Journal of Statistical Software

June 2008, Volume 26, Issue 5. *http://www.jstatsoft.org/*

Model Averaging Software for Dichotomous Dose Response Risk Estimation

Matthew W. Wheeler
CDC/NIOSH

A. John Bailer
Miami University and CDC/NIOSH

Abstract

Model averaging has been shown to be a useful method for incorporating model uncertainty in quantitative risk estimation. In certain circumstances this technique is computationally complex, requiring sophisticated software to carry out the computation. We introduce software that implements model averaging for risk assessment based upon dichotomous dose-response data. This software, which we call Model Averaging for Dichotomous Response Benchmark Dose (**MADr-BMD**), fits the quantal response models, which are also used in the US Environmental Protection Agency benchmark dose software suite, and generates a model-averaged dose response model to generate benchmark dose and benchmark dose lower bound estimates. The software fulfills a need for risk assessors, allowing them to go beyond one single model in their risk assessments based on quantal data by focusing on a set of models that describes the experimental data.

Keywords: bootstrapping, information criteria, model uncertainty.

1. Introduction

Risk assessors are frequently concerned in finding an exposure level which is associated with some specified level of excess risk from exposure to a hazardous agent. This level of exposure is often estimated using the benchmark dose method (BMD, Crump 1984). Here a point estimate of the dose, the BMD, or a lower limit on this endpoint (i.e., $100(1-\alpha)\%$ lower confidence limit), the benchmark dose lower bound (BMDL), is typically reported. The benchmark dose estimate is dependent on the parametric model used to fit the dose-response data, with the BMDL often being more sensitive to modeling assumptions than the BMD. Further, as one model is frequently selected to estimate the BMD, while other plausible models are ignored, model uncertainty is part of any risk assessment. This problem is well noted in the literature, and is often problematic when two or more models that describe the data similarly produce

widely varying estimates of the benchmark dose. A technique which allows one to estimate the BMD from more than one model that has been put forth in recent years, has been Bayesian model averaging (Raftery 1995), and its related frequentist analogue model averaging (MA) (Buckland, Burnham, and Augustin 1997).

Model averaging, for risk assessment, was first proposed by Kang, Kodell, and Chen (2000) for continuous microbial dose-response data, and further considered by Bailer, Noble, and Wheeler (2005a) for dichotomous dose-response data. These papers suggested a model averaged benchmark dose (MA-BMD) formed by taking a weighted average of model specific BMDs. The corresponding $100(1-\alpha)\%$ lower bounds are also derived in this manner. This method was extensively studied through computer simulation experiment by Wheeler and Bailer (2009) and this averaging procedure was frequently found to calculate a lower limit that failed to have the nominal coverage probability on the true BMD value. This failure to effectively characterize the lower limit on the BMD estimate was most evident in cases when the exposed agent was highly hazardous, e.g. a steep dose-response relationship, which lead to questions about the broader applicability of MA in risk assessment. Subsequently Wheeler and Bailer (2007) modified the above procedure to a more computationally intensive methodology. They found this modified MA procedure, which averaged individual dose-response curves instead of the individual BMDs, more effectively described both the dose-response curve as well as the given benchmark dose often yielding estimated BMDs with minimal bias and lower bound estimates that achieved coverage at the nominally specified $100(1-\alpha)\%$ level.

Although this work was promising, the computational complexity of the method makes the general use of such a method difficult to all but the most computationally sophisticated risk assessors. Consequently, this paper introduces a software program that implements the "average-model" MA procedure of Wheeler and Bailer (2007) for dichotomous risk assessment and is freely available to the risk analysis community. The paper proceeds by briefly describing this form of model averaging in the context of risk assessment applied to dichotomous dose-response data, and in the process the software is described. It then highlights the model averaging for dichotomous response benchmark dose (**MADr-BMD**) software against features and performance of the popular US EPA benchmark dose software package (**BMDS** US Environmental Protection Agency 2001).

2. Materials and methods

2.1. Modeling dichotomous dose-response data

In the context of risk assessment the term "dichotomous response data" is used to describe a two level response denoting the presence or absence of a potentially unwanted outcome (e.g., tumor response, death). Frequently, the probability of adverse response is modeled by some set of covariates. For experimental data, a single covariate corresponding to the dose administered is often considered. Dichotomous dose-response modeling describes the probability of the adverse outcome given a specific dose administered to subjects under study. In this context, parametric models are used to link a dose to the observed dichotomous response under the assumption that the observed data are binomially distributed with a probability of response related to the dose.

