Continuous model averaging for benchmark dose analysis: Averaging over distributional forms

Peer-reviewed author version

Wheeler, Matthew W.; Abrahantes, Jose Cortinas; AERTS, Marc; Gift, Jeffery S. & Davis, Jerry Allen (2022) Continuous model averaging for benchmark dose analysis: Averaging over distributional forms. In: ENVIRONMETRICS, 33 (5) (Art N° e2728).

DOI: 10.1002/env.2728
Handle: http://hdl.handle.net/1942/37504

Published in final edited form as:

Environmetrics. 2022 August ; 33(5): . doi:10.1002/env.2728.

Continuous Model Averaging for Benchmark Dose Analysis: Averaging Over Distributional Forms

Matthew W. Wheeler[1], **Jose Cortinas**[2], **Marc Aerts**[3], **Jeffery S. Gift**[4,*], **J. Allen Davis**[5,*]

[1]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, RTP, NC, USA

[2]European Food Safety Authority

[3]Center for Statistics, Hasslet University

[4]National Center for Environmental Assessment,US Environmental Protection Agency, RTP, NC, USA

[5]National Center for Environmental Assessment, U.S. Environmental Protection Agency, Cincinnati, OH, USA

Abstract

When estimating a benchmark dose (BMD) from chemical toxicity experiments, model averaging is recommended by the National Institute for Occupational Safety and Health, World Health Organization and European Food Safety Authority. Though numerous studies exist for Model Average BMD estimation using dichotomous responses, fewer studies investigate it for BMD estimation using continuous response. In this setting, model averaging a BMD poses additional problems as the assumed distribution is essential to many BMD definitions, and distributional uncertainty is underestimated when one error distribution is chosen a priori. As model averaging combines full models, there is no reason one cannot include multiple error distributions. Consequently, we define a continuous model averaging approach over distributional models and show that it is superior to single distribution model averaging. To show the superiority of the approach, we apply the method to simulated and experimental response data.

Keywords

Bayes Factors; Dose-Response Analysis; Distributional Uncertainty; Quantitative Risk Analysis

Correspondence: Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, RTP, NC, USA, matt.wheeler@nih.gov.

*Equally contributing authors.

Supporting Information

For additional information on the simulation, results and modeling choices, we refer the interested reader to the Web Supplement. The R software used in the simulation and example is available at https://github.com/NIEHS/ToxicR/releases/tag/v1.0.0.

1 | INTRODUCTION

Model averaging (MA) [1, 2, 3, 4, 5, 6, 7, 8] is rapidly becoming the gold standard dose-response modeling technique. The National Institute for Occupational Safety and Health, the World Health Organization (WHO) and the European Food Safety Authority (EFSA) recommend using model averaging as a default for quantitative risk assessments [9, 10, 11]. The use of MA is appealing because it incorporates information across multiple models to account for model uncertainty. That is, when multiple models are fit to data, it allows for multi-model inference. In most cases, this provides for model-derived estimates of the point of departure, like the benchmark dose (BMD), to be more accurately estimated. Ignoring this uncertainty results in less than optimal BMD estimates[12], and MA methodologies improve these operating characteristics.

For dichotomous dose-response data, literature exists that investigates the performance of MA, see (Kang et al. [4], Wheeler and Bailer [5], Piegorsch et al. [6], Simmons et al. [7] and Wheeler et al. [8], and references therein). These studies show that MA outperforms single model selection approaches, but despite advancements in MA for dichotomous data, few studies exist that develop MA for continuous data. To our knowledge, only Shao and Gift [13] studied MA in this setting (Varewyck and Verbeke [14] and Shao and Shapiro [15] describe software for continuous response MA but do not provide information on performance). Shao and Gift's study was small and only considered cases where the distributional assumption was known (i.e., the studies assumed normal variance in the simulation and the model average) and was similar to dichotomous methodologies focusing on MA over the mean response (i.e., every model assumed a normal distribution). If the normal distribution does not adequately represent the data, many distributional assumptions (random error in the model) other than the normal may produce better BMD estimates, and MA may further improve reliability.

For BMD estimation, some continuous response BMD definitions (e.g., the hybrid definition Crump [16]) use the tails of the distribution, and when MA ignores distributional assumptions, it may not fully capture the uncertainty in the data, which was the conclusion of Shao et al. [17]. In that work, it was shown that there was a difference in BMD estimates when considering the tails of the distribution (e.g., in the hybrid case), but not when using the relative deviation definition, i.e., which uses the mean. Additionally, Wheeler et al. [18], who looked at quantile estimators for the BMD, showed that distributional assumptions in standard models impacted BMD estimation.

Although Shao et al. [17] only looks at one error model, there is no requirement that all models use the same error distribution. Instead, we investigate an approach using MA with multiple distributions and different mean models. For example, we include the Exponential-5 model [19], which defines a hexible sigmoidal shaped dose-response function; and we use this with normal and log-normal error models. This method is investigated using the Laplace approach of Wheeler et al. [8], as well as Markov chain Monte Carlo (MCMC) methodologies. In our simulation study, the proposed method better captures the uncertainty in estimating the BMD (e.g., more reliable confidence intervals) than the single distributional form MA approach. In the following sections, we describe the

model averaging approach for BMD estimation, apply this to multiple real dose-response studies and analyze MA performance in a simulation study.

