Quantifying Dose-Response Uncertainty Using Bayesian Model Averaging

Melissa Whitney and Louise Ryan

Department of Biostatistics

Harvard School of Public Health

Resources for the Future

October 22-23, 2007

1 Introduction

Regulatory agencies such as the United States Environmental Protection Agency (EPA) often face the difficult task of establishing environmental standards and regulations in contexts where substantial knowledge and/or data gaps lead to significant uncertainty regarding the right decision. Developing effective strategies for decision-making in such settings has become a major thrust for the EPA in recent years. Indeed, decision-making in the face of uncertainty has been the focus of a number of workshops and national academy reports.

Recently, Resources for the Future convened a workshop entitled *Uncertainty Modeling in Dose Response*, with the goal of addressing the specific question of how uncertainty issues can be appropriately reflected in dose response modeling. In preparation for the workshop, invited attendees were asked to tackle a variety of interrelated uncertainty issues individually, with the final goal of comparing their various approaches to uncertainty analysis at the workshop. To explore different methodologies for uncertainty analysis, four bioassay datasets were utilized, each characterizing an unique uncertainty issue. The first dataset was taken from the Benchmark Dose Software available through the EPA's website, and this dataset was utilized for its simplicity of only three dose groups and no extraneous covariates. Researchers were then left with the task of choosing the appropriate structure for the model when there are only three groups of subjects, with only 50 animals per group, and a range of 1-20 responses per group. The second problem set posed to researchers consists of two datasets, one for male and another for female rats, each exposed to four (differing) dose levels. To model this dataset, researchers must determine (1) whether separate dose-response curves are needed for each sex, and (2) to what extent the decision to combine or report separate estimates affects the overall uncertainty in risk estimation. Similarly, a third dataset presents results for two separate

outcomes, two different types of tumor formation. Researchers must decide whether to combine these endpoints for dose-response modeling or whether each possible outcome (i.e. type of tumor formed) should be modeled separately. Finally, the last dataset consists of two studies and leaves researchers to determine whether these studies should be combined for purposes of risk estimation. These four datasets were designed to demonstrate some of the recurring dilemmas researchers face, namely what structural forms should be used to model the data and whether data from different experiments, animal types, or outcome types should be combined for purposes of dose-response modeling and subsequent risk analysis.

Our strategy for tackling the challenges posed by the workshop datasets was to use Bayesian Model Averaging (BMA). A number of authors have discussed the use of BMA as an effective strategy for addressing uncertainty associated with model choice in a variety of practical settings (Viallefont et al., 2001; Brown et al., 1998; Volinsky et al., 1997; Raftery et al., 1997). BMA is appealing in the sense that quantities of interest are averaged with respect to a set of candidate models, with weights proportional to the posterior probability that each model is correct, given the observed data. Consequently, the approach gives more weight to estimates obtained from models that fit the data fairly well, while estimates corresponding to poorly fitting

models are downweighted. Because of these appealing properties, Morales et al. (2006) proposed the use of BMA as an effective tool for quantifying model uncertainty in a risk assessment setting. Their work was motivated by an analysis of some epidemiological data from southwestern Taiwan, related to cancer risks associated with exposure to arsenic in drinking water. There were a number of problems with the arsenic dataset, including the fact that it lacked individual-level exposure information and had no information related to important confounders such as smoking behavior. Morales et al. (2000) calculated Benchmark Dose (BMD) estimates (the dose that leads to a specified increase in cancer risk, compared with unexposed subjects) and reported considerable sensitivity of the results to the choice of underlying dose-response model. Depending on whether dose was modeled linearly or on the log scale, as well as whether risk was quantified on an additive or multiplicative scale, estimated BMDs for male lung cancer ranged from 42 to 70 parts per billion (ppb), with non-overlapping confidence intervals. In contrast, the BMD obtained through model averaging lead to an estimate of 60 ppb, with a slightly wider confidence interval than those obtained from individual models, reflecting the additional uncertainty due to model choice (Table 1) (Morales et al., 2006).