Many models exist that can be used in this situation. The **MADr-BMD** software package estimates a variety of models that are commonly used in risk assessment to fit dichotomous dose response data. The following list describes the dose-response models, as well as the default parameter bounds that are used, in the software package.

$$\text{logistic:} \quad \pi_1(d) = \frac{1}{1 + exp[-(\alpha + \beta \times d)]} \tag{1}$$

$$\text{log-logistic:} \quad \pi_2(d) = \gamma + \frac{(1-\gamma)}{1 + exp[-(\alpha + \beta \times d)]} \quad \begin{matrix} 0 \leq \gamma \leq 1 \\ \beta \geq 0.5 \end{matrix} \tag{2}$$

$$\text{gamma:} \quad \pi_3(d) = \gamma + \frac{1-\gamma}{\Gamma(\alpha)} \int_0^{\beta d} t^{\alpha-1} e^{-t} dt \quad \begin{matrix} 0 \leq \gamma \leq 1 \\ \alpha \geq 1, \beta \geq 0.5 \end{matrix} \tag{3}$$

$$\text{multistage:} \quad \pi_4(d) = \gamma + (1-\gamma)(1 - exp(-\theta_1 d - \theta_2 d^2 \ldots)) \quad \begin{matrix} 0 \leq \gamma \leq 1 \\ \theta_1 \geq 0, \theta_2 \geq 0, \ldots \end{matrix} \tag{4}$$

$$\text{probit:} \quad \pi_5(d) = \Phi(\alpha + \beta d) \tag{5}$$

$$\text{log-probit:} \quad \pi_6(d) = \gamma + (1-\gamma)\Phi(\alpha + \beta ln(d)) \quad 0 \leq \gamma \leq 1, \beta \geq 0 \tag{6}$$

$$\text{quantal-linear:} \quad \pi_7(d) = \gamma + (1-\gamma)(1 - exp(-\beta d)) \quad 0 \leq \gamma \leq 1, \beta \geq 0 \tag{7}$$

$$\text{quantal-quadratic:} \quad \pi_8(d) = \gamma + (1-\gamma)(1 - exp(-\beta d^2)) \quad 0 \leq \gamma \leq 1, \beta \geq 0 \tag{8}$$

$$\text{Weibull:} \quad \pi_9(d) = \gamma + (1-\gamma)(1 - exp(-\beta d^\alpha)) \quad \begin{matrix} 0 \leq \gamma \leq 1 \\ \beta \geq 0.5 \end{matrix} \tag{9}$$

Here $\pi(d)$ represents the probability of adverse response given the dose d, $\Phi(x)$ is the cumulative distribution function of a standard normal random variable at x, and $\pi_i(d) = \gamma$ when $d = 0$ for the log-logistic (2) and the log-probit (6) models. The bounds stated for the above models allow for model families having a wide range of curvature that includes both sub-linear and supra-linear behavior. These constraints also represent a wider range of curvature than available, by default, in the US EPA software. In particular the Weibull, gamma, log-probit, and log-logistic models allow a wider degree of curvature, when fitting dichotomous response data. The increased curvature is added, based upon the results of Wheeler and Bailer (2009). Here they suggest that the low-dose linear behavior is better estimated, using model averaging, given this wide range of curvature.

2.2. Model averaging

Model averaging (MA) is a statistical technique that combines estimates from different dose-response models. This is accomplished through a weighted average of the dose-response functions considered in the analysis, where the weights reflect the relation of the fitted function to the observed experimental data. For instance, a model's posterior probability can be used as a weight, and in this case this is a form of model averaging known as Bayesian model averaging (here the posterior probability represents the probability that a given model is the true model given the observed data). The weights are used to combine the models considered into one central model form.

Given the individual model weights the "averaged-model," is defined as $\hat{\pi}_{ma}(d) = \sum_{i=1}^{K} w_i \cdot \hat{\pi}_i(d)$, where $\hat{\pi}_i(\cdot)$ represents a dose-response function evaluated at the MLE, and w_i represents the weight of the i^{th} model. Conceptually the function $\hat{\pi}_{ma}(d)$ can be thought of as a smoothed dose-response, which includes information from all models considered in the analysis.

The model weights, used in the above computation, can be estimated using a variety of methods. The **MADr-BMD** software program is designed to accommodate weights formed using the AIC (Akaike 1978), the BIC (Schwartz 1978), and the KIC (Cavanaugh 1999) information criterion. These criteria are defined as $-2\log(L) + K$, and $-2\log(L)$ represents the maximum value of the -2 log likelihood for a particular model, $K = 2P$ for the AIC, $K = P\log(n)$ for the BIC, and $K = 3P$ for the KIC, here n represents the total sample size (i.e., the number of observations not the number of dose groups), and P represents the number of parameters in the model. These criteria were used in MA for risk assessment by Kang *et al.* (2000), Bailer *et al.* (2005a) Bailer, Wheeler, Dankovick, Noble, and Bena (2005b), and Moon, Kim, Chen, and Kodell (2005). Given one of these criteria the weights are formed using the following formula:

