application to animal toxicity studies, there is no reason to suggest that the results are not applicable to a wider range of risk estimation methods. It is important to note that model uncertainty is not overcome through the use of MA. In fact the uncertainty is shifted from the model chosen to the model space chosen. If the selected model space does not adequately describe the true dose–response relationship, then the MA procedure will exhibit poor behavior. The choice of adequate model space becomes paramount, and thus the use of supra-linear as well as sub-linear models, in effort to fully describe the underlying true model, is recommended for producing reliable risk estimates. We also note that the use of MA does not preclude the expert knowledge of the practitioner to use biologically based mechanistic models for estimating risk. Models which are based upon some mechanistic understanding of the underlying system should be preferred when such knowledge is available. Thus, content and biological understanding can be the first filter applied when selecting the model space.

Finally, the MA procedure described herein averaged excess risk estimates derived from each model's functional form, and were not based upon the average of the individual models themselves, which would correspond to an averaged dose–response model of the form $\bar{\pi}(d) = \sum_{i=1}^{k} w_i \pi_i(d)$, where w_i represents the weight for the ith model and $\pi_i(d)$ is functional form of that model. This was primarily done based upon the computational complexity of estimating risk from this model; however, calculating the corresponding BMD and BMDL from the average dose–response model may more appropriate, and may provide less bias and better coverage properties than were observed in this study. Research is being conducted to investigate the benefits of such a procedure over the current MA procedure (Wheeler and Bailer, 2007).

References

Akaike H (1978) A Bayesian analysis of the minimum AIC procedure. Ann Inst Stat Math 30:9–14
Bailer AJ, Noble RB, Wheeler M (2005a) Model uncertainty and risk estimation for quantal responses. Risk Anal 25:291–299
Bailer AJ, Wheeler M, Dankovick D, Noble R, Bena J (2005b) Incorporating uncertainty and variability in the assessment of occupational hazards. Int J Risk Assess Manage 5:344–357
Buckland ST, Burnham KP, Augustin NH (1997) Model selection: an integral part of inference. Biometrics 53:603–618
Crump KS (1984) A new method for determining allowable daily intakes. Fundemental Appl Toxicol 4:854–871
Kang SH, Kodell RL, Chen JJ (2000) Incorporating model uncertainties along with data uncertainties in microbial risk assessment. Regulat Toxicol Pharmacol 32:68–72
Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. Stat Sci 14:382–417
Raftery AE (1995) Bayesian model selection in social research. Sociol Methodol 25:111–163
Raftery AE, Madigan D, Hoeting JA (1997) Bayesian model averaging for linear regression models. J Am Stat Assoc 92:179–191
Razzaghi M, Kodell RL (2000) Risk assessment for quantitative responses using a mixture model. Biometrics 56:519–527
Schwartz G. (1978) Estimating the dimension of a model. Ann Stat 6:461–464
R Development Core Team (2005) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria
USEPA (2001) Help manual for Benchmark Dose Software. Version 1.3. Office of Research and Development, US Environmental Protection Agency, Washington, DC 20460
Wheeler M, Bailer AJ (2007) Properties of model-averaged BMDLs: a study of model averaging in dichotomous response risk estimation. Risk Anal 27:659–670

Author Biographies

Matthew W. Wheeler received a M.S. degree in Statistics and a B.S. in Computer Science and Systems Analysis, both from Miami University in Oxford, OH, USA. He is a member of the Risk Evaluation Branch at the National Institute for Occupational Safety and Health in Cincinnati, OH, USA, and is currently a doctoral student, studying biostatistics, at the University of North Carolina at Chapel Hill. His current interests include risk estimation and computational statistics.

A. John Bailer received a Ph.D. degree in Biostatistics from the University of North Carolina in Chapel Hill. He joined the faculty of the Department of Mathematics and Statistics at Miami University in Oxford, OH, USA in 1988 following a post-doctoral position at the US National Institute of Environmental Health Sciences in RTP, NC. From 1996, he has been Professor of Statistics. He is also a Co-director of the Center for Environmental Toxicology and Statistics, an affiliate faculty member of the Department of Zoology, and a senior researcher in the Scripps Gerontology Center, all at Miami University. His research interests include quantitative risk estimation, and the design and analysis of occupational and environmental health studies.