Title: Quantitative risk assessment: Developing a complete Bayesian Approach to Dichotomous Dose-Response Model Averaging

Authors: Matthew W. Wheeler[1], Kan Shao[2], Jeffrey S Gift[3], J. Allen Davis[4], Bruce C Allen[5], Todd Blessinger, Louis Olszyk,

[1] National Institute for Occupational Safety and Health
Risk Evaluation Branch
MS C-15
1090 Tusculum Ave
Cincinnati OH, 45226
mwheeler@cdc.gov

[2]Indiana University
Department of Environmental Health
School of Public Health
1025 E. Seventh Street, Room C040
Bloomington, IN 47405
kshao@indiana.edu

[3] US Environmental Protection Agency
National Center for Environmental Assessment
EPA (B243-01)
RTP, NC 27711
Gift.Jeff@epa.gov

[4] National Center for Environmental Assessment
U.S. Environmental Protection Agency
26 West Martin Luther King Drive
Cincinnati, OH 45268, MC A110
Davis.Allen@epa.gov

[5]Bruce C Allen Consulting
Bruce Allen Consulting,
Chapel Hill, NC, USA
bruce_allen@frontier.com

Abstract:
Model averaging for dichotomous dose-response estimation is superior to estimating the benchmark dose from a single model; however, there remain several challenges with regard to implementing these methods in general analyses before model averaging becomes ready for risk assessment practice. Among these challenges, questions remain on the number and type of the models considered, what to do when model degeneracy occurs within the set of models considered, and the comparison of model averaging to other alternative methods such as nonparametric dose response modeling. For benchmark dose estimation, there is a scant literature of Bayesian techniques that allow the inclusion of prior model information for both the models and the parameters of the constituent models, which would take full use of the Bayesian paradigm. This manuscript introduces an approach that addresses all of these questions while providing a fully Bayesian model averaging framework; further, in contrast to posterior-sampling methods, we approximate the posterior distribution of the parameter of interest (the benchmark dose). The approximation allows for accurate computation while maintaining the speed of maximum likelihood estimation, which is crucial in many applications such as the screening of massive high throughput datasets. In what follows, we develop the method, apply this method to real data, and compare it to other approaches through simulation study under a large variety of true underlying dose-response curves, some of which are avoid parametric model specification as they are generated from monotone stochastic processes. Through the simulation study, the method is shown to be superior to a number of published software tools that represent competing potential and traditional methods for the dose-response analysis of dichotomous data.

Keywords: Benchmark Dose Estimation, Monte Carlo Simulation, Quantitative Risk Estimation

1 Introduction

Model averaging [1-4] is a technique for inference over multiple parametric models; it estimates predictor-response relationship as a convex weighted sum of individual models and is one solution to the problem of model uncertainty in risk assessment. There are many different model averaging approaches dedicated to dose response and benchmark dose estimation [5-11]. Recent research showing that traditional quantitative risk assessments based upon a single "best model" have poor statistical properties [6, 9] and that model averaging is superior to the single model approach [6-8] has led some regulatory Agencies to recommend model averaging over the single model approach [12]. Despite its superiority, there are some remaining technical challenges that need to be addressed before model averaging can be comfortably adopted as standard risk assessment practice. This article proposes a methodology that overcomes these challenges while demonstrating its superiority over suggested practice defined in the US EPA's benchmark dose (BMD) technical guidance document [13] as well as other approaches currently in the literature [10, 14].

The challenges seen in model averaging are twofold: the number and type of parameters in the model and the issue of model degeneracy for nested models. Model averaging is based upon individual parametric models chosen by the modeler, and the performance of the approach is dependent upon the models chosen. Wheeler and Bailer [7] investigate this approach using two model sets showing that the statistical results differ depending on the number and type of the parametric models included in the average. Despite other studies [8] showing the difference is minimal in practical terms, the question still remains "what models should be included in the model average to estimate risk?" As risk assessors could conceivably change the models included in the average resulting in different results simply based upon the modeler's choices, this is a significant concern for risk assessment practice.

The observation that many of the models degenerate into other models as special cases compounds this problem. Some models have bounds on parameters such that when some parameters are

estimated as equal to a bound, the models degenerate into a single model; this leads to concerns of implicit bias in the results by essentially including the same model in the average multiple times. For example, the Weibull and the multistage 2-degree model can degenerate into the quantal linear. Because of this issue, there are problems in the construction of the model weights and inference.

To solve these problems, this article proposes a Bayesian [15] approach different from previously suggested Bayesian approaches [9, 10]. Our proposed approach solves the problems through the inclusion of prior information. Specifically, strict bounds are replaced by "soft bounds" defined by mildly informative prior distributions for the individual parameters of the models included in the analysis. These distributions put low prior probability on regions often defined outside the boundary of the parameters. For example, the US EPA's BMD technical guidance [13] recommends constraining the bounds of the shape parameter of the Weibull model to be greater or equal to 1, because values less than 1 lead to an infinite slope of the dose-response curve at dose zero. Our proposed priors allow the shape parameter to take any positive value, but place less than 1% prior probability to values less than 1. This model will still fit supralinear curves, but such shapes will only get a high weight if models that are more parsimonious do not describe the data well and the data support them. However, in the cases where there are limited data, the shapes of the models are limited to dose-response shapes that frequently seen in practice.