$$w_j = \frac{exp(-0.5 \cdot IC_j)}{\sum_{i=1}^{K} exp(-0.5 \cdot IC_j)}.$$

,

which was used by both Raftery (1995) and Buckland *et al.* (1997). It is important to note that three weighting criterion options are available to the **MADr-BMD** program, even though six options can be specified. These extra three options represent a modification to the AIC, BIC and KIC; we call these the AICB, BICB and the KICB respectively, and they are based upon the number of non-bounded parameters in the model. It is often the case that some of the models parameters are estimated at their lower, or upper, bounds, and in this case the effective number of parameters in the model is reduced. Although it can be argued that there is still the original number of parameters in the model, the effective number of parameters may be a better estimate than the total number of parameters. Though the effective number of parameter method is the default behavior of the US EPA **BMDS** software it is not clear which construction is more correct. Consequently the **MADr-BMD** software provides both estimates for exploration and further research; however, all results presented below are formed using the full number of parameters originally specified in the model. Thus we recommend using this approach, and leave the other three options for further research.

2.3. Benchmark dose estimation

Given the model-averaged dose-response curve $\hat{\pi}_{ma}(d)$, the benchmark dose (BMD) is defined as the dose that increases the risk above the background rate by some predefined level. Excess risk, for dichotomous data, is represented by the probability of adverse response above the background response rate. The value, which represents a specified increase in the probability of response, is known as the benchmark response (BMR), and uniquely determines the benchmark dose; the BMD is defined as the dose d that satisfies the equation:

$$BMR = \frac{\hat{\pi}_{ma}(d) - \hat{\pi}_{ma}(0)}{1 - \hat{\pi}_{ma}(0)},$$

where the BMR is typically set at values of 1%, 5% and 10%. The above formulation is known as the extra risk specification of the BMD, and can be thought of as the dose that increases the probability of response, or risk, above background by the BMR, given that the

event would not occurred in the absence of exposure. An alternative specification, also known as the added risk, is the dose d that satisfies the equation:

$$BMR = \hat{\pi}_{ma}(d) - \hat{\pi}_{ma}(0),$$

which represents the absolute increase in risk, relative to no-exposure. The software package **MADr-BMD** can estimate either risk type of the BMD specified above.

Although the MA-BMD value can be readily estimated using the above formula, no such closed form formula exists in calculating the model averaged benchmark dose lower bound (MA-BMDL). Consequently the **MADr-BMD** calculates this value using a parametric bootstrap (Efron and Tibshirani 1993). Here parametric bootstrap resamples are taken for $\hat{\pi}_{ma}(d)$, and the calculated $100(1-\alpha)\%$ lower bound is obtained using the bias corrected and adjusted confidence limits (BCa). The parametric bootstrap assumes that the response, at each dose d, is distributed binomially having success probability $\hat{\pi}_{ma}(d)$ given n trials specified by the original experiment (note that it is possible, though rare, for $\hat{\pi}_{ma}(d)$ to be equal to zero, specifically at the background dose, in this case the bootstrap proceeds by generating zero positive responses for the dose d) . The acceleration constant, which is required when computing the BCa, is estimated through the jackknife procedure also described by Efron and Tibshirani (1993).

MADr-BMD computes these estimates (i.e., the MA-BMD and its corresponding MA-BMDL) automatically. It also assumes a monotonic increasing dose-response, and in cases where shallow or no dose-response is evident the software can produce BMD estimates that are greater than the maximum dose administered. Further, in cases of flat or negative dose-response the software will produce an arbitrarily large BMD estimate of 10^8. As these estimates are often far outside of the doses administered care should be taken on the risk assessor's part on interpreting the output when very shallow dose-response data are observed.

2.4. Implementation details

Estimation of the MA-BMD and the corresponding MA-BMDL is computationally intensive requiring large amounts of CPU time. Consequently, the **MADr-BMD** software is written in C++ and compiled into machine readable code as a stand-alone command line program that is executable under the Microsoft Windows operating system's shell command prompt. The program itself utilizes many routines that are under public license or in the public domain. The maximum likelihood estimation is done using the Fortran subroutine `dmngb` (Dennis, Gay, and Welsch 1981), which is available at `http://www.netlib.org/`. This routine is the same routine used within the US EPA's benchmark dose software. Further the C++ function, which calls `dmngb` routine in the benchmark dose software, is borrowed from the source code of the US EPA's software, and is available at the US EPA's web site (`http://www.epa.gov/ncea/bmds/`). The GNU scientific library **GSL** Galassi, Davies, Theiler, Gough, Booth, and Rossi (2005), was used for all other numerical routines (e.g., the matrix algebra and numerical integration routines were **GSL** algorithms) except those involved in differentiation. In this case, a finite difference algorithm (Press, Teukolsky, and Flannery 1992), which was borrowed with slight modifications from the EPA's benchmark dose software (US Environmental Protection Agency 2001), was used. The finite difference algorithm was tested by applying it to a range of functions. These test functions included Gaussian, probit and Weibull dose-response functions with known parameters, i.e. functions whose derivatives