2 | MODEL

2.1 | Benchmark Dose modeling

Assume one has $Y = (y_1, y_2, \ldots, y_n)'$, which are n observations each taken with error on $\mathbb{R}^+$. Each y_i, $1 \leq i \leq n$, corresponds to a dose x_i from an animal toxicology experiment. Here, assume all observations are independently drawn from a common error distribution such that the central tendency changes as a smooth function of dose. This function, $f(x)$, is the dose-response. It determines changes in the response as the dose increases, and its exact role depends on the data distribution. For example, if the y_i are normally distributed, $f(x_i)$ is the mean given x_i, and if one assumes each y_i follows a log-normal distribution, $f(x_i)$ is the median given x_i.

We are interested in determining risk and thus the BMD. Though there are many definitions of risk for continuous responses, we focus on those dependent upon higher-order moments and look at the hybrid and standard deviation definitions of the BMD [16]. Other BMD definitions exist in the literature and can be applied using our proposed methodology, but they only require knowledge of $f(x)$ (e.g., the relative risk definition [19] measures the absolute change in the dose-response from the mean/median.) We do not consider such definitions further.

2.1.1 | Hybrid BMD—In defining risk, the hybrid BMD definition is a direct analog to the dichotomous BMD extra-risk approach [20]. Given a benchmark response(BMR), the hybrid approach estimates the BMD to be the dose that solves

$$\frac{\Pr(Y > y_0 \mid x = \text{BMD}) - \Pr(Y > y_0 \mid x = 0)}{1 - \Pr(Y > y_0 \mid x = 0)} = \text{BMR}\,. \tag{1}$$

Here, $\Pr(Y > y_0 \mid x = x_0)$ is the probability that the response is greater than y_0 at dose $x = 0$ and $\Pr(Y > y_0 \mid x)$ is stochasticly ordered such that $\Pr(Y > y_0 \mid x = x_0) \leq \Pr(Y > y_0 \mid x = x_1)$ when $x_0 < x_1$. Further, for (1), y_0 is a level of adverse response, and the BMR ϵ (0, 1) and is defined prior to the analysis. In this case, the BMR is the increase in probability that a response is adverse, relative to the probability that the response at control would not be adverse (i.e., it is analogous to the extra risk BMR definition for dichotomous response data). In most cases, instead of defining the cut-point y_0 directly, the value is specified as the cut point where the $100 \times \Pr(Y > y_0 \mid x = 0)\%$ of the population exhibit a response at least as extreme as y_0. To define the hybrid approach, either $1 - \Pr(Y > y_0 \mid x = 0)$ or y_0 is defined *a-priori*. In what follows, we specify $\Pr(Y > y_0 \mid x = 0) = 0.025$ as a default.

2.1.2 | Standard Deviation BMD—Finding valid confidence intervals for (1) is difficult. Consequently, the standard deviation definition as a surrogate of the hybrid approach [16]. In the standard deviation (SD) definition, the BMD is the value solving

$$f(x = \text{BMD}) - f(x = 0) = \text{BMR} \cdot \sigma_0, \tag{2}$$

where σ_0 is the standard deviation at dose zero and BMR ≥ 0, which is the number of standard deviations $f(x)$ must change to be classified as adverse. Here, unlike the hybrid approach, only the BMR needs to be specified.

Both (1) and (2) assume an increasing stochastic ordering with dose. When a decreasing ordering occurs similar definitions of the BMD can be made by looking at the lower tails in (1) or interchanging $f(x = \text{BMD}) - f(x = 0)$ with $f(x = 0) - f(x = \text{BMD})$. In what follows, the appropriate definition is obvious in context.

2.2 | Bayesian Model Averaging

Solving (1) and (2) require knowledge of $f(x)$ and higher-order moments (e.g., knowledge of σ_0, etc.); thus when the distribution is unknown, there is additional uncertainty introduced when estimating the BMD. We define a Bayesian MA approach that calculates the BMD across multiple dose responses and underlying error distributions.

Bayesian inference for a single model proceeds by finding the posterior distribution of a vector of parameters θ given Y, a data generating mechanism with log-likelihood $l(Y \mid \theta, f(dose \mid \theta))$, and prior distribution $p(\theta)$. This is done using Bayes' rule,

$$\Pr(\theta \mid\mid Y) = \frac{\exp\{l[Y \mid \theta, f(x \mid \theta)]\}p(\theta)}{p(Y),} \tag{3}$$

where $p(y) = \int \exp\{l[Y \mid \theta, f(x \mid \theta)]\}p(\theta)\mathrm{d}\theta$. Though the posterior distribution is defined for θ, one can use it to derive the posterior distribution for any function of θ, i.e., $g(\theta)$. We investigate inference on the quantity $g(\theta) = \text{BMD}$, defined in (1) and (2) in what follows.

When $f(x \mid \theta)$ and $l(Y \mid \theta,)$ are unknown, Bayesian MA [1, 2] can be used to define a multi-model posterior distribution on the BMD using individual posteriors. Let $\Pr(BMD \mid \mathscr{M}_1, Y)$,…, $\Pr(BMD \mid \mathscr{M}_M, Y)$ be M posterior distributions of the BMD given $\mathscr{M}_1, \ldots, \mathscr{M}_M$ and Y. Additionally, let $\{\pi_1, \ldots, \pi_M\}$ be M probabilities defining the posterior distribution of the M models (the computational formula for these probabilities is discussed in the next section). The posterior distribution of the BMD averaged over all M models is

$$\Pr(\text{BMD} \mid Y) = \sum_{m=1}^{M} \pi_m \Pr(\text{BMD} \mid Y, \mathscr{M}_m). \tag{4}$$

There is no requirement that the distribution used to construct (3) be the same across all models. We consider the case where the model, $\mathscr{M}_m$, is the tuple $\{f_m(x \mid \theta), l_m[Y \mid \theta, f_m(x \mid \theta)]\}$, where $l_m(\cdot)$ represents a different log-likelihood for each m. In what follows, "model" refers to both the dose-response function and the data error model.