A further advantage of this approach is that it allows for a single model suite across all data sets. Because the parameters are restricted through their prior distributions, the model average will work for models that have more parameters than there are data points. As the models can be included regardless of the degeneracy issue and the number of data points, the approach allows for a large suite of model across many different study designs.

The manuscript is organized as follows: Section 2, describes the model averaging method and the prior choices as well as justifications for their use. Section 3 gives an analysis on several real

dichotomous datasets. Section 4 outlines a comprehensive simulation study of the method and gives the results comparing it to current practice.

2 Model

Consider an animal toxicology experiment with m doses $d_1,\ldots d_m$ and $n_1,\ldots,n_m$ animals per dose group. For this experiment, let $y_1,\ldots y_m$ be the number of positive responses observed in each dose group. It is frequently assumed that $y_i \sim \text{binomial}(\pi(d_i), n_i)$, where $\pi(d_i)$ is the probability of adverse response at dose d_i. To estimate $\pi(d_i)$ given $y_1,\ldots y_m$, $\pi(d_i)$ is often assumed to be a parametric function of dose. For example, the current US EPA Benchmark Dose Software (BMDS) [16] can estimate $\pi(d)$ using one of nine dose-response functions. Picking a single model (*e.g.*, any single model in the US EPA BMDS suite of models) may result in misrepresentation of the true underlying dose-response relationship and significant model uncertainty. Bayesian Model averaging [1, 3] develops a probabilistic framework to incorporates inference from the models considered.

We use these same models in the model averaging procedure, but place priors over the parameters of each model. The priors place a high probability over shapes commonly seen in practice and lower probability on other dose-response curves that may be unreasonable. For example, when using the method of maximum likelihood to estimate the parameter shape parameter, α, of the Weibull model (described below), EPA constrains this value to be less than 18 [16]. Values near 18 result in a hockey stick-shaped dose-response curve that implies the probability of an adverse event goes from background to 100% in an extremely small dose range. Thus, high values of this parameter are unlikely, and the proposed prior puts exponentially decreasing weight on values of α near 18 and higher. This results in the Bayesian estimate of this parameter to be smaller than its equivalent estimate made using maximum likelihood. This is especially true in cases where there are limited data on the dose-response curve; however, when there are sufficient data one sees minimal differences between the Bayesian estimate and the method of maximum likelihood. Further, by placing priors that ensure positivity of the

parameters, instead of the strict bounds seen with maximum likelihood, the prior prevents models from degenerating into other models. Further, if one model is close to degenerating into another model, the more parsimonious model will be preferred over the more complicated model in the average.

2.1 Models Considered

The nine models used in the modeling suite for the present methodology are listed in Table 1, along with the priors for the individual model parameters. All priors are specified for dose-response curves for doses on [0,1] interval where 1 is the maximum tested dose. For dose-responses not in this range, doses are rescaled in relation to 1 being the maximum dose; the new data set is used in all fitting procedures with all BMD and BMDL values rescaled in relation to the maximum dose.

Model Averaging

For a given dataset D and a model M_k, we fit each model M_k individually and compute the posterior distribution of the BMD, i.e., $g_k(\text{BMD} \mid M_k, D)$. The posterior density of the model averaged BMD is

$$g_{ma}(BMD) = \sum_{k=1}^{9} \pi_k(M_k|D) g_k(BMD|M_k, D), \quad (1)$$

where π_k is the posterior probability of model M_k given the data. The BMD and BMDL are then computed from this posterior distribution. More specifically, the point estimate of the BMD is taken as the median of this density and the BMDL is taken as the $100*\gamma^{th}$ percentile for appropriately low confidence level γ. Model weights π_k are approximated from using the Laplace approximation [3].

The posterior distribution described in (1) is approximated using a Laplacian approximation [17, 18]. The approximation is similar to the Model Averaged Profile Likelihood (MAPL) approach of Fletcher and Turek [19]. However, while MAPL relies on deterministic calculations, our approach incorporates prior information in that it uses the marginal profile density [18, 20] of the BMD. In other

words, both the likelihood and prior are used. The method consists of two steps. First, one develops model-specific posterior distributions for the parameter of interest (eliminating other, "nuisance" parameters); and second, a weighted combination of the model-specific distributions is taken. The model-specific distribution is defined by treating profile distribution bounds as quantiles of a marginal posterior distribution for the parameter of interest, and the relation to the present approach and the MAPL approach is justified asymptotically [21, 18]. The full explanation of this approximation is described in the supplement.

Our approach can be related to the MAPL framework by substituting the posterior distribution for the likelihood in each of the steps. Instead of information criteria based weight (e.g., see the model averaging literature [1, 3, 4]) we use the Laplace approximation for model weights. This method approximates an integrated likelihood using the posterior maximum a posteriori (MAP) estimate and Hessian of the log-posterior. Further the profile posterior density is used instead of profile likelihood.