were known. In all cases the numerical derivative produced by the algorithm agreed with a high degree of precision, e.g. an absolute difference $< 10^{-6}$, to the known derivatives. The algorithm was further tested within the context of maximum likelihood estimation. Here 2000 randomly simulated data sets were generated and models (1)–(9) were fit using the finite difference algorithm and compared to the US EPA's **BMDS** suite. In most cases the likelihood evaluated at the solution to the maximized likelihood estimating equations was identical to that based on the **BMDS** software, and differences, when they occurred, were small with the test algorithm and the **BMDS** software outperforming the other approximately 50% of the time.

3. Software specifications

The **MADr-BMD** program is invoked through the command line argument `madrbmd`, taking its input through a user specified text file. Given no input arguments (i.e., typing `madrbmd` at the command prompt) it assumes an input file named `input.txt` is located in the present working directory. Alternative input files can be specified as an additional command line argument that names file. For example, invoking the program with the command `madrbmd ethychlor.txt` would prompt the program to take its input from the file `ethychlor.txt`.

```
250 1e-8 1e-8
0 0 1 1 1 1 1 1 1
101
2 1 0.1
0.95 5000 0
2
4
0         50      0
22.5      50      3
45.0      50      10
90.0      50      29
```

The structure of the input file is relatively straightforward and is outlined above, where the meaning of each line completely described in Table 1. The file specifies which models, out of the models (1)–(9), to include in the analysis, as well as the MA weighting criterion (i.e., AIC, BIC, KIC etc.), and the specifics on the benchmark dose calculation. Individual analyses can be specified by only including one model in the specification. In these cases, the benchmark dose, and the corresponding lower bound, represents estimates based upon the single model and the lower bound based on the bootstrap. Here the choice of the weighting criterion is irrelevant to the calculation, as the single model will always receive 100% of the weight.

MADr-BMD, by default, outputs all calculations to the command prompt; however, output can be saved to a file using redirection at the command prompt. For example the command `madrbmd ethychlor.txt > output.txt` would take data from the file `ethychlor.txt` and output the data to the file `output.txt`. The output can then be viewed using any text editor. Partial output from the **MADr-BMD** program is shown below.

```
Model Fit Statistics
-------------------------------------------------------------------------
Model                    Weight  -2log(L)           AIC            BIC
-------------------------------------------------------------------------
Multistage               0.157     140.77          146.77          156.66
Logistic                 0.094     143.79          147.79          154.38
Probit                   0.164     142.68          146.68          153.28
Weibull                  0.157     140.77          146.77          156.67
Log-Probit               0.129     141.16          147.16          157.06
Log-Logistic             0.148     140.89          146.89          156.79
Gamma                    0.151     140.85          146.85          156.74
-------------------------------------------------------------------------

----------------------------------------
'Average-Model' Benchmark Dose Estimate
----------------------------------------
Nominally Specified Confidence Level:0.950
Weighting Criterion: AIC
BMD Calculation: Added Risk
BMR: 0.100000
BMD: 31.553854847521
BMDL(BCa):19.104950496818
BMDL(Percentile):23.954508771702
Acceleration: -0.038029
Bootstrap Resamples: 5000
Random Seed: 101
----------------------------------------

----------------------------------------
'Average-Model' Goodness of Fit Test
----------------------------------------
Test Statistic   : 0.237900
Bootstrap P-Value: 0.780800
----------------------------------------
```

The output itself describes fit statistics of each model, along with the computed MA benchmark dose output. The model weights formed based upon the user specified weighting criterion, as well as the -2 log likelihood the BIC and AIC (computed using the full number of parameters) are shown in the initial output table created by the program. The subsequent output describes the calculation of the MA-BMD, its corresponding lower bound, as well as information that corresponds to its calculation. Specifically the term labeled "acceleration" represents the estimated acceleration constant used in the BCa calculation, and the random seed number is the value passed to the random number generator when computing the bootstrap resamples. Both values are both provided for transparency in the calculation. Specifying this value allows the researcher to replicate the results sin the future. Not surprisingly, if the same random number seed is not used, these values may change slightly each time the program is executed. Finally the goodness of fit statistic, i.e. $\sum \frac{(0-E)^2}{E}$, is estimated with and