In this paper, we further explore the use of BMA as a tool for quantify-

ing uncertainty in the contexts of both (1) model structure choice and (2) model covariate selection for any given, particular model structure. Two data analyses are used to illustrate these concepts. Dataset I consists of EPA BMD test data for which researchers must determine an appropriate model structure choice and possible dose transformations when fitting the data. Dataset II consists of separate data for male and female subjects and poses the model covariate selection question of whether sex should be taken into account when calculating risk due to exposure. We apply multiple techniques for conducting Bayesian Model Averaging to these two datasets and subsequently use the BMA method to calculate a measure of excess risk due to exposure known as the Benchmark Dose (BMD).

1.1 Bayesian Model Averaging for Toxicological Data and Benchmark Dose Estimation

Although the above example utilized environmental epidemiological data, the flexible BMA framework can be adapted for quantal response toxicology data analysis. Using the notation of Clyde et al. (2004) and Hoeting et al. (1999), let Δ be the quantity we wish to estimate using Bayesian Model Averaging. In these analyses, we are interested in estimating a Benchmark Dose (BMD) estimate which accounts for model selection uncertainty by

incorporating information from a variety of reasonably fitting dose-response models. To estimate $\Delta = BMD$, the excess risk due to exposure, we utilize the additive risk definition of the Benchmark Dose when analyzing Datasets I and II. Thus, the BMD is defined as the dose which increases the probability of an adverse effect from P_0 in an unexposed subject to $P_0 + BMR$ where BMR is the benchmark response, a level of response deemed epidemiologically/biologically relevant (Butz-Jorgensen et al., 2000). For this analysis, we chose $BMR = 0.1$, consistent with common EPA practice (U.S. EPA, 2000). We calculated the lower bound of the BMD, the BMDL, using bootstrap methods. In particular, to calculate the BMDL for each of model, we generated 1,000 dataset samples drawn at random under a given model, calculated the BMD for each sample, and then estimated the BMDL as the lower 5% cutoff value for the BMDs calculated from the 1,000 samples.

To carryout BMA to find a model-averaged exposure-response curve and the resulting, averaged BMD estimate, we first specify a set of suitable models. We consider K total models, which are typically weighted equally a priori, i.e. $p(M_k) = 1/K$ for $k = 1, 2, ...K$. Next, we compute posterior model probabilities, $\Pr(M_k \mid Data)$, which reflect the likelihood that a model holds "true" given the observed data. The posterior model probability for

a given model, k, can be written as:

$$\begin{aligned} \Pr(M &= k \mid Data) = \frac{p(Data \mid M = k)p(M_k)}{\sum_{j=1}^{K} p(Data \mid M_j)p(M_j)} \qquad (1) \\ \text{where } p(Data &\mid M_k) = \int p(Data \mid \boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k \mid M_k)d\boldsymbol{\theta}_k, \end{aligned}$$

which follows from Bayes Theorem. In addition to calculating posterior model probabilities for each model, we use classical estimation methods to calculate an estimate of excess risk for each model run, $\widehat{\Delta}_k$.

Finally, we average the results using posterior probabilities as weights, such that poorly-fitting models get downweighted and better-fitting models contribute more strongly to the final, averaged estimate of risk. After obtaining posterior model probabilities using Equation (1), $\Pr(M_k \mid Data)$, and using classical estimation procedures to obtain a quantity of interest for each model, $\widehat{\Delta}_k$, we can then use posterior model probabilities as model weights to obtain the BMA estimates of the unconditional expectation and variance, i.e. the averaged estimate of risk over all models examined.

The unconditional expected value and variance of Δ given the posterior model probabilities, $p_k = \Pr(M_k \mid Data)$, are:

$$E(\Delta \mid Data) = \sum_{k=1}^{K} \widehat{\Delta}_k p_k, \text{ where } \widehat{\Delta}_k = E(\Delta \mid M_k, Data) \text{ and} \qquad (2)$$

$$Var(\Delta \mid Data) = \sum_{k=1}^{K} Var(\widehat{\Delta}_k \mid M_k, Data) p_k + \sum_{k=1}^{K} (\widehat{\Delta}_k - \widehat{\Delta})^2 p_k \qquad (3)$$

where K is the total number of models considered. The variance formula above separates out the variance into two components: the variance due to the estimation procedure for each individual model, $\sum_{k=1}^{K} Var(\widehat{\Delta}_k \mid M_k, Data)p_k$, and the uncertainty due to model selection, $\sum_{k=1}^{K}(\widehat{\Delta}_k - \widehat{\Delta})^2 p_k$ (Hoeting et al., 1997).