Practical considerations are necessary when averaging the BMD. Equation (3) implicitly assumes this is a distribution over finite quantities; however, there are situations where

definitions (1) and (2) will result in infinite estimates of the BMD due to asymptotes in $f(x)$. This plateau results in no posterior mean and an upper bound of the BMD (BMDU) being infinity. For these cases, we determine if the distribution is finite up to the median. If not, we remove these models before averaging and assign these cases a posterior probability of zero. For this reason, we do not consider the BMDU in this simulation.

2.3 | Posterior Model Probability Computation

The posterior distribution model probabilities, π_j, are computed as

$$\pi_j = \frac{p(Y \mid \mathcal{M}_j)}{\sum_{m=1}^{M} p(Y \mid \mathcal{M}_m)}, \tag{5}$$

which requires the normalizing constant $p(Y \mid \mathcal{M}) = \int l_{\mathcal{M}}(Y \mid \theta) p(\theta) \mathrm{d}\theta$. This constant is not analytically available for the dose-responses considered, and calculating this quantity using numerical integration is difficult. Likewise, simulation-based techniques like reversible jump MCMC [21] or bridge sampling [22] requiring special proposal algorithms that are often not easily generalized because they are usually tailored for a specific analysis. Though approximations using the Bayesian Information Criterion [2] have been proposed to estimate this value, it is known to be $\mathcal{O}(1)$ consistent. Instead, we use the Laplace approximation. That is

$$p(Y \mid \mathcal{M}) \approx (2\pi)^{\frac{r}{2}} \left|\hat{\Sigma}\right|^{\frac{1}{2}} \exp[l(Y \mid \mathcal{M}, \hat{\theta})] p(\hat{\theta} \mid \mathcal{M}),$$

where $\hat{\theta}$ is the maximum *a posteriori* estimate for model $\mathcal{M}$, $\hat{\Sigma}$ is the inverse of the negative Hessian of $\Pr(\hat{\theta} \mid Y)$, and r is the number of parameters in θ. This approximation has an $\mathcal{O}\left(n^{-1}\right)$ relative error [23] and it was used in [8] to compute the posterior distribution model probabilities. We refer the reader to the supplement for more information on the models considered in the model average.

Following the U.S. EPA's Benchmark Dose Software 3.2 (BMDS) package, we consider the following set of dose-response models:

$$f_{hill}(x \mid \theta = \{a, b, c, d\}) = a + \frac{bx^d}{c^d + x^d}, \tag{6}$$

$$f_{exp-3}(x \mid \theta = \{a, b, d\}) = a\left[\exp\left(-\{bx\}^d\right)\right], \tag{7}$$

$$f_{exp-5}(x \mid \theta = \{a, b, c, d\}) = a\left[c - (1-c)\exp\left(-\{bx\}^d\right)\right], \tag{8}$$

$$f_{\text{power}}(x \mid \theta = \{a, b, d\}) = a + bx^d. \tag{9}$$

These functions represent a diverse suite of dose-response models for continuous data. Noticeably absent from this list are polynomial models. We do not consider these models as monotone restrictions on polynomials are challenging to enforce, and non-monotone functions lead to problems when evaluating the benchmark dose. For more information on these models, we refer the reader to the supplement.

2.4 | Model Prior Specification

Prior distributions for a specific models' parameters, i.e. θ, should place higher probability over dose-response curves expected to be encountered in practice. As one observes continuous dose-response data on various scales, e.g., one may see liver weights in the hundreds and blood clinical chemistry in the tens, we develop priors to be scaled to the response so that they may generally apply to a large variety of analyses.

For the coefficients determining the mean model, we define our priors under the assumption that the data are scaled so that the mean control response is 1. For a given data set, we rescale the prior distribution based upon the observed response. For example, if a $N(a, \tau^{-1})$ prior is placed over parameter b, where it enters the model as $b \times x$, then a is scaled by μ_0 and the variance τ^{-1} has μ_0^2 as a scaling factor. Some parameters like the shape parameter d of the Hill, Exponential, and Power models are scale-invariant and do not change based upon the response.

For Bayesian MA, placing priors over the dispersion parameters (e.g., σ^2 in the normal model) is challenging because small changes in the prior specification will inhuence the model's posterior weights. Initial numerical experiments suggested that the prior placed on the variance term may disproportionately impact final model weightings, which is true even when individual model estimates and inference are qualitatively identical for two different prior specifications. To minimize this effect, we place a data-based prior over the dispersion parameter and center the dispersion parameter equal to the estimated variance, i.e., the sampled variance or sample geometric variance. We center the background parameter, a, at 1. This implies the prior dose-response analysis is centered at the observed mean at zero dose, assuming variance equal to the sampled variance.

Defining general priors for Bayesian model-averaged dose-response analyses is a complex issue, which likely will require more research. The prior weights impact the posterior model weighting distribution, and we try to specify informative priors due to issues with Lindley's paradox Shafer [24]. In this manuscript, we attempted to make our priors informative over a response range typically seen in gross in-vivo responses (e.g., liver weight). For example, a priori, we would do not expect much more than a fold change relative to the background given exposure. Thus, we use an $N(0, 1)$ or $N(0, 2)$ on most parameters to quantify this idea.