`250 1e-8 1e-8`	Maximum number of iterations, rel. convergence, general convergence
`0 0 1 1 1 1 1 1 1`	MA specification: `1` = included, `0` = not-included Model order: (Quantal-linear, Quantal-quadratic, Multistage Logistic, Probit, Weibull, Log-probit, Log-logistic, Gamma)
`101`	Random Seed (specifying `0` implies current clock time will be used)
`2 1 0.1`	Averaging criterion (`1` = BIC, `2` = AIC, `3` = KIC, `4` = BICB, `5` = AICB, `6` = KICB), Risk type (`1` = added risk, `2` = extra risk), BMR
`0.95 5000 0`	Type I error rate, Number of bootstrap resamples, output bootstrap resamples (`0` = no, `1` = yes)
`2`	Degree of multistage polynomial.
`4`	Number of data lines

`0.0 50 0`
`22.5 50 3`
...
	Data specification:
	dose, number of experimental units, number of observed responses

Table 1: Description of each line of input when specifying the model average (MA) dose response analysis in the **MADr-BMD** program. It is important to note that the degree of the multistage polynomial must be specified regardless of its inclusion in the model averaging. Further no error checking is done on this value, and thus one can specify a polynomial that has more parameters than there are data, which may cause unpredictable results.

corresponding significance level estimated from the bootstrap resamples assuming the function $\hat{\pi}_{ma}(d)$ generated data. The program produces further output, not reported here, which includes model-specific parameter estimates as well as the individual bootstrap resamples.

4. Software comparison

4.1. Software features

Table 2 compares the basic feature set of the US EPA **BMDS** and the **MADr-BMD** software programs for fitting dichotomous dose-response data, and highlights the similarities, and differences between the two programs. Both packages rely on the same optimizer, and thus produce nearly identical results in most situations. In the situations when the two methods do not give identical parameter estimates, a difference in initial estimates along with the existence of local maxima, within the likelihood of the given model being fit, is usually the cause. A small simulation study by Wheeler and Bailer (2007) showed that there was no clear advantage between either the **BMDS** software and the **MADr-BMD** optimization strategy. In most all cases, the software converged to the same value, and when the two failed to achieve the same results, the larger likelihood value was found approximately half the time in the US EPA's software, and half the time in the **MADr-BMD** package.

Table 2 also highlights the fact that the **MADr-BMD** package can be used to fit individual

	MADr-BMD	US EPA **BMDS**
Dichotomous models fit	Models (1)–(9)	Models (1)–(9)
Parameter bounds*	Gamma: $0 \leq \gamma \leq 1$, $\alpha \geq 0.5, \beta \geq 0$ Weibull: $0 \leq \gamma \leq 1$, $\alpha \geq 0.5, \beta \geq 0$ Log-probit: $0 \leq \gamma \leq 1$, $\beta \geq 0$ Log-logistic: $0 \leq \gamma \leq 1$, $\beta \geq 0.5$	Gamma: $0 \leq \gamma \leq 1$, $\alpha \geq 1, \beta \geq 0$ Weibull: $0 \leq \gamma \leq 1$, $\alpha \geq 1, \beta \geq 0$ Log-probit: $0 \leq \gamma \leq 1$, $\beta \geq 1$ Log-logistic: $0 \leq \gamma \leq 1$, $\beta \geq 1$
BMDL construction	Parametric bootstrap**	Profile likelihood
Maximization algorithm	`dmngb`	`dmngb`, `donlp2`
Numeric algorithms	**GSL** 1.7	**CDFlib**, **BLAS**, **LAPACK**
Primary interface	Command prompt-text input file	Windows shell program/ command prompt
Batch processing	Available	Available
Model averaging	Available	Not available

Table 2: The table compares the available features of the **MADr-BMD** and US EPA benchmark dose software programs. * Only models that have different default parameter bounds are described. Further the **BMDS** bounds are the default bounds and can be modified by the user. ** BMDLs for individual models can be constructed by choosing a single model (i.e., the model of interest) in the model average.

models. This is done, as described previously, by simply including the desired model in the model specification line of the input file, while excluding all other models. If this is done the estimated BMD will be based upon the maximum likelihood, of the desired model, while the $100(1-\alpha)\%$ BMDL will be estimated through parametric bootstrap. This is an important difference from the method used by the US EPA's software; in that they use the method of profile likelihoods to compute the lower bound. BCa and profile methods yield BMDLs with similar coverage properties; however, the value of these estimated lower limits may differ, with the difference more pronounced with small sample sizes. Finally, we note that only the **MADr-BMD** package can calculate MA-BMD estimates.