2. 2 Weight Calculation

In previous approaches to benchmark dose inference using model averaging (*e.g.*, Bailer *et al.* [5], weights were calculated using either the BIC or AIC, where the AIC is used primarily in frequentist model averaging [2, 4]. The proposed approach generates weights using the Laplace approximation to the marginal density of the data. That is for model M_k, $1 \leq k \leq 9$, with parameter vector θ_k of length s, one approximates the marginal density as

$$I_k = (2\pi)^{s/2}|\Sigma_k|^{1/2}\ell\left(D\middle|\hat{\theta}_k, M_k\right)g(\hat{\theta}_k|M_k), \qquad (2)$$

where Σ_k is the negative inverse Hessian matrix, $\hat{\theta}_k$ is the MAP estimate, $\ell\left(D\middle|\hat{\theta}_k, M_k\right)$ is the likelihood of the model given the data *D*, and $g(\hat{\theta}_k|M_k)$ is the prior density for θ_k.

For each model M_k, one calculates the MAP and calculates I_k using equation (2). The posterior probability of the model is

$$\pi_k(M_k|D) = \frac{g(M_k)I_k}{\sum_{i=1}^{9} g(M_k)I_k},$$

where $g(M_k)$ is the prior probability of model M_k (e.g., 1/9 if each of 9 models is treated as equally plausible *a priori*).

2.3 Computation of the Model-Averaged BMDL and BMD Point Estimate

Our model-averaged BMD point estimate is the weighted average of MAP estimates from individual models, weighted by posterior weights $\pi_k(M_k|D)$. This is equivalent to the median of the posterior distribution defined in equation (1). For the BMDL, equation (1) is integrated. A 100α% BMDU or 100(1 - α) % BMDL is the value BMD_α such that:

$$\alpha = \int_{-\infty}^{BMD_\alpha} \Pr(\text{BMD}\,|\,D)d\text{BMD} \ . \qquad (3)$$

For full details on the approximation defined in equation (3), see the supplement included with the online version of this manuscript.

3 Data Example

To illustrate the approach applied to experimental data, we choose two datasets that present challenges to the risk assessor. Using current methodologies solutions to these challenges may not be fully satisfactory, but are reasonable using the proposed method. These challenges include limited data, parameter bounds, and supra-linearity. All analyses were conducted using a BMR of 10% extra risk, with $\alpha = 0.05$, and all models are given equal prior probability of being the true model. The supplement gives an excel spreadsheets that implement this method and can reproduce the results below.

3.1 N-Nitrosomorpholine

Ketkar *et al*. [22] exposed Syrian golden hamsters to N-Nitrosomorpholine in their drinking water. Four dose groups consisting of 50, 28, 30, and 30 rats were given of 0, 1.36, 6.82, and 13.60 mg/kg/day in their drinking water. Respiratory hyperplasia is the endpoint of interest, where 0, 14, 16, and 22 of the animals had the adverse health effect. With the current BMDS system, this dataset presents many challenges for the analyst. In this dataset, none of the constrained model, i.e., models that are not supralinear, adequately describe the data; and, unconstrained models have unacceptably small BMDLs with all of the supra linear models containing zero for the lower bound estimates.

For the analysis, the final weight assigned to the log logistic model was 82.9% with the quantal linear and log-probit each having just under 6% of the final weight; the benchmark dose is 0.16 mg/kg/day (0.01). Figure 1 gives the estimated model average dose-response (black) and constituent models (shades of grey) that form the average. Here darker shaded curves have more weight. From a visual inspection of the fit the model average dose-response is well within the error bars of the given data sets, indicating the model adequately describes the data; this is significant as none of the constrained models adequately fit the data. Further, this analysis provides a reasonable estimate of the BMDL at 0.01 ppm, as opposed to 0 for all of the unconstrained dose-response datasets.

3.2 Methyl Isocyanate

Dodd and Fowler [23] conducted a sub chronic vapor inhalation study of Methyl Isocyanate with Fischer 344 rats. In this study, four dose groups of 10 rats each were exposed to 0, 0.15, 0.60 and 3.1 ppm of Methyl Isocyanate in the air. This study observed non-neoplastic lesions only in the highest dose group; all 10 animals had these lesions, i.e., the observed lesion count was 0, 0,0,10 respectively. Unlike the first example, this study misses the dose response curve entirely. Additionally, the low sample sizes increase the probability that no animals experience lesions at low doses. The resultant likelihood is completely separable, and analyses using the Weibull and similar models force the shape

parameter α to be as large as possible. For the log-logistic, gamma, and Weibull models the BMDS system estimate the shape parameter at its upper bound; this results in dose-response curves that are essentially on/off switches. Though the BMDL is similar across these models the BMD is determined by the maximum bound programmed into the BMDS system and will tend towards 3.1 ppm as this bound is increased. In many cases, the bound is arbitrary and often set based upon computer precision.