To implement Bayesian Model Averaging, one must calculate posterior model probabilities, $p_k = \Pr(M_k \mid Data)$, for each of the k models using Equation (1). However, this requires solving an an integral that is difficult to calculate in all but the simplest cases, and many posterior distributions for models used in environmental epidemiological studies do not have closed form solutions (Clyde and George, 2004). Both fully Bayesian methods as well as Frequentist approximations to posterior model probabilities have been utilized to carryout BMA. Fully Bayesian methods include

closed-form solutions (which rarely exist when modeling toxicology data) and Markov Chain Monte Carlo Methods such as Reversible Jump MCMC, Carlin & Chib Method, Stochastic Search Variable Selection, and Gibbs Variable Selection. The most common Frequentist analytical approximation used to estimate posterior model probability is the BIC approximation described below. For Dataset I, the BIC approximation to the posterior model probability is utilized. For Dataset II, both the BIC approximation and Gibbs Variable Selection are employed. These methods are not exhaustive. Researchers must assess the computational feasibility of these various techniques when determining which model averaging methods are most appropriate for a given dataset.

1.1.1 The BIC Approximation

Several analytical techniques have been used to approximate the marginal distribution for Y, $p(Data \mid M_k)$. Raftery (1995) derived the following approximation using Schwarz's Bayesian Information Criteria (BIC) (Schwarz, 1978) to approximate model posterior probabilities:

$$p(Data \mid M_k) \propto e^{-0.5BIC(M_k)},$$

and thus

$$\Pr(M_k \mid Data) \doteq \frac{p(M_k)exp(-0.5BIC(M_k))}{\sum_{k=1}^{K} p(M_k)exp(-0.5BIC(M_k))}$$

with BIC defined as $BIC(M_k) = -2\log(\text{maximized likelihood}|\ M_k) + \dim(M_k)log(n)$, where $\dim(M_k)$ is the number of parameters for M_k and n is the sample size. This approximation tends to work well in settings with independent covariates and moderate to large sample sizes (Wasserman, 2000).

1.1.2 Simulation Techniques

The most common way to calculate posterior model probabilities is to simulate data from a posterior distribution using Markov Chain Monte Carlo (MCMC) methods. The simplest simulation techniques, however, are designed for sampling from distributions of parameters with fixed dimensionality. In the model averaging framework we aim to characterize the posterior distributions of models with different numbers of parameters. Standard MCMC theory does not apply when the dimension of the parameter space

is allowed to vary, and model averaging may violate conditions necessary to ensure Markov chain convergence (Carlin and Chib, 1995). Gibbs Variable Selection (GVS) (Dellaportas et al, 2002; George and McCulloch, 1993) addresses the problem of varying sample space dimension. The Dataset II analysis utilizes GVS, and its implementation is described below in the context of covariate selection.

2 Dataset I Analysis: Uncertainty Due to Choice of Link Function/Dose-Response Model

For the analysis of Dataset I (presented in Table 2), we consider a variety of models, most of which are included in EPA's BMD software, BMDS Version 1.4.1b. We consider only non-zero background models, both for their biological plausibility and computational issues with calculating BMD for zero background dose models when $P(\text{response at } d = 0) = 0$. We fit Logit, Probit, Multistage, and Weibull models, and consider non-transformed dose d, log-transformed dose $\log(d)$, additive background effects, adding an effective dose term to the models, and simple empirical models for the data. For each of these models, we adopted the common constraints on parameter values suggested by the U.S. EPA Benchmark Dose Software (2007) and

Gart et al. (1986). In total, ten models are fit to the data (Table 3).

These ten models can be fitted easily using maximum likelihood, which involves choosing the values of the unknown parameters which maximize $ll = \sum_{j=1}^{J}\{r_j(\log(P(d_j))exp + (n_j - r_j)(\log(1 - P(d_j))\}$, where r_j is the number of responses observed among the n_j subjects exposed at the j^{th} dose level, d_j. We used the non-linear optimizing function *nlminb* in R (Version 2.6.2) to fit all models. Table 4 shows the estimated parameters for each of the 10 models and Figure 1 depicts these models graphically. We calculated the posterior probability for each of the 10 models using the BIC Approximation to posterior model probability, as described above (Table 4). Both the table and graph demonstrate that similarly well-fitting models have nearly identical posterior model probabilities. Models 1, 2, 5, and 6 provide the best fit to the data, are quite similar in structure, and have the highest shared posterior model probabilities (approximately .18 each, or over 0.70 combined posterior probability). Those models that fit the data poorly, particularly the simple logistic and probit models, models 9 and 10, have the lowest posterior probabilities of 0.02 and 0.03, respectively. Figure 1 also reveals that these models fit the data poorly.