4.2. Data example

Consider the National Toxicology Program study where male F344/N rats were exposed to 2,4-hexadienal via gavage (National Toxicology Program 2001). For this study four groups of 50 rats were exposed to one dose of 2,4-hexadienal for two years. The observed proportion of rats exhibiting squamous cell papilloma of the fore stomach was 0/50, 3/50, 10/50, and 29/50, which corresponded to hexadienal doses of 0, 22.5, 45, and 90 mg/kg/day.

Table 3 describes the estimated BMDs computed from models (1)–(9) for a BMR of 1% and 10%, using the added risk BMD specification, where the estimates were computed using both the US EPA's BMD software and the **MADr-BMD** package. A model average using the AIC as a weighting criterion and all the models except the quantal-linear (7) and quantal-quadratic

	10% BMR				1% BMR			
	MADr-BMD		**BMDS**		**MADr-BMD**		**BMDS**	
	BMD	BMDL	BMD	BMDL	BMD	BMDL	BMD	BMDL
Quantal linear	15.5	12.1	15.5	12.1	1.5	1.2	1.5	1.2
Quantal quadratic	31.1	27.3	31.1	27.5	9.6	8.4	9.6	8.5
Multistage	30.1	22.6	30.1	20.1	8.1	2.2	8.1	2.3
Logistic	39.0	32.3	39.0	32.7	8.2	5.4	8.2	5.5
Probit	36.4	29.9	36.4	30.4	8.1	5.0	8.1	5.1
Weibull	30.2	20.1	30.2	20.2	8.9	3.4	8.9	3.8
Log-probit	29.9	21.2	29.9	22.3	13.5	7.0	13.5	7.8
Log-logistic	30.4	21.2	30.4	21.9	10.7	4.6	10.7	5.1
Gamma	30.1	19.8	30.1	21.3	10.7	3.6	10.7	4.4
Model averaging	31.6	19.1	N/A	N/A	9.7	3.5	N/A	N/A

Table 3: Benchmark dose comparisons for squamous cell papillomas in rats exposed to 2,4-hexadienal.(16) Data were analyzed using both the **MADr-BMD** program and the US EPA **BMDS** software, with benchmark responses of 1% and 10%. The added risk BMD formulation was used for all calculations. Finally for the model averaging data were analyzed with the software package **MADr-BMD**, using the AIC as weights and with the BCa lower bound confidence interval reported (see Table 2 for the models and the weights used in the model average).

(8) models is also reported. For both packages, the estimated BMD for models (1)–(9) are nearly identical. The MADr-BMD BMDL estimates tend to be slightly less than the BMDL values estimated using the corresponding profile likelihood counterparts calculated using the US EPA software. These differences are most noticeable in the Weibull, gamma, log-probit, and log-logistic. Here supra-linear fits are allowed, which frequently produce BMDL much smaller than the estimated BMD, and results in **MADr-BMD** BMDL estimates that are frequently smaller than the US EPA **BMDS** by an order of magnitude, though this behavior is not observed in this example.

4.3. Coverage comparison

We also consider the differences of model averaging using **MADr-BMD** compared to fitting the true model using the US EPAs benchmark dose software. For this comparison the true dose-response model is assumed to be known; though this is something that is seldom, if ever, the case in practice. The comparison is instructive as it shows that model averaging can frequently estimate the $100(1-\alpha)\%$ lower bound of the BMD at a rate which is similar to that which would be obtained using the true dose-response model.

The coverage data for the true model fits, using the US EPA software, was obtained from the study of Wheeler and Bailer (2009), where the MA-BMDL coverage was obtained from the

subsequent study of Wheeler and Bailer (2007). These studies were preformed using identical simulation conditions. In these studies, dichotomous dose-response experiments with 4 dose groups and 50 experimental units per dose group were simulated 2000 times and the BMD/BMDL was recorded. In these simulations, models (1)–(9) were used as the true underlying dose-response forms, and for each model six unique parametric forms were used given the underlying model, and the observed coverage (i.e., $\mathsf{P}(BMDL \leq BMD_{\text{true}})$) was reported for the BMDL estimated from the true model, as well as the observed coverage from two different "average-model" MA-BMDLs families of models. The first model space, from which the MA was constructed, consisted of three flexible models which included the 2-stage multistage (4), the log-probit (6), and the weibull (9). The second model space consisted of seven models adding the quantal linear (7), the quantal quadratic (8), the probit (5), and the logistic (1) models to the three model space. Finally, we note that even though the **MADr-BMD** software program can calculate BCa MA-BMDL estimates, the lower bound estimates described were calculated using the percentiles of the bootstrap distribution due to the increased computational demand required for the BCa estimate and the size of the simulation. For a comprehensive review of the simulation design, as well as other implementation details, the reader is referred to the aforementioned manuscripts (Wheeler and Bailer 2009, 2007).