Of all parameters, a prior distribution over the shape parameter, i.e., d in all models, is the most challenging, but our prior distribution comes from the fact that we do not expect large amounts of curvature (i.e., abrupt changes in response between dose groups). In practice, this typically limits d to be between 1 and about 3. However, there are cases where it is most probably less than 1 or greater than 3, so we place a prior that puts about 80% probability

of response a priori on (1, 3); however, when $d < 1$ exposure may be more hazardous, so we constructed a prior placing about 13% of the probability mass below 1, with the remaining 7% placed for values above 3. We admit this is a prior distribution based upon convenience, and other prior distributions may be superior; however, these choices tend to perform well with real-world in-vivo data. We do not expect these choices to be optimal in all situations (e.g., toxicogenomic data), and a sensitivity analysis is always warranted.

The web supplement contains information on the exact prior distributions used and the scaling performed on each parameter/distribution for each model.

3 | SIMULATION

3.1 | Simulation Design

We investigate the approach by analyzing 240, 000 simulated data sets generated with various dose-response shapes, data-generating mechanisms, and experimental designs. Our study's primary purpose is to investigate the performance of MA over distributional assumptions, so we assume three possible data-generating scenarios: normal, lognormal, and inverse-Gaussian, under a variety of dose-response conditions. The first two assumptions are within the model suite, but the inverse-Gaussian is a right-skewed distribution not in the standard modeling suite. Finally, the normal VPM is not included in the simulation. We felt its inclusion was not as important as having a variance assumption that is not in the model suite, and due to the size of the study and computer resources available, we did not consider it further.

To provide a realistic simulation, we base our study on the data provided in Piao et al. [25], which looked at different organ weights for unexposed rats over their life cycle. Using this as the basis of our simulation, we define a total of twenty true dose-response conditions. We report the results based upon experiments with four non-parametric dose-responses. All simulation and results are fully described in the supplement.

For each simulated experiment, we use 10 observations per dose group under different experimental designs. For each data set, the maximum dose tested is 100. Experimental dose spacing conditions were geometric or even, with 4 or 5 non-control doses. For the geometric spacing designs, the dose points were 0, 6.25, 12.5, 25, 50, and 100, where 6.25 was absent for the 4 dose-group design. For the evenly spaced designs, the doses considered were 0, 20, 40, 60, 80 and 100 for the 5 dose-group design and 0, 25, 50, 75, and 100 for the 4 dose-group design. A total of 1000 data sets were generated for each experimental design/dose-response/distributional assumption. For additional information on all of the simulation conditions, we refer the reader to the supplemental material.

Given there may be discrepancies seen between estimation methods, the simulation compares the Laplace approach to MCMC estimation. This study fits MA-1 and MA-2 for both the BMD standard deviation and hybrid BMD definitions. The standard deviation BMD is based upon a benchmark response that is a one SD shift from the control mean. For the hybrid approach, we define adverse responses as those occurring in 2.5% of the control's

tail distribution, and the benchmark response represents a 5% increase in the probability of being adverse.

3.2 | Performance Evaluation

There are a variety of metrics one may use to evaluate the performance of MA. Many regulatory agencies use the BMDL as the POD, and this estimate's frequentist performance should be at or near the nominally specified coverage level of 95%. Our primary metric of evaluation is the lower bound's observed coverage, e.g., Pr(BMDL ≤ BMD), where BMD is the actual benchmark dose.

3.3 | Simulation Results

The non-parametric simulations give insight into the performance of Bayesian MA when the true dose-response or actual distribution is not used. Table 1 shows all results when estimating the SD BMD, and table 2 gives the results for the hybrid model. Both tables provide results for the five non-zero dose simulations. Overall coverage is close to or greater than the nominally specified 95% confidence level with some changes to performance for even or geometric dose spacings. In general, the normal distribution conditions provide either more conservative coverage (i.e., greater than 98%) across all simulation conditions or results that are slightly less than nominal coverage, with some important exceptions. Here, the results for the I-Spline 2 simulation condition can be sub-optimal. In most cases, coverage between MA-1 and MA-2 is similar, when MA-2 is extremely anti-conservative (i.e., < 85%) MA-1 is much closer to nominal (e.g., Dose-response condition 3). In cases where MA-2 is closer to nominal, the difference in coverage is typically between the 1 to 4%.

The SD BMD results are observed to be anti-conservative (i.e., < 95%) at a higher rate than the hybrid BMD results. The primary reason for this is that one SD may be close to the plateau, whereas a 5% increase in the tails represents a more modest response. Thus, for the hybrid methodology, the BMD is more reliably estimated.

In these tables, particularly for the SD results, the I-Spline 2 log-normal and inverse-Gaussian conditions have the worst performance. Table 1 shows this for the SD BMD. Many MA-1 results perform poorly with observed coverage between 78% and 99%, but the normal MA-2 condition's performed markedly worse. For these conditions, observed coverage was between 10 to 20 percent less than MA-1, particularly for the SD benchmark dose condition.

There are multiple reasons for this behavior. Here, even though the true underlying dose-response plateaus, the MA places high weights on the power and exponential models, which are typically near-linear for these data. Figure 1 shows radar charts giving the average model weights over simulations for dose-response condition 2 using MA-1 (dark purple) and MA-2 (yellow) for the inverse-Gaussian and normal simulations for the even five dose group condition. Here, the Hill and Exponential-5 models receive significantly more weight on average for the normal simulation than the inverse-Gaussian simulation. Interestingly, for the MA-1 condition, the Exponential-3 model receives on average 66% of the weight when the data are generated using the inverse-Gaussian. In this case, the first moment is incorrectly

estimated due to the heavy right skew in the observations. As a consequence, this causes problems calculating the higher-order moments and thus the BMD.