For each model, we then calculated the BMD by solving for x, such that $p(dose = x) - p(dose = 0) = BMR$, with $BMR = 0.1$. Lower limits on x,

the BMDLs, were computed via parametric bootstrapping (Table 4). Finally, utilizing the posterior model probabilities and individual model BMDs and BMDLs reported in Table 4, we calculated a model-averaged BMD, $\Delta_{BMD} = 5.3112$, and corresponding averaged lower limit, $\Delta_{BMDL} = 2.9071$, which we calculated using equation (2). These averaged values capture and quantify the model selection uncertainty and overall variability observed in the risk estimates and the graphical depictions of the 10 exposure-response curves fit above.

3 Dataset II Analysis: Uncertainty Due to Covariate Adjustment

To analyze Dataset II (Table 5), we use Bayesian Model Averaging to address whether males and females should be combined when estimating risk due to Frambozadrine exposure, or, rather, whether separate dose/response curves are necessary. Whereas the Dataset I analysis utilized BMA to examine model structure uncertainty, the Dataset II analysis uses BMA to address model covariate selection uncertainty given a particular model structure, logistic regression. Under a traditional approach to data analysis, researchers

would (1) consider fitting two separate models versus pooling the data to form a combined model, (2) examine the model fits, and then (3) draw inferences from the "final" model(s) chosen, ignoring the alternative estimates and the uncertainty arising from the model-selection process. Past research has demonstrated that this model selection process can underestimate true variability and uncertainty and thereby result in over-confident decision-making (Draper, 1995). The goals of the Dataset II analysis are two-fold. First, we address model uncertainty arising from covariate selection by comparing two strategies for implementing model averaging. Second, we use the averaged model to calculate benchmark doses that account for the uncertainty involved in modeling Dataset II.

To carryout Bayesian Model Averaging, we approach the dataset as a covariate selection problem and compare the three (multiple) logistic regression models:

$$M_1 \quad : \quad p(response) = g(\beta_0 + \beta_1 dose)$$

$$M_2 \quad : \quad p(response) = g(\beta_0 + \beta_1 dose + \beta_2 sex)$$

$$M_3 \quad : \quad p(response) = g(\beta_0 + \beta_1 dose + \beta_2 sex + \beta_3 dose * sex)$$

As depicted above, Model 1 (M_1) combines males and females to calculate a single dose-response curve and Model 2, M_2, assumes a common dose effect, but also allows for an additive gender effect. Finally, the interaction model, M_3, is the equivalent of fitting two entirely separate dose-response curves for males and females.

Under a standard, naïve analysis, the three models would be compared using standard model selection criterion, such as the AIC. The model with the lowest AIC would be chosen as the final model from which all inferences would be made. In this case, the combined model (M_1) would be utilized with the lowest AIC value of 36.37, with no indication that researchers also considered separate dose-response curves for males versus females (Table 6). Bayesian Model Averaging can be utilized to combine information across these three models in lieu of traditional model/covariate selection methodology.

As with the analysis of Dataset I, the BIC approximation to the posterior model probability is utilized. In addition to this analytical approximation, we use a simulation technique, Gibbs Variable Selection, to conduct the model averaging and compare results. Gibbs Variable Selection proceeds by noting that regression models with various numbers of included/excluded covariates can be written as:

$$\begin{aligned} \text{logit}(p(y &= 1)) = \sum_{j=1}^{p} g(j) X_j \beta_j \ + \boldsymbol{\varepsilon} \\ \text{where} \ \ g(j) &= 1 \text{ if the } j^{th} \text{ variable is included in the model, } g(j) = 0 \text{ otherwise} \end{aligned} \tag{1}$$

Introducing the variable indicator function $g(\cdot)$ reduces the framework to one of fixed dimensionality. We can then utilize standard simulation techniques to estimate $g(\cdot)$ and $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \tau)$ for all models, M_k, $k = 1, 2, 3$. Using the following framework, GVS was implemented for Dataset II analysis using WinBUGS 14.