Table 4 compares the observed coverage of the BMDL, with a BMR of 10%, when the true model is fit as compared to using model averaging BMDL calculation. Further, the nominally specified coverage rate was 95% for all calculations. With the exception of the quantal linear case, model averaging using the **MADr-BMD** software performed similarly in terms of observed coverage when calculating the BMDL. In some cases, the log-probit case in particular, MA's performance was superior to knowing and fitting the true model. Potential causes for this observed behavior warrant further investigation.

5. Conclusions

The software package **MADr-BMD** gives risk assessors increased flexibility when estimating risk from dichotomous dose-response data. By combining estimates from multiple models, MA gives researchers and regulators alike a method that accounts for statistical variability as well as model uncertainty. This method has been shown by Wheeler and Bailer (2007) to yield point estimates with minimal bias and confidence limits with nominal coverage properties, while producing estimates that are superior to picking one single model. This is true even if this model describes the data "better" than the other models used. Further, as shown above, in most cases model averaging performs similarly, in terms of observed coverage, to actually knowing the true model. The **MADr-BMD** software package thus fills a need within quantitative risk assessment community. By allowing researchers to estimate risk, and corresponding benchmark doses readily, researchers will be less encumbered by difficult questions pertaining to which single model should be used, and instead will be able to focus, on the far more manageable question "what curvature is plausible, given the current state of knowledge on the hazard, and what models should be fit prior to averaging?"

With this said, it is important to note that the **MADr-BMD** software package does not constitute a "magic-bullet" when estimating risk from dichotomous dose-response data. Many questions should be addressed with further research, and care should be taken when using values estimated from the software package. For example, research questions about the best model space to use, as well as diagnostic methods for the appropriateness of the BCa cal-

			True	7-model MA	3-model MA		True	7-model MA	3-model MA		True	7-model MA	3-model MA
Condition 1	Quantal	linear	0.96	0.83	0.90	Log-probit	0.92	0.91	0.94	Gamma	0.93	0.95	0.96
Condition 2			0.94	0.78	0.90		0.90	0.93	0.95		0.92	0.94	0.95
Condition 3			0.95	0.83	0.91		0.90	0.94	0.96		0.95	0.91	0.94
Condition 4			0.98	0.88	0.95		0.95	0.97	0.99		0.97	0.97	0.99
Condition 5			0.95	0.81	0.93		0.88	0.99	0.98		0.94	0.98	0.97
Condition 6			0.95	0.80	0.91		0.89	0.99	0.98		0.94	0.94	0.94
Condition 1	Quantal	quadratic	0.96	0.94	0.96	Probit	0.96	0.95	0.97	Logit	0.93	0.97	0.98
Condition 2			0.95	0.96	0.96		0.94	0.98	0.96		0.92	0.99	0.98
Condition 3			0.94	0.98	0.96		0.93	0.98	0.95		0.93	0.99	0.96
Condition 4			0.97	0.98	0.99		1.00	0.96	0.99		0.97	0.96	0.98
Condition 5			0.95	1.00	1.00		0.95	0.98	0.96		0.92	0.99	0.97
Condition 6			0.95	0.99	0.96		0.93	0.98	0.95		0.93	0.98	0.94
Condition 1	Weibull		0.93	0.94	0.95	Multistage	0.97	0.94	0.96	Log-logit	0.94	0.93	0.95
Condition 2			0.93	0.94	0.95		0.98	0.94	0.94		0.93	0.93	0.94
Condition 3			0.94	0.89	0.92		0.94	0.86	0.91		0.95	0.93	0.95
Condition 4			0.98	0.98	0.99		1	0.98	0.99		0.98	0.97	0.99
Condition 5			0.94	0.98	0.97		0.98	0.97	0.96		0.95	0.98	0.97
Condition 6			0.94	0.92	0.93		0.95	0.88	0.91		0.94	0.96	0.95

Table 4: Comparison of observed coverage (i.e., $\mathsf{P}(BMDL \leq BMD_{\text{true}})$) between two different model averaging(MA) model spaces and the benchmark dose lower bound (BMDL) formed from fitting the true model to the data. The above data compares simulation results from the two studies of Wheeler and Bailer (2009, 2007). Here the true models were fit using the US EPA **BMDS** software, and the model averaging was conducted using a modified version of the **MADr-BMD** package with the AIC used as a weighting criterion. Further the MA-BMDL was computed using percentile based confidence intervals.

culation in apparent shallow dose-response curves, are still open. Further, by incorporating uncertainty, the software does not immediately give one the license to estimate BMDs based upon an "averaged-model" that constitute low dose extrapolations beyond the range of the data. This was best illustrated in Wheeler and Bailer (2007) In their study, the model space directly influenced the ability of the "averaged-model" to recover the BMD and BMDL, for the logistic and probit conditions. Here the 3-model average covered the true BMD at the nominal 95% level when the BMR was set to 10%, but failed to cover the true BMD when the BMR was set at 1%. However the 7-model average condition covered, at the nominal level, in both situations. This was primarily because the fact that the true model was included in the 7-model condition, which stabilized the lower bound estimate. Finally, the current implementation assumes an underlying binomial response distribution. Extensions to over dispersed responses could be considered, and it is hoped that the software may provide a template for this extension.