This is not the case for the proposed method. Figure 2 gives the model average plot as the black line and the corresponding individual fits as grey lines. The BMD for all models is well-defined and the estimated curves do not resemble on/off switches. The prior provides information that shrinks the curve fit back to the mean of the prior, and though different priors would produce different results, the motivation behind the prior choices becomes apparent; the priors favor dose-response curves that do not increase arbitrarily rapidly. As a result, the BMD is lower because the method shrinks the results back to dose-response curves with higher prior probability when data do not adequately define the shape of the curve. Despite the difference, the BMDL is in line with the results from a BMDS analysis, which bound the BMDL to be somewhere between 0.33 and 0.57 ppm; in the case of the MA the BMDL is 0.41 ppm.

4 Simulation

To investigate this approach, we created simulations from 34 different dose response curves assuming an experimental condition designed to mimic chronic bioassays. Simulation results are provided for the described MA approach using the priors defined above (denoted as "Prior 1a" in supplemental material Appendix 3) and for the MA approach using several alternative sets of model parameter priors (see supplemental material Appendix 3, Table SA3-1). We use four dose groups with 50 observations per group with geometric spacing between doses (0, 0.25, 0.5, and 1.0) and analyze 2000 datasets investigating coverage, bias (% of true BMD) and BMD/BMDL ratio (see supplemental, material, Appendix 3, Tables SA3-2, SA3-3 and SA3-4). Further, to investigate the sensitivity of the model to the prior model weight choice, two model weighting schemes were assessed. The first, denoted as the MA

248 "even" alternative in Tables SA3-1, SA3-2 and SA3-3, assumes all models are equally likely *a priori*;
249 the second condition, denoted as the MA "QL = 0.5" alternative in Tables SA3-1, SA3-2 and SA3-3,"
250 places 50% of the weight on the quantal linear model with the remaining eight models given equal
251 6.25% weighting. The "QL = 0.5" alternative is referred to as the MAQ approach below. The
252 development of the second condition follows from the literature, which suggests near linear dose-
253 response curves are the most difficult to account for in model averaging approaches [7, 9].

254 For comparison purposes, simulation results are also provided for the approach recommended in
255 the US EPA BMD technical guidance for the selection of a "best model" [13]. For that simulation
256 analysis, all models described above were fit, except the Hill model, which was not fit due to
257 convergence issues. The models that fit the data were considered further (i.e., having p-value > 0.1). The
258 BMD and BMDL from the model with the lowest AIC was chosen unless the range of BMDLs from
259 adequately fitting models was more than 3-fold, in which case the BMD and BMDL from the model
260 with the lowest BMDL was chosen. Additionally, simulation results are provided for a competing
261 Bayesian model averaging method from Shao and Shapiro [10] (denoted as the MAKS approach), using
262 the same priors as describe above. This methodology fits all models except the hill model using the same
263 priors as defined above and uses a model averaging approach as defined in that manuscript. Finally, for
264 an additional comparison, simulation results are provided for the non-parametric (NP) method of Guha
265 *et al.* [14].

266 **4.1 True Dose Response Curves**

267 To simulate a range of plausible dose-response relationships, we define 34 dose response curves
268 from a variety of shapes. The shapes varied from simple parametric forms, to weighted averages of
269 parametric models, to smooth monotone curves generated from stochastic processes. The shapes mimic
270 plausible curves that may be seen in a dose-response analysis as well as certain cases that might be non-
271 standard, which can be used as a benchmark to diagnose possible problems with any particular method.

4.1.1 Single Parametric Models in the Model Suite

To mimic a flexible parametric model that may be included in the model suite, we use the multistage 3-parameter model to form various true shapes. As this model has limitations, we add the log-logistic model to simulate the single, but flexible, parametric model. The 3-parameter multistage is

$$\pi_{ms3}(d) = \gamma + (1-\gamma)[1 - \exp(-\beta_1 d - \beta_2 d^2 - \beta_3 d^3)],$$

and the Log-logistic is defined above. Figure 3 (M1-M14) and Figure 5 (M24-M26) show the range of shapes considered. Though all dose-response curves are monotone, we include three nonstandard curves in this set of models. Models M3, M8 and M12 all increase at low and high dose ranges, but plateau somewhere in the mid-dose range. These curves, though not expected to represent dose-response curves routinely encountered in practice, give an indication of data sets where the methods may have difficulty fitting the data. The exact form of the dose-response curves is in Table SA2-1 of the supplement.

4.1.2 Convex Sum of Multiple Parametric Models

In addition, we investigate cases where the true dose-response is a convex combination of the underlying dose response curve. Though these dose-responses are representable by the proposed methodology, one should not consider these as directly in the model space using the proposed methodology. As the sample size goes to infinity, model averaging converges on the single model that minimizes the Kulback-Leibler divergence within the data generating mechanism [24], implying that for large *n* there is a single model to which the average model converges with probability 1. For the first set of dose-responses considered (M15-M17 and M23), we look at a convex sum of

$$\pi_{s1}(d) = \frac{1}{1+\exp(3-4d)},$$

and

$$\pi_{s2}(d) = 0.02 + 0.98 \times [1 - \exp(-1.5d)],$$

which is the logistic and quantal linear model respectively. Table SA2-2 of the supplement gives the different convex sums considered in these conditions and the corresponding BMD, while Figure 4 gives the range of the dose response models considered (M15-M17 and M23).