Likelihood :

$$\begin{aligned} Y[i] &\sim binom(p[i], n[i]) \\ \text{logit}(p[i]) &= \beta_0 + g(1) * \beta_1 * dose + g(2) * \beta_2 * sex + g(3) * \beta_{12} * dose * sex \end{aligned}$$

Priors :

$$\begin{aligned} g(j) &\backsim Bernoulli(0.5) \text{ for } j = 1, 2, 3 \\ \beta_j &\backsim N(0, \tau) \\ \tau &\backsim dgamma(0.1, 0.1) \end{aligned}$$

Gibbs Sampling first samples each variable indicator $g(j)$, then $\boldsymbol{\beta}_j$, and finally, τ. Table 6 gives posterior model probabilities for the three models of interest estimated using the GVS method in WinBUGS. The combined data model had the highest posterior probability ($p_k = 0.546$). The model which allowed for an additive sex effect had posterior probability 0.443, and the model allowing for a sex*dose interaction (i.e. the equivalent of modeling male and female data via two separate models) only had a posterior probability of 0.011.

Next, the BIC approximation was used to estimate posterior model probability. We implemented the BIC approximation method using Chris Volinsky's R BMA Package (function BIC.GLM). The final column of Table 6 gives estimated posterior model probabilities for the three models of interest estimated using the BIC Approximation. As with GVS, the BIC approximation finds that the combined model has the highest posterior probability (0.639). M_2 has posterior probability 0.266, and fitting separate models via M_3 is least-supported by the data, with posterior probability 0.095. A Comparison of the GVS method and the BIC approximation model-averaged estimates are reported in Table 7. Table 7 gives the estimated posterior probabilities for inclusion of variables in the true model, the model-averaged coefficient estimates, and the standard deviations calculated using Equations

(2) and (3) for both methods. The GVS and BIC results are quite similar in terms of model-averaged dose estimation and posterior probability of a dose effect, with probability of dose inclusion in the "true" model of approximately 1, and model-averaged coefficient values of 0.0317 and 0.0315 for dose, respectively. Nonetheless, these methods deviate with respect to inclusion of a sex term and dose-sex interaction term.

4 Discussion

We have demonstrated that benchmark dose estimates are highly dependent upon the particular dose-response curve calculated. By adopting a Bayesian Model Averaging Framework, we have accounted for the additional variability due to choosing a "final" dose-response curve, and we have used BMA to provide benchmark doses that more accurately reflect uncertainty in our understanding of the effects of exposure on the occurrence of adverse responses.

For the Dataset I analysis, model averaging performed well, as the best-fitting models all had high posterior probabilities and thus larger "weight" in the final, averaged model estimates of risk due to exposure. In addition, the models with very similar structural forms had nearly the same posterior probabilities. The top four models accounted for over 70% of the total

posterior probability and had an average BMD value of 2.673. Nonetheless, researchers must be careful in choosing which models to include for model averaging. The overall model-averaged BMD including ill-fitting, lower-weighted models was 5.311, revealing the penalty for including models structures ill-suited to the data. Naively averaging over ill-performing models may result in biologically implausible model fits and can lead to nonmonotonic curves for which risk estimates such as BMDL can not be obtained. In addition, researchers must include plausible restraints for models to ensure they perform reasonably when fitting the data. Weibull models were particularly unstable absent suitable constraints.

For the Dataset II analysis, both the BIC approximation method and the GVS method for carrying out BMA performed similarly in terms of calculating posterior model probabilities, i.e. the "weight" of the evidence in favor of modeling male and female rat data separately. In this case, small posterior probabilities for the full logistic regression model with an interaction term (0.011 using GVS and 0.095 using BIC) demonstrate that separate estimates of excess risk due to exposure are probably unnecessary. Nonetheless, the BMA approach is worthwhile because it accounts for the uncertainty arising from the decision-making process of fitting multiple models. This contrasts with the traditional approach, which ignores plausible alternatives when re-

porting results and thus may result in over-confident inferences regarding risk.