The model's weightings change when we add more observations or decrease the variance. However, it requires many observations per dose group to guarantee a plateaued dose-response estimate or a significant reduction in the variability at the control dose. We could only get a plateau with certainty when we decreased the control dose variance by a quarter or raised the number of observations per dose-group to 150, which is unrealistic in practice. This result suggests that in cases where there is heavy right skew in the observed data, considerable caution should be made to interpret the BMD.

In the dose-response 2 inverse-Gaussian condition, this behavior only occurs because the dose-response is essentially "missed." That is, when all of the change occurs in between experimental dose groups, and there is a large amount of right skew. For the even dose spacing conditions, the problems with coverage occur because the BMD is between the zero and the lowest dose 20; however, for the geometric condition, which has experimental doses of 6.25 and 12.5 coverage is conservative (i.e., 100%).

These simulations give insight into the performance of the MCMC and Laplace estimation approaches. No method is uniformly superior, but in general, the Laplace approach tends to be more conservative than its MCMC counterpart. In certain situations, this results in the Laplace attaining nominal coverage where the MCMC fails to achieve the specified rate; however, there are other situations where this results in overly conservative behavior for the Laplace methodology where the MCMC method is closer to or at nominal. Generally, coverage is noticeably worse for MCMC when the distribution is right-skewed.

The additional simulation results in the supplement are similar to the results presented here. From these and the results presented here, we can make several observations: first, adding more distributional forms to the modeling suite improved or, at worst, provided equivocal coverage for a given simulation condition. Sometimes, like that of the non-parametric dose-response condition 2 with inverse-Gaussian data, the improvement was significant. Further, there is often little difference between the Laplace and MCMC approaches in practice. However, the MCMC central estimate is larger than the Laplace MAP estimate and may exhibit anti-conservative coverage. When MCMC performs worse, this is caused by the fact 1 SD may be on the asymptote for the Hill and Exponential-5 models. Here, the MCMC estimate becomes unreliable because there are many infinite BMDs sampled in the posterior distribution. Finally, we note, MA tends to offer conservative coverage, which may approach 100% in certain situations.

4 | DATA ANALYSIS

We investigate continuous model averaging using a 90 day study of Fischer 344 rats exposed to airborne concentrations of dimethylformamide. In this experiment, animals were exposed to dimethylformamide concentrations of 0, 50, 100, 200, 400, and 800 ppm for 13 weeks. Ten female and ten male rats were in each dose group. We investigate the cholesterol blood levels measured at 13 weeks and combine the male and female observations. Data were

obtained from the NTP CEBS database [26]. As described above, we use MA-1 and MA-2 modeling suites for this analysis. The methods and data used for this analysis are available in the supplement. Additionally, adverse response levels for the hybrid and SD approaches are specified as in the simulation.

Table 3 gives the BMD estimates for the 10 continuous models considered and two model averaging approaches. It also provides the posterior weights for each MA approach. This table gives the standard deviation definition of the benchmark dose. This table shows large differences between the posterior weighting of the two model sets. When we include model distributions other than normal, normal models receive only 5% percent of the weight. Further, in the normal model space, the power dose-response model has 29.0% of the posterior probability. When we include the normal variance proportional to the mean and log-normal models, the power model recieves only 2.3% of the weight. In this case, the Log-normal models appear to be superior in describing the data.

In terms of the model average, the BMD estimates are different. Table 3 shows, depending on the estimation method, the BMD central is greater when using only normal models (e.g., 62.4 ppm vs. 116.1 ppm using the Laplace MA methodology). Additionally, the table shows differences between the Laplace and MCMC individual models in terms of BMD results. The BMD posterior distributions have heavy right tails. As a result, the estimated MCMC BMD is higher than the Laplace estimate in all cases as shown in the table.

Though the BMD estimates are different between MA-1 and MA-2, there is very little difference between the predicted dose-response. This behavior is clearly evident in figure (2), which shows this for Laplace estimates. This figure shows the dose-response(solid blue line), and the horizontal line show estimates of the BMDL, BMD, and BMDU from left to right, with the BMD being the dose at the triangle. Using MA-1 it is seen that the BMD distribution is different from the BMD distribution constructed using MA-2. This figure supports the conclusion reached by Shao et al. [17], which suggested the SD approach was more dependent upon the distributional than the relative deviation, which only looks at the dose-response to define risk.

5 | DISCUSSION

We have provided a comprehensive comparison of continuous model averaging when the benchmark dose is defined using higher-order moments of the response distribution. The results are promising and show that adding additional distributional assumptions to the MA produces better coverage results.

Additional research should investigate how increasing the number of models, removing specific models, and adding distributional assumptions impact the method’s performance. Though there may be diminishing returns once the model/distributional space is saturated, the current model suite may be too small in some situations, and the use of model spaces like that described in Aerts et al. [27] may be appropriate. Placing historical prior information on the parameters and prior model probabilities may be a promising area of research. Though we attempted to develop a procedure that was both reasonable and general,

significant gains may be seen by studying the behavior of the priors on individual models and modifying this procedure. Finally, all simulations were conducted assuming the original data are available. There are many situations where only sufficient statistics based upon a normal distribution are available. In those situations, the log-normal distribution's sufficient statistics can only be estimated. As this is frequently done in practice, it is crucial to understand the impact of this procedure on MA.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The author's would like to thank Sooyeong Lim and Dr A. John Bailer for providing the code used in all MA graphics. Additionally we would like to thank Drs Todd Blessinger, Dustin Kapraun, Gary Larson, and Fred Parham for comments on an earlier version of this manuscript. This research was supported [in part] by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. The study was reviewed by the Center for Public Health and Environmental Assessment and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use. Views expressed in this article are the authors' and do not necessarily rehect the US EPA's views or policies.

references

[1]. Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 1997;92(437):179–191.