In addition to the two model convex combination, we consider another set of models (M18-M22) composed of a four-parameter convex sum. In this case, the four true dose response conditions are

$$\pi_{s3}(d) = \Phi(-1.6 + 2.5d),$$

$$\pi_{s4}(d) = 0.02 + 0.98[1 - \exp(-1.6d)],$$

$$\pi_{s5}(d) = 0.02 + \frac{0.98}{1+\exp[-1.3-2\times\log(d)]},$$

and

$$\pi_{s6}(d) = 0.02 + 0.98[1 - \exp(-1.5d^{2.2})],$$

which are a probit, quantal linear, log-logistic and Weibull models respectively. Table SA2-3 of the supplement gives the different convex sums considered in these models and the corresponding BMD, while Figure 4 gives the range of curves estimated (M18-M22). These conditions all form near linear dose-response conditions found to be problematic model averaging cases by Wheeler and Bailer [7].

4.1.3 Models Out of the Model Suite.

We investigate models not representable as any function in the model suite; these are denoted in Figure 5 as simulation models M27-M34. Models M27-M32 are generated from a smooth monotone stochastic process over a basis set (e.g. see Higdon et al. [25]); in these simulations, random coefficients for each basis were generated in a manner that guaranteed monotonicity. To guarantee the plausibility of the dose-response, each curve was visually inspected and found to be a reasonable dose-response shape. In addition to the non-parametric curves M27-M32, two additional cases, M33 and M34, were considered. M33 uses an exponentially modified Gaussian distribution, which has a history in analytical chemistry [26]. For M34, a multistage 3 degree model was created to define a case of high dose

downturn. For these simulation, the generation of each curve is available in an R program [27] in the supplement. Figure 5 gives the range of curvature defined using these functions (M27-M34) and shows that a large range of curvature was considered when constructing the simulations.

4.2 Simulation Results

For the simulation, we investigated the observed coverage Pr(BMDL < BMD_{TRUE}), the relative BIAS percentage $100 \times E\left[\frac{\widehat{BMD}}{BMD_{TRUE}}\right]$ %, and the expected ratio between the lower bound estimate and the estimated BMD as a measure of spread $E\left[\frac{\widehat{BMDL}}{\widehat{BMD}}\right]$. We note that the more commonly suggested ratio of the lower and upper bound (BMDL/BMDU) was not used as it was not available in all of the methods investigated (e.g. the BMDS modeling results). Additionally, the statistics were computed for BMRs of 1 and 10%. As there are 34 true dose-response curves, two BMRs for each curve, and five methods tested, not all of the results are presented here, but are available in the supplement.

Tables 2-4 give the observed coverage for all of the methods for the BMR = 10% ER. The simulation results for the 1% ER BMR, given in the supplement, are in line with the 10% results, and are not discussed further. Overall, the proposed model averaging approach and the non-parametric approach of Guha *et al* [14] are similar; these two approaches frequently achieve near nominal coverage, i.e., ≥ 90% across the simulations. In contrast, the current BMDS approach failed to achieve nearly nominal coverage in most simulations, and the model averaging approach of Shao and Shapiro [10] usually performed worse than proposed approach and the NP approach. Unsurprisingly, all methods performed poorly for simulation conditions M3, M8 and M12, which are conditions where the response increases, plateaus, and then increases again at higher doses. In many of these cases, the coverage is 0%, which is a result of the poor fit of the parametric methods. Even the non-parametric monotone approach performed poorly in these conditions because the NP method linearly interpolates between observed points. In cases of concave dose-responses between dose groups, a linear interpolation will systematically

underestimate the true dose-response curve and the corresponding benchmark dose. For the NP approach, this pattern is also seen in simulations M1, M23, and M24; all conditions are concave between zero and the first tested dose of 0.25.

The simulations also examined the effect of placing *a priori* weight on the quantal linear model of 0.5, and these results show that by using this weighting scheme coverage may improve for many dose-response that are very similar to dose-responses observed in practice. For example, for simulation conditions M15-M25, coverage is improved to nominal or near-nominal rates with little impact on the coverage for curves that are clearly sublinear. This indicates such weighting schemes may help in modeling the BMD for most dose-response data seen in practice.

Though the MA, MAQ and NP approaches obtain similar coverage for many models, there are differences in the methodologies' performance. Simulation M1 obtains observed coverage of 0% using the NP approach as compared to 97.9% using MA. When this occurs in the simulations (M1, M23, M24, and M31) it can be traced back to the linearization performed in the specific MA approach. In cases where the NP approach clearly outperforms MA, that is M4, M13 and M33, the true dose-response is nearly linear (i.e., directly proportional to the dose). The constituent parametric models in the MA do not support the shape, whereas the linear interpolation of the NP approach appropriately models the true dose response curve as it assumes a linearity between observed doses.