For Dataset II, the resulting model-averaged coefficient estimates and posterior probability estimates differ somewhat for GVS and BIC approximation methods. One must exercise caution in applying the BIC approximation to smaller datasets where covariates are not independent, such as in our Dataset II analysis which includes an interaction term. (Wasserman, 2000). In addition, the BIC approximation exacts a strong penalty for model complexity, giving lower posterior probability to models with increasing number of parameters. Accordingly, researchers must consider which method(s) for implementing BMA are most appropriate for a given dataset.

This paper addresses uncertainty due to model structure and model covariate selection. There are other sources of uncertainty worthy of exploration for quantal response toxicology data, such as uncertainty due to dose levels where data are sparse or risk estimate uncertainty that may arise from the presence of outliers or lack of monotonicity. As always, researchers must struggle with the question of when a formal uncertainty analysis will be applicable for a particular dataset. We have demonstrated that Bayesian Model Averaging can serve as a useful tool for accounting for several sources

of uncertainty. Ongoing work aims to utilize alternative simulation methods to calculate BMDs, BMDLs, and resulting variance estimates for both Datasets I and II. In addition, we plan to utilize Reverse Jump MCMC simulation method for implementing BMA for Dataset I in order to compare this method with the BIC Approximation to posterior model probability calculated above.

5 References

Brown PJ, Vannucci M, Fearn T. Multivariate Bayesian Variable Selection and Prediction. Journal of the Royal Statistical Society, Series B, 60:627-642 (1998).

Butz-Jorgensen E, Grandjean P, Keiding N, White RF, Weihe P. Benchmark Dose Calculations of Methylmercury-Associated Neurobehavioural Deficits. Toxicology Letters 112-113:193-199 (2000).

Clyde M. Model Averaging. In Subjective and Objective Bayesian Statistics (Editor: James Press), Wiley, 320-333 (2003).

Crump KS. A New Method for Determining Allowable Daily Intakes. Fund. Appl. Toxicol. 4:854-871 (1984).

Draper D. Assessment and Propagation of Model Uncertainty. Journal of the Royal Statistical Society, Series B, 57:45-97 (1995).

Gart JJ, Krewski D, Lee PN, Tarone RE, Wahrendorf J. The Design and Analysis of Long-Term Animal Experiments. Volume III - The design and analysis of long-term animal experiments. International Agency for Research on Cancer (IARC) Scientific Publications No. 79 (1986).

George EI, McCulloch RE. Gibbs Variable Selection via Gibbs Sampling. Journal of the American Statistical Association, 88:881-889 (1993).

Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial. Statistical Science, 14:382-417 (1999).

Morales KH, Ibrahim JG, Chen CJ, Ryan LM. Bayesian model averaging with application to benchmark dose estimation for arsenic in drinking water. Journal of the American Statistical Association, 101(473):9-17 (2006).

Morales KH, Ryan L, Kuo TL, Wu MM, Chen CJ. Risk of internal cancers from arsenic in drinking water. Environmental Health Perspectives, 108(7):655-661 (2000).

NRC. Science and Judgment in Risk Assessment: Student edition. Washington, DC: National Academy of Sciences (1996).

The R Project. R Software Version 2.6.2. Available at: http://www.r-project.org.

Raftery AE, Madigan D, Hoeting JA. Bayesian Model Averaging for Linear Regression Models. Journal of the American Statistical Association, 92:179-191 (1997).

Raftery AE. Bayesian Model Selection in Social Research (with discussion). In Sociological Methodology 1995, ed. P.V. Marsden (1996).

Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics, 6(2):461-64 (1978).

U.S. EPA. Benchmark Dose Technical Guidance Document. External Review Draft EPA/630/R-00/001. Available at: http://www.epa.gov/ncea/pdfs/bmds/BMD-External_10_13_2000.pdf (2000).

U.S. EPA. An Examination of EPA Risk Assessment Principles and Practices. EPA/100/B-04/001. Washington, DC: US Environmental Protection Agency (2004).

U.S. EPA. Benchmark Dose Software (BMDS) Version 1.4.1b. Available at: http://www.epa.gov/ncea/bmds (2007).

Viallefont V, Raftery AE, Richardson S. Variable Selection and Bayesian Model Averaging in Epidemiological Case-Control Studies. Statistics in Medicine, 20:3215-3230 (2001).

Volinsky C, Madigan D, Raftery AE, Kronmal RA. Bayesian Model Averaging in Proportional Hazard Models: Predicting the Risk of a Stroke. Applied Statistics, 46:443-448 (1997).