Finally note that the development of this software does not remove the need for expert judgment on the part of the researcher when conducting an analysis using the **MADr-BMD** software. The researcher should carefully choose the model space, where mechanistic model knowledge should be used if available, and should approach low-dose extrapolations with considerable care when using model averaging. The software advances risk assessors ability to address model uncertainty, but should not be used as a substitute for scientific reasoning.

Acknowledgments

The authors would like to thank Woody Setzer, Bob Noble, Raghupathy Ramanathan and two anonymous referees for their comments and suggestions on an earlier version of this manuscript.

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.

References

Akaike H (1978). "A Bayesian Analysis of the Minimum AIC Procedure." *Annals of the Institute of Statistical Mathematics*, **30**, 9–14.

Bailer AJ, Noble RB, Wheeler MW (2005a). "Model Uncertainty and Risk Estimation for Quantal Responses." *Risk Analysis*, **25**, 291–299.

Bailer AJ, Wheeler MW, Dankovick D, Noble R, Bena J (2005b). "Incorporating Uncertainty and Variability in the Assessment of Occupational Hazards." *International Journal of Risk Assessment an Management*, **5**, 344–357.

Buckland ST, Burnham KP, Augustin NH (1997). "Model Selection: An Integral Part of Inference." *Biometrics*, **53**, 603–618.

Cavanaugh JE (1999). "A Large-Sample Model Selection Criterion Based on Kullback's Symmetric Divergence." *Statistical Probability Letters*, **42**, 333–343.

Crump KS (1984). "A New Method for Determining Allowable Daily Intakes." *Fundemental and Applied Toxicology*, **4**, 854–871.

Dennis JE, Gay DM, Welsch RE (1981). "Algorithm 573 – An Adaptive Nonlinear Least-Squares Algorithm." *ACM Transactions on Mathematical Software*, **7**, 369–383.

Efron B, Tibshirani RB (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Galassi M, Davies J, Theiler J, Gough Band Jungman G, Booth M, Rossi F (2005). ***GNU Scientific Library*** *Reference Manual*. 2nd edition. Network Theory Ltd, 15 Royal Park, Bristol BS8 3AL, UK.

Kang SH, Kodell RL, Chen JJ (2000). "Incorporating Model Uncertainties Along with Data Uncertainties in Microbial Risk Assessment." *Regulatory Toxicology and Pharmacology*, **32**, 68–72.

Moon H, Kim HJ, Chen JJ, Kodell RL (2005). "Model Averaging Using the Kullback Information Criterion in Estimating Effective Doses for Microbial Infection and Illness." *Risk Analysis*, **25**, 1147–1159.

National Toxicology Program (2001). "Toxicology and Carcinogenisis Studies of 2,4-hexadienal in F344/N Rats and B6C3f Mice." *Technical Report TR 509*, NTP TR 509.

Press WH, Teukolsky S Aand Vetterling WT, Flannery BP (1992). *Numerical Recipes in C: Tthe Art of Scientific Computing*. 2nd edition. Cambridge University Press., Shaftsbury Road. Cambridge. CB2 2RU. UK.

Raftery AE (1995). "Bayesian Model Selection in Social Research." *Sociological Methodology*, **25**, 111–163.

Schwartz G (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, **4**, 461–464.

US Environmental Protection Agency (2001). *Help Manual for* ***Benchmark Dose Software*** *Version:1.3*. US Environmental Protection Agency, Research Triangle Park, NC, EPA 600/R-00/014F. URL http://www.epa.gov/ncea/bmds/.

Wheeler MW, Bailer AJ (2007). "Properties of Model-Averaged BMDLs: A Study of Model Averaging in Dichotomous Risk Estimation." *Risk Analysis*, **27**, 659–670.

Wheeler MW, Bailer AJ (2009). "Comparing Model Averaging with Other Model Selection Strategies for Benchmark Dose Estimation." *Environmental and Ecological Statistics*. Forthcoming.

Affiliation:

Matthew W. Wheeler
National Institute for Occupational Safety and Health
Education and Information Division
Risk Evaluation Branch
4686 Columbia Pkwy, MS C-15
Cincinnati, OH 45226-1998, United States of America
E-mail: MWheeler@cdc.gov

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 26, Issue 5 *Submitted:* 2007-04-20
June 2008 *Accepted:* 2008-02-04