The simulation results also investigated the bias of the methods. The MAQ results exhibited less bias than the MA approach and typically had less bias than the NP approach, which had more conservative point estimates for sub linear dose-response relationships. For example, conditions M7 and M32 were very sub-linear dose-response functions; the NP approach had point estimates that were 35.4% and 34.9% of the true BMD whereas the MAQ approach had point estimates that were 78.6% and 79.6% of the true BMD, which were identical to the MA values. The BMDS and approach of Kan and Shapiro performed better than they did in the case of coverage, but the results were not noticeably better

than either the MA or MAQ approaches. Though the MAQ weightings make the BMDL more conservative with respect to the MA, these results show that the MAQ change the point estimate very little, and possibly making the estimate less biased. A point that is seen with regard to the ratio statistic, fully described in the supplement. Nearly all of the MAQ BMDLs are closer to the BMD, which argues that the MAQ weighting scheme also increases the stability of the estimates.

5 Discussion

The proposed dichotomous model averaging method solves various problems that have not been properly addressed within the literature. As seen in the data examples, it allows the use of unconstrained models without problems in the estimation of the lower bound, something that occurs frequently in model averaging using unconstrained models. By using a Bayesian approach, it allows the fitting of models that have more parameters than there are data points, and this allows the use of a consistent model suite across dose-response data sets, which increases transparency as it prevents modelers from advantageously picking a set of models that may support a conclusion deemed appropriate *a priori*.

The MA method, with our proposed priors, performs favorably against many of the current state-of-the-art methods, and it does so in a comprehensive simulation study using several alternative sets of parameter priors and representing 34 plausible dose-response curves that are both in and out of the MA modeling suite. Though one may contend that these results are based upon the use of informative priors, which bias the result in the favor of the proposed method, it would be difficult to construct a set of priors tailored to all of the simulations conditions simultaneously; thus, the results are more indicative of general properties of the method than the specific priors used. Additionally, the method of Shao and Shapiro [10] use the same priors as proposed in the MA method, which gives a reasonable comparison as to the effect of the priors with an alternative approach. Further, we contend that the priors are not very informative for the range of dose responses normally considered reasonable by most toxicologists. The priors only affect the results when the dose response exhibits a very steep response (i.e., when the dose-

response relationship is not captured by the experiment and some prior information should be used to make sure sensible estimates are generated), or when there is very little data exist to inform the dose response (i.e., when sample sizes are small, as in the second data example). This is not to say that more appropriate priors cannot be developed in the future to produce better results in certain situations, but the given results show the current priors offer a significant improvement over traditional analyses (BMDS), and little bias when compared to other methods.

Finally, we mention that the method was developed with regard to practical considerations, including the need for consistency across dose-response analyses and the need for fast analytic methods to model very large datasets (e.g., many high throughput toxicity datasets). The proposed MA approach should promote consistency by removing many of the decisions a risk assessor needed to make in performing a traditional dose-response analysis, including manually running multiple individual models and choosing a "best model." It is also much faster than previously published MA approaches. For this approach, individual model results are fit in milliseconds, with all model averaging results (BMD and BMDL estimates) computed within a half of a second or less on a modern desktop. This is in contrast to previous model average approaches such as Wheeler and Bailer [7] that require half a minute to complete, or full MCMC based approaches such as Shao and Shapiro [10] that may require even longer run times depending convergence.

References

1. Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. Journal of the American Statistical Association, 1997; 92:179-91.
2. Buckland ST, Burnham KP, Augustin NH. Model selection: An integral part of inference. Biometrics, 1997; 53 (2):603-18.
3. Hoeting JA, Madigan D, Raftery AE et al. Bayesian model averaging: A tutorial. Statistical Science, 1999; 14 (4):382-417.
4. Claeskens G, Hjort NL. Model selection and model averaging. Cambridge, England: Cambridge University Press; 2008.
5. Bailer AJ, Noble RB, Wheeler MW. Model uncertainty and risk estimation for experimental studies of quantal responses. Risk Analysis, 2005; 25 (2):291-9.
6. Wheeler M, Bailer AJ. Comparing model averaging with other model selection strategies for benchmark dose estimation. Environmental and Ecological Statistics, 2009; 16 (1):37-51.
7. Wheeler MW, Bailer AJ. Properties of model-averaged bmdls: A study of model averaging in dichotomous response risk estimation. Risk Analysis, 2007; 27 (3):659-70.
8. Shao K, Gift JS. Model uncertainty and bayesian model averaged benchmark dose estimation for continuous data. Risk Analysis, 2013; 34 (1):101-20.
9. Simmons SJ, Chen C, Li X et al. Bayesian model averaging for benchmark dose estimation. Environmental and Ecological Statistics, 2015; 22 (1):5-16.
10. Shao K, Shapiro AJ. A web-based system for bayesian benchmark dose estimation. Environmental Health Perspectives, 2018; 126 (1):017002.
11. Faes C, Aerts M, Geys H et al. Model averaging using fractional polynomials to estimate a safe level of exposure. Risk Analysis, 2007; 27 (1):111-23.
12. Hardy A, Benford D, Halldorsson T et al. Update: Use of the benchmark dose approach in riskassessment. 2017; 15 (1).
13. U.S. EPA. Benchmark dose technical guidance. Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum Report No.: EPA/100/R-12/001.
14. Guha N, Roy A, Kopylev L et al. Nonparametric bayesian methods for benchmark dose estimation. Risk Analysis, 2013; 33 (9):1608-19.
15. Gelman A, Carlin JB, Stern HS et al. Bayesian data analysis. 3rd ed. Boca Raton, FL: CRC Press; 2014.
16. U.S. EPA. Benchmark dose software (bmds). Version 2.7.0.4. In., Series Benchmark dose software (bmds). Version 2.7.0.4. Washington, DC: U.S. Environmental Protection Agency, National Center for Environmental Assessment; 2017.
17. Leonard T, Hsu JSJ, Tsui KW. Bayesian marginal inference. Journal of the American Statistical Association, 1989; 84 (408):1051-8.
18. Hsu JSJ. Generalized laplacian approximations in bayesian inference. Canadian Journal of Statistics, 1995; 23 (4):399-410.
19. Fletcher D, Turek D. Model-averaged profile likelihood intervals. Journal of Agricultural, Biological, and Environmental Statistics, 2012; 17 (1):38-51.
20. Hu B, Ji Y, Tsui KW. Bayesian estimation of inverse dose response. Biometrics, 2008; 64 (4):1223-30.
21. Severini T. On the relationship between bayesian and non-bayesian interval estimates. Journal of the Royal Statistical Society: Series B (Methodological), 1991; 53:611-8.
22. Ketkar MB, Holste J, Preussmann R et al. Carcinogenic effect to nitrosomorpholine administered in the drinking water to syrian golden hamsters. Cancer Letters, 1983; 17:333-8.
23. Dodd DE, Fowler EH. Methyl isocyanate subchronic vapor inhalation studies with fischer 344 rats. Toxicological Sciences, 1986; 7:502-22.
24. Yao Y, Vehtari A, Simpson D et al. Using stacking to average bayesian predictive distributions. Bayesian Analysis, 2017; 2017.