Wasserman L. Bayesian Model Selection and Model Averaging. Journal of Mathematical Psychology, 44(1):92-107 (2000).

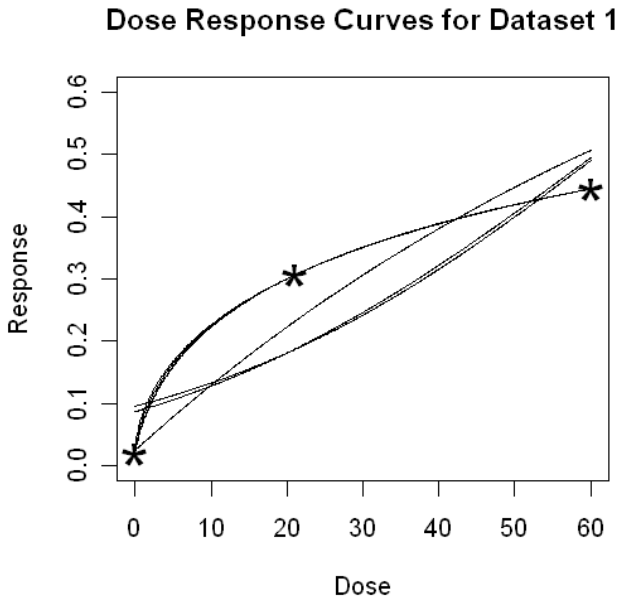

Figure 1. All 10 Models, Predicted Dose-Response Curves for Dataset I

Table 1. Bayesian Model Averaging of Arsenic Data and Resulting Averaged BMD Estimate

Individual models	BMD (95%CI)
Add, linear dose	42 (40,43)
Multi, linear dose	91 (78,107)
Multi, log dose	70 (60,83)
Model averaged	60 (47, 74)

Table 2. Dataset I: BMD Technical Guidance Data

Dose	# of Subjects	# of Responses
0	50	1
21	49	15
60	45	20

Table 3: Ten Models Considered in Dataset I Analysis

$p(d)$:	*constraints*
1. $\beta_3 + (1-\beta_3)\frac{e^{\beta_1+\beta_2\log(d)}}{1+e^{\beta_1+\beta_2\log(d)}}$	$\beta_2 > 0; 0 < \beta_3 < 1$
2. $\beta_3 + (1-\beta_3)\Phi(\beta_1 + \beta_2\log(d))$	$\beta_2 > 0; 0 < \beta_3 < 1$
3. $\beta_3 + (1-\beta_3)(1-e^{-\beta_1 d - \beta_2 d^2})$	$\beta_1, \beta_2 > 0; 0 < \beta_3 < 1$
4. $\beta_3 + (1-\beta_3)(1-e^{-\beta_1 d^{\beta_2}})$	$\beta_2 > 1; 0 < \beta_3 < 1$
5. $\frac{e^{\beta_1+\beta_2\log(d+\beta_3)}}{1+e^{\beta_1+\beta_2\log(d+\beta_3)}}$	$\beta_3 > 0$
6. $\Phi(\beta_1 + \beta_2\log(d+\beta_3))$	$\beta_3 > 0$
7. $1 - e^{-\beta_1(d+\beta_3) - \beta_2(d+\beta_3)^2}$	$\beta_1, \beta_2, \beta_3 > 0$
8. $1 - e^{-\beta_1(d+\beta_3)^{\beta_2}}$	$\beta_2 > 1; \beta_3 > 0$
9. $\frac{e^{\beta_1+\beta_2 d}}{1+e^{\beta_1+\beta_2 d}}$	$\beta_2 > 0$
10. $\Phi(\beta_1 + \beta_2 d)$	$\beta_2 > 0$

Table 4. For Each of the 10 Models Considered: Estimated Parameters, Estimated Posterior Probabilities Using BIC Approximation, Resulting Benchmark Dose for 10 Models Considered, and BMDLs Calculated Using Bootstrap Methods