[2]. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. Statistical science 1999;p. 382–401.

[3]. Claeskens G, Hjort NL, et al. Model selection and model averaging. Cambridge Books 2008;.

[4]. Kang SH, Kodell RL, Chen JJ. Incorporating model uncertainties along with data uncertainties in microbial risk assessment. Regulatory Toxicology and Pharmacology 2000;32(1):68–72. [PubMed: 11029270]

[5]. Wheeler MW, Bailer AJ. Properties of Model-Averaged BMDLs: A Study of Model Averaging in Dichotomous Response Risk Estimation. Risk Analysis 2007;27(3):659–670. 10.1111/j.1539-6924.2007.00920.x. [PubMed: 17640214]

[6]. Piegorsch WW, An L, Wickens AA, Webster West R, Peña EA, Wu W. Information-theoretic model-averaged benchmark dose analysis in environmental risk assessment. Environmetrics 2013;24(3):143–157. [PubMed: 24039461]

[7]. Simmons SJ, Chen C, Li X, Wang Y, Piegorsch WW, Fang Q, et al. Bayesian model averaging for benchmark dose estimation. Environmental and Ecological Statistics 2015;22(1):5–16.

[8]. Wheeler MW, Blessinger T, Shao K, Allen BC, Olszyk L, Davis JA, et al. Quantitative Risk Assessment: Developing a Bayesian Approach to Dichotomous Dose–Response Uncertainty. Risk Analysis 2020;40(9):1706–1722. [PubMed: 32602232]

[9]. Daniels RD, Gilbert SJ, Kuppusamy SP, Kuempel ED, Park RM, Pandalai SP, et al. NIOSH practices in occupational risk assessment Department of Health and Human Services, National Institute for Occupational Health; 2020.

[10]. Committee ES, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen KH, et al. Update: use of the benchmark dose approach in risk assessment. EFSA Journal 2017;15(1):e04658. [PubMed: 32625254]

[11]. Organization WH. Chapter 5 : Dose-Response Assessment and Derivation of Health-Based Guidance Value Organization WH, editor, Geneva: World Health Organizations; 2020.

[12]. West RW, Piegorsch WW, Peña EA, An L, Wu W, Wickens AA, et al. The impact of model uncertainty on benchmark dose estimation. Environmetrics 2012;23(8):706–716. [PubMed: 23794799]

[13]. Shao K, Gift JS. Model uncertainty and Bayesian model averaged benchmark dose estimation for continuous data. Risk Analysis 2014;34(1):101–120. [PubMed: 23758102]

[14]. Varewyck M, Verbeke T. Software for benchmark dose modelling. EFSA Supporting Publications 2017;14(2):1170E.

[15]. Shao K, Shapiro AJ. A web-based system for Bayesian benchmark dose estimation. Environmental health perspectives 2018;126(1):017002. [PubMed: 29329100]

[16]. Crump KS. Calculation of benchmark doses from continuous data. Risk Analysis 1995;15(1):79–89.

[17]. Shao K, Gift JS, Setzer RW. Is the assumption of normality or log-normality for continuous response data critical for benchmark dose estimation? Toxicology and applied pharmacology 2013;272(3):767–779. [PubMed: 23954464]

[18]. Wheeler MW, Shao K, Bailer AJ. Quantile benchmark dose estimation for continuous endpoints. Environmetrics 2015;26(5):363–372. 10.1002/env.2342.

[19]. Slob W Dose-response modeling of continuous endpoints. Toxicological Sciences 2002;66(2):298–312. [PubMed: 11896297]

[20]. Crump KS. A new method for determining allowable daily intakes. Toxicological Sciences 1984;4(5):854–871.

[21]. Fan Y, Sisson SA. Reversible jump MCMC. Handbook of Markov Chain Monte Carlo 2011;p. 67–92.

[22]. Gelman A, Meng XL. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Statistical science 1998;p. 163–185.

[23]. Kass RE, Raftery AE. Bayes factors. Journal of the american statistical association 1995;90(430):773–795.

[24]. Shafer G Lindley's paradox. Journal of the American Statistical Association 1982;77(378):325–334.

[25]. Piao Y, Liu Y, Xie X. Change trends of organ weight background data in Sprague Dawley rats at different ages. Journal of toxicologic pathology 2013;26(1):29–34. [PubMed: 23723565]

[26]. Lea IA, Gong H, Paleja A, Rashid A, Fostel J. CEBS: a comprehensive annotated database of toxicological data. Nucleic acids research 2017;45(D1):D964–D971. [PubMed: 27899660]

[27]. Aerts M, Wheeler MW, Abrahantes JC. An extended and unified modeling framework for benchmark dose estimation for both continuous and binary data. Environmetrics 2020;31(7):e2630. [PubMed: 36052215]