25. Higdon D, editor Space and space-time modeling using process convolutions; 2002. 37-56 p. (CW Anderson; V Barnett; PC Chatwin et al. editors. Quantitative methods for current environmental issues).
26. Pauls RE, Rogers LB. Band broadening studies using parameters for an exponentially modified gaussian. Analytical Chemistry, 1977; 49 (4):625-8.
27. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Model	Constraints	Priors	Notes
Quantal linear $\pi_1(d) = \gamma + (1-\gamma)(1-\exp[-\beta d])$	$\beta > 0$ $0 \le \gamma \le 1$	log(β) ~ Normal(0,1) Ψlogit(γ) ~ Normal(0,2)	$\gamma = \frac{1}{1+\exp(-\Psi)}$
Multistage $\pi_2(d) = \gamma + (1-\gamma)(1-\exp[-\beta_1 d - \beta_2 d^2])$	$\beta_1 > 0$ $\beta_2 > 0$ $0 \le \gamma \le 1$	log(β_1) ~ Normal(0,0.25) log(β_2) ~ Normal(0,1) Ψlogit(γ) ~ Normal(0,2)	Note the prior over the β_1 parameter expresses the belief that the linear term should be positive if the quadratic term is positive in the two hit model of carcinogenesis.
Weibull $\pi_3(d) = \gamma + (1-\gamma)(1-\exp[-\beta d^{\alpha}])$	$\beta > 0$ $\alpha > 0$ $0 \le \gamma \le 1$	log(β_1) ~ Normal(0,1) log(α) ~ Normal(log(2),0.18) logit(γ)Ψ ~ Normal(0,2).	Here the prior over α is set so that there is only a 0.01 prior probability the power parameter will be < 1. This allows for models that are supra-linear, but requires a large amount of data for the α parameter to go much below 1.
Gamma $\pi_4(d) = \gamma + \frac{1-\gamma}{\Gamma(\alpha)} \int_0^{\beta d} t^{\alpha-1} \exp(-t)\, dt$	$\beta > 0$ $\alpha > 0$ $0 \le \gamma \le 1$	log(β) ~ Normal(0, 1) log(α) ~ Normal(log(2),0.18) logit(γ)Ψ ~ Normal(0,2)	Here the prior over α is designed such that there is only a 0.01 prior probability the power parameter will be less than 1. This allows for models that are supra linear; however, it requires a large amount of data for the parameter to go much below 1.
Dichotomous Hill $\pi_5(d) = \gamma + \frac{\nu(1-\gamma)}{1+\exp[-a-b\log(d)]}$	$0 \le \gamma \le 1$ $0 \le \nu \le 1$ $-\infty < a < \infty$ $b > 0$	a ~ Normal(0, .25) b ~ Normal(log(10),0.0625) logit(γ)Ψ ~ Normal(0,2) ν ~ Normal(4,2)	$\gamma = \frac{1}{1+\exp(-\Psi)}$
Logistic $\pi_6(d) = \frac{1}{1+\exp[-\beta_0-\beta_1 d]}$	$-\infty < \beta_0 < \infty$ $\beta_1 > 0$	β_0 ~ Normal(0, 1) log(β_1) ~ Normal(0,2)	
Log-Logistic $\pi_7(d) = \gamma + \frac{1-\gamma}{1+\exp[-\beta_0-\beta_1\log(d)]}$	$-\infty < \beta_0 < \infty$ $\beta_1 > 0$	β_0 ~ Normal(0, 1) log(β_1) ~ Normal(log(2),0.25) logit(γ)Ψ ~ Normal(0,2).	$\gamma = \frac{1}{1+\exp(-\Psi)}$
Probit $\pi_8(d) = \Phi(\beta_0 + \beta_1 d)$	$-\infty < \beta_0 < \infty$ $\beta_1 > 0$	β_0 ~ Normal(0,1) log(β_1) ~ Normal(0,1)	
Log-Probit $\pi_9(d) = \gamma + (1-\gamma)\Phi[\beta_0 + \beta_1 \log(d)]$	$-\infty < \beta_0 < \infty$ $\beta_1 > 0$	β_0 ~ Normal(0, 1) log(β_1) ~ Normal(log(2),0.25) logit(γ)Ψ ~ Normal(0,2)	$\gamma = \frac{1}{1+\exp(-\Psi)}$