Model	Model Fit ($\widehat{\boldsymbol{\beta}} = \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$)	$\Pr(M_k \mid Data)$	BMD ($\widehat{\Delta}_k$)	BMDL
1. $\beta_3+(1-\beta_3)\frac{e^{\beta_1+\beta_2\log(d)}}{1+e^{\beta_1+\beta_2\log(d)}}$	(-0.2692,0.5874,0.0200)	0.1772	2.3406	0.0287
2. $\beta_3+(1-\beta_3)\Phi(\beta_1+\beta_2\log(d))$	(-0.1685,0.3612,0.0200)	0.1772	2.8418	0.0562
3. $\beta_3+(1-\beta_3)(1-e^{-\beta_1 d-\beta_2 d^2})$	(0.6805,0.00001,0.0251)	0.0597	9.5423	7.9702
4. $\beta_3+(1-\beta_3)(1-e^{-\beta_1 d^{\beta_2}})$	(0.6805,1.0000,0.0251)	0.0597	9.5422	8.0666
5. $\frac{e^{\beta_1+\beta_2\log(d+\beta_3)}}{1+e^{\beta_1+\beta_2\log(d+\beta_3)}}$	(-0.2240,0.5685,0.0016)	0.1772	2.5798	0.0064
6. $\Phi(\beta_1+\beta_2\log(d+\beta_3))$	(-0.1413,0.3525,0.0044)	0.1772	2.9295	0.0065
7. $1-e^{-\beta_1(d+\beta_3)-\beta_2(d+\beta_3)^2}$	(0.6805,0.00001,0.0374)	0.0597	9.5423	7.9041
8. $1-e^{-\beta_1(d+\beta_3)^{\beta_2}}$	(0.6805,1.0000,0.0374)	0.0597	9.5422	7.9854
9. $\frac{e^{\beta_1+\beta_2 d}}{1+e^{\beta_1+\beta_2 d}}$	(-2.2419,2.2017)	0.0215	22.6390	19.6551
10. $\Phi(\beta_1+\beta_2 d)$	(-1.3587,1.3452)	0.0310	20.9702	18.0580

Table 5. Dataset II: Frambozadrine Data

	Dose (mg/kg-day)	# of Rats	# of Response, Hyperkeratosis
Male			
	0	47	2
	1.2	45	6
	15	44	4
	82	47	24
Female			
	0	48	3
	1.8	49	5
	21	47	3
	109	48	33

Table 6. Traditional Approach to Covariate Selection vs. BMA Approach: (1) Akaike Information Criteria for Model Selection, (2) Estimated Posterior Model Probabilities, ($\widehat{p_k} = \widehat{\Pr}(M_k \mid Data)$), Using Gibbs Variable Selection, (3) Estimated Posterior Model Probabilities Using BIC Approximation

Logistic Regression Model - Full Model:

$\Pr(response) = g(\beta_0 + \beta_1 dose + \beta_2 sex + \beta_{12} dose * sex)$

Model (k)	β_0	β_1	β_2	β_{12}	AIC	$\widehat{p_k}(GVS)$	$\widehat{p_k}(BIC)$
1. All Data Combined	-2.639	0.0313	0	0	36.37	0.546	0.639
2. Sex Effect Only	-2.531	0.0315	-0.175	0	38.04	0.443	0.266
3. Separate Models	-2.501	0.0309	-0.229	0.001	40.02	0.011	0.095

Table 7. Comparison of Two BMA Methods: BIC Approximation and MCMC - GVS Method. EV: $E(\beta \mid Data) = \sum_{k=1}^{K} \widehat{\beta}_k p_k$, where $\widehat{\beta}_k = E(\beta \mid M_k, Data)$ and p_k is posterior model probability. SD is sqrt(variance) with variance defined as above: $\sum_{k=1}^{3} Var(\widehat{\beta}_k \mid M_k, Data)p_k + \sum_{k=1}^{3}(\widehat{\beta}_k - \widehat{\beta})^2 p_k$.

Variable	GVS Method (Fully Bayesian)	BIC Approx.
Dose	Pr(dose) = 1.0	Pr(dose) = 1.0
	EV: 0.0317	EV: 0.03149
	SD: 0.00344	SD: 0.00389
Sex	Pr(sex) = 0.454	Pr(sex) = 0.292
	EV: -0.0311	EV: -0.0552
	SD: 0.2249	SD: 0.212
Dose*Sex	Pr(Interaction) = 0.011	Pr(Interaction) = 0.27
	EV: -0.00249	EV: -0.000218
	SD: 0.249	SD: 0.00289