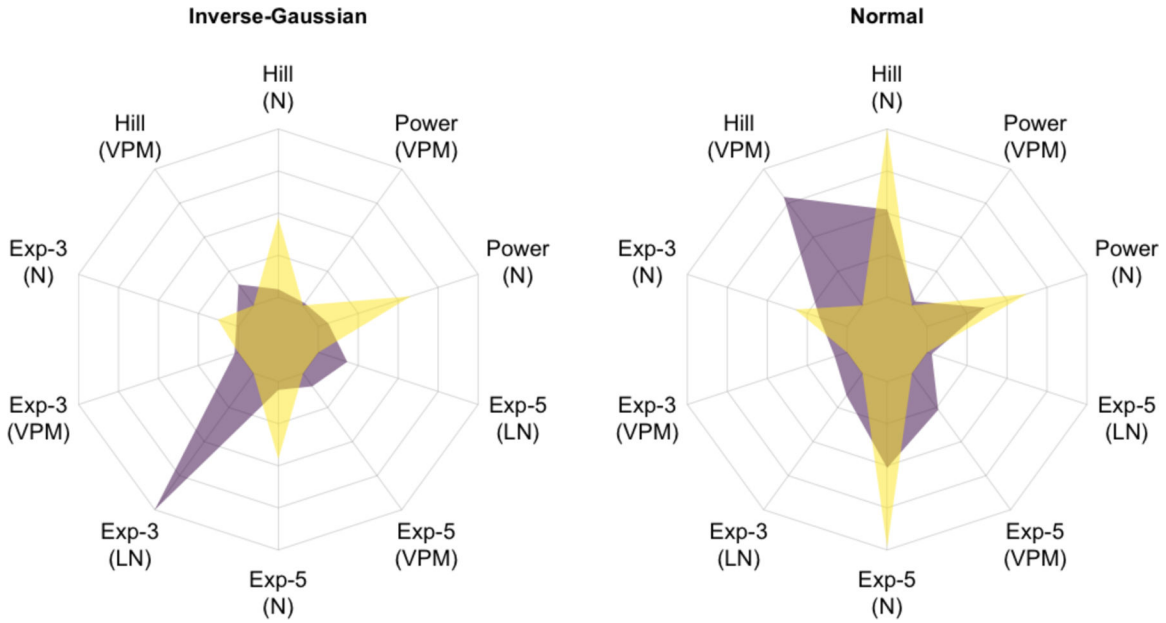

FIGURE 1.

Radar charts showing the average relative model weighting for the MA-1 (dark purple) and MA-2 (yellow) model suites across 1000 simulations for the non-parametric dose-response condition. Here, normal (N), normal variance proportional to the mean (VPM), and log-normal (LN) distributions were fit to the data.

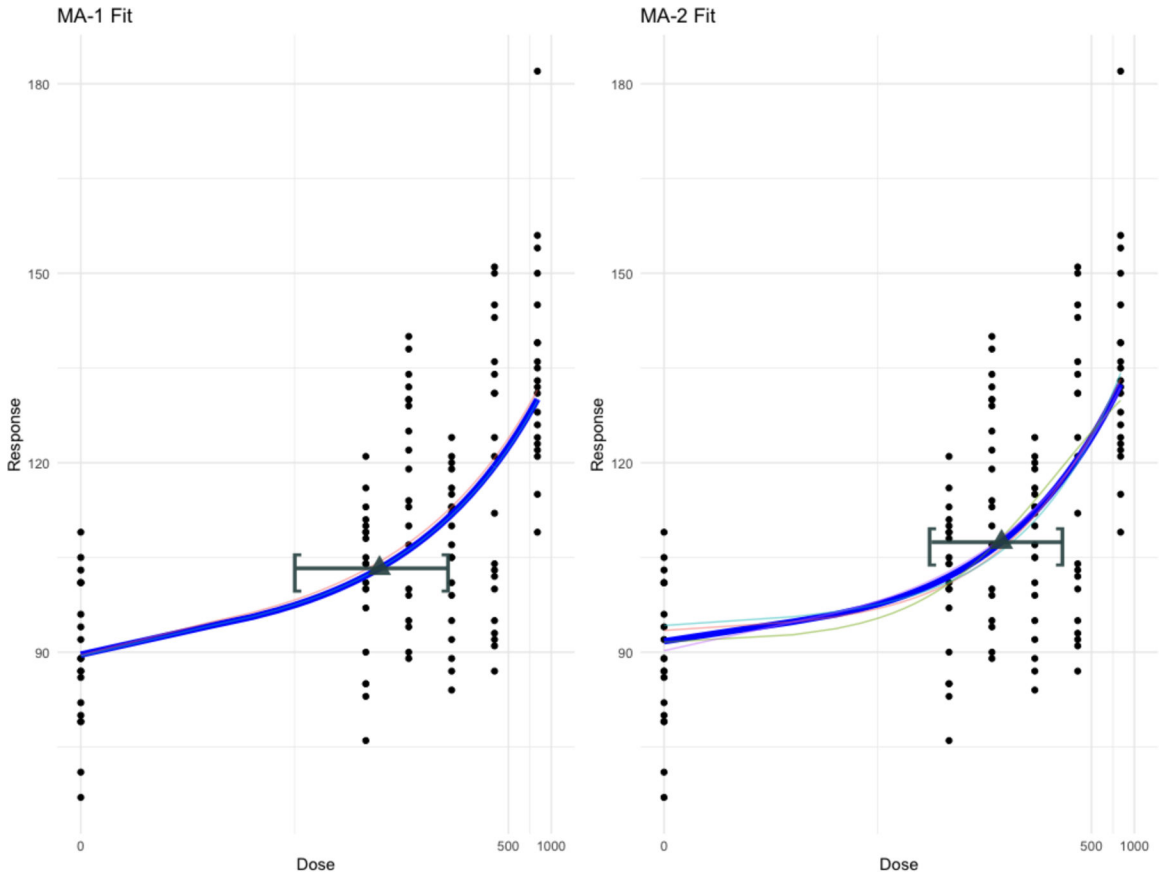

FIGURE 2.

Model average (MA) dose-response estimate of blood cholesterol levels for Fischer 344 rats exposed to dimethylformamide in the air for 90 days using ten models in the MA (left pane) and four model MA (right pane). For both plots, the standard deviation definition of the BMD is used with a benchmark response of 1 standard deviation. The dark blue line represents the MA dose-response and smaller multicolored lines represent dose-respones recieving greater than 5% weight. The triangle represents the response corresponding to the BMD estimate and the horizontal grey line represents the 90% confidence region.

TABLE 1

Coverage percentages using the standard deviation BMD approach across all three error models. Simulations were generated assuming a I-spline dose-response curves using 5 dose groups and a control. Additionally, conditions 1 and 2 represent liver weight changes (increase) and conditions 3 and 4 represent body weight changes (decrease)

		Even Spacing				Geometric Spacing			
		Laplace		MCMC		Laplace		MCMC	
	Simulation Condition	MA-1	MA-2	MA-1	MA-2	MA-1	MA-2	MA-1	MA-2
Normal	I-Spline 1	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
	I-Spline 2	98.5	98.1	95.4	93.2	98.3	97.7	95.3	93.8
	I-Spline 3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	I-Spline 4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Log Normal	I-Spline 1	98.1	98.1	98.7	98.3	99.4	99.4	99.5	99.5
	I-Spline 2	82.0	73.1	75.5	51.0	86.1	80.2	79.9	63.0
	I-Spline 3	94.9	95.5	87.7	91.8	93.0	94.0	84.3	88.2
	I-Spline 4	88.5	91.8	83.0	89.0	90.1	92.2	79.3	86.8
Inverse Gaussian	I-Spline 1	98.3	98.3	98.5	98.2	99.7	99.6	99.8	99.8
	I-Spline 2	87.0	75.2	78.1	56.2	90.7	83.7	81.1	67.6
	I-Spline 3	94.8	95.7	86.5	92.7	94.8	95.1	84.1	89.8
	I-Spline 4	91.5	93.8	80.8	88.8	93.3	93.3	79.4	87.4

TABLE 2

Coverage percentages using the hybrid BMD approach across all three error models. Simulations were generated assuming a I-spline dose-response curves using 5 dose groups and a control. Additionally, conditions 1 and 2 represent liver weight changes (increase) and conditions 3 and 4 represent body weight changes (decrease)

		Even Spacing				Geometric Spacing			
		Laplace		MCMC		Laplace		MCMC	
	Simulation Condition	MA-1	MA-2	MA-1	MA-2	MA-1	MA-2	MA-1	MA-2
Normal	I-Spline 1	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
	I-Spline 2	99.5	99.4	99.1	98.6	99.7	99.4	99.4	98.8
	I-Spline 3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	I-Spline 4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Log Normal	I-Spline 1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	I-Spline 2	96.6	96.0	98.6	94.8	98.9	98.3	98.8	96.1
	I-Spline 3	92.0	95.0	92.6	91.7	92.9	93.6	91.2	89.5
	I-Spline 4	86.6	91.2	86.7	86.5	88.0	93.3	87.2	89.5
Inverse Gaussian	I-Spline 1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	I-Spline 2	90.0	88.3	95.1	81.6	94.6	92.6	95.0	85.7
	I-Spline 3	96.5	97.3	96.2	95.7	98.5	98.0	96.7	95.6
	I-Spline 4	91.8	94.8	91.5	92.2	91.7	94.6	91.7	90.7

TABLE 3

Model average and individual model BMD estimates for the dimethylformamide analysis. BMD values are for the standard deviation definition of the benchmark dose. Model space $\mathscr{M}_1$ represents all models and distributions where as model space $\mathscr{M}_2$ represents only the normal distribution.. Estimates include the central estimate and the 90% credible intervals.

	$P(\mathscr{M} \mid Y, \text{MA-1})$	$P(\mathscr{M} \mid Y, \text{MA-2})$	**Laplace**	**MCMC**
Hill Normal	0.5 %	10.0 %	113.4 (47.3, 265.7)	178.8 (72.5, 381.4)
Exponential-3 - Normal	2.6 %	54.7 %	88.4 (25.8, 221.7)	121.2 (39.9, 278.4)
Exponential-5 - Normal	0.3 %	6.3 %	164.5 (66.7, 348.5)	191.9 (81.5, 387.1)
Power - Normal	1.6 %	29.0 %	189.6 (78.1, 377.1)	236.5 (98.3, 469.7)
Hill - Normal-VPM	1.8 %	-	68.0 (29.9, 157.1)	97.0 (38.9, 246.5)
Exponential-3 - Normal-VPM	5.4 %	-	49.0 (12.5, 141.4)	71.5 (21.2, 196.0)
Exponential-5 - Normal-VPM	1.1 %	-	93.8 (35.9, 227.8)	49.6 (42.6, 49.7)
Power - Normal-VPM	0.7 %	-	110.9 (40.4, 262.4)	144.7 (52.8, 359.4)
Exponential-3 - Log-Normal	84.1 %	-	60.9 (15.8, 169.1)	90.7 (26.5, 236.7)
Exponential-5 - Log-Normal	2.8 %	-	113.7 (43.2, 268.9)	123.2 (49.1, 290.5)
Model Average MA-1			62.4 (16.2, 185.2)	92.0 (27.1, 250.8)
Model Average MA-2			116.1 (37.7, 310.5)	160.1 (50.0, 365.4)