462 **Table 1**: Individual models used in the model averaging method and their respective parameter priors. Note that $\text{logit}(\gamma) = \log\left(\frac{\gamma}{1-\gamma}\right)$.

Test Condition	MA	MAQ	BMDS	NP	MAKS
1	97.9%	97.0%	99.1%	0.0%	93.6%
2	99.7%	99.7%	90.8%	100.0%	100.0%
3	0.0%	0.0%	0.0%	0.0%	0.0%
4	52.1%	55.6%	36.2%	92.7%	67.9%
5	98.8%	98.8%	77.8%	99.2%	100.0%
6	100.0%	100.0%	91.4%	100.0%	100.0%
7	100.0%	100.0%	100.0%	100.0%	100.0%
8	100.0%	100.0%	97.0%	0.0%	0.0%
9	87.4%	91.0%	88.0%	94.1%	68.9%
10	100.0%	100.0%	92.6%	99.7%	99.3%
11	100.0%	100.0%	100.0%	100.0%	100.0%
12	29.7%	6.2%	48.8%	33.5%	0.0%

Table 2: Observed coverage probabilities for the test conditions M1-M12 with BMR = 10% for multiple methods: MA is the proposed method with equal weighting, MAQ is the proposed method with 50% prior weight assigned to the quantal linear, BMDS is the current algorithm used by the US EPA, and fitting procedure recommended by the US EPA [13], NP is the non-parametric Bayesian procedure of Guha et al. [14], and MAKS is the fully Bayesian model averaging approach of Shao and Shapiro [10].

Test Condition	MA	MAQ	BMDS	NP	MAKS
M13	67.7%	83.5%	80.9%	98.8%	82.9%
M14	94.9%	95.0%	91.6%	100.0%	98.8%
M15	94.9%	95.0%	57.8%	97.2%	78.8%
M16	88.2%	95.1%	56.3%	94.1%	82.3%
M17	91.6%	96.9%	81.2%	89.2%	61.9%
M18	91.6%	93.1%	65.6%	98.4%	88.5%
M19	95.5%	98.3%	73.6%	97.1%	89.6%
M20	97.2%	97.9%	76.2%	99.0%	94.4%
M21	91.5%	92.7%	78.7%	99.2%	88.4%
M22	92.7%	94.5%	61.6%	98.3%	88.6%
M23	89.6%	90.5%	87.5%	83.7%	50.9%
M24	97.1%	99.9%	67.7%	65.8%	99.9%
M25	100.0%	100.0%	99.7%	100.0%	100.0%
M26	95.8%	98.8%	53.1%	95.4%	96.5%

Table 3: Observed coverage probabilities for test conditions M13-M26 with BMR = 10% for multiple methods. Here MA is the proposed method with equal weighting, MAQ is the proposed method with 50% prior weight assigned to the quantal linear, BMDS is the current algorithm, and fitting procedure recommended by the US EPA [13], NP is the non-parametric Bayesian procedure of Guha et al. [14], and MAKS is the fully Bayesian model averaging approach of Shao and Shapiro [10].

Test Condition	MA	MAQ	BMDS	NP	MAKS
M27	99.5%	99.8%	95.1%	99.6%	99.5%
M28	100.0%	100.0%	100.0%	100.0%	100.0%
M29	100.0%	100.0%	100.0%	100.0%	100.0%
M30	92.7%	97.3%	93.2%	99.8%	97.6%
M31	95.7%	99.0%	67.3%	56.1%	92.6%
M32	95.9%	100.0%	77.7%	100.0%	100.0%
M33	0.9%	36.4%	59.6%	96.9%	46.4%
M34	80.7%	99.8%	99.7%	98.9%	83.7%

Table 4: Observed coverage probabilities for the test conditions M27-M34 with BMR = 10% for multiple tested method. Here MA is the proposed method with equal weighting, MAQ is the proposed method with 50% prior weight assigned to the quantal linear, BMDS is the current algorithm, and fitting procedure recommended by the US EPA [13], NP is the non-parametric Bayesian procedure of Guha et al. [14], and MAKS is the fully Bayesian model averaging approach of Shao and Shapiro [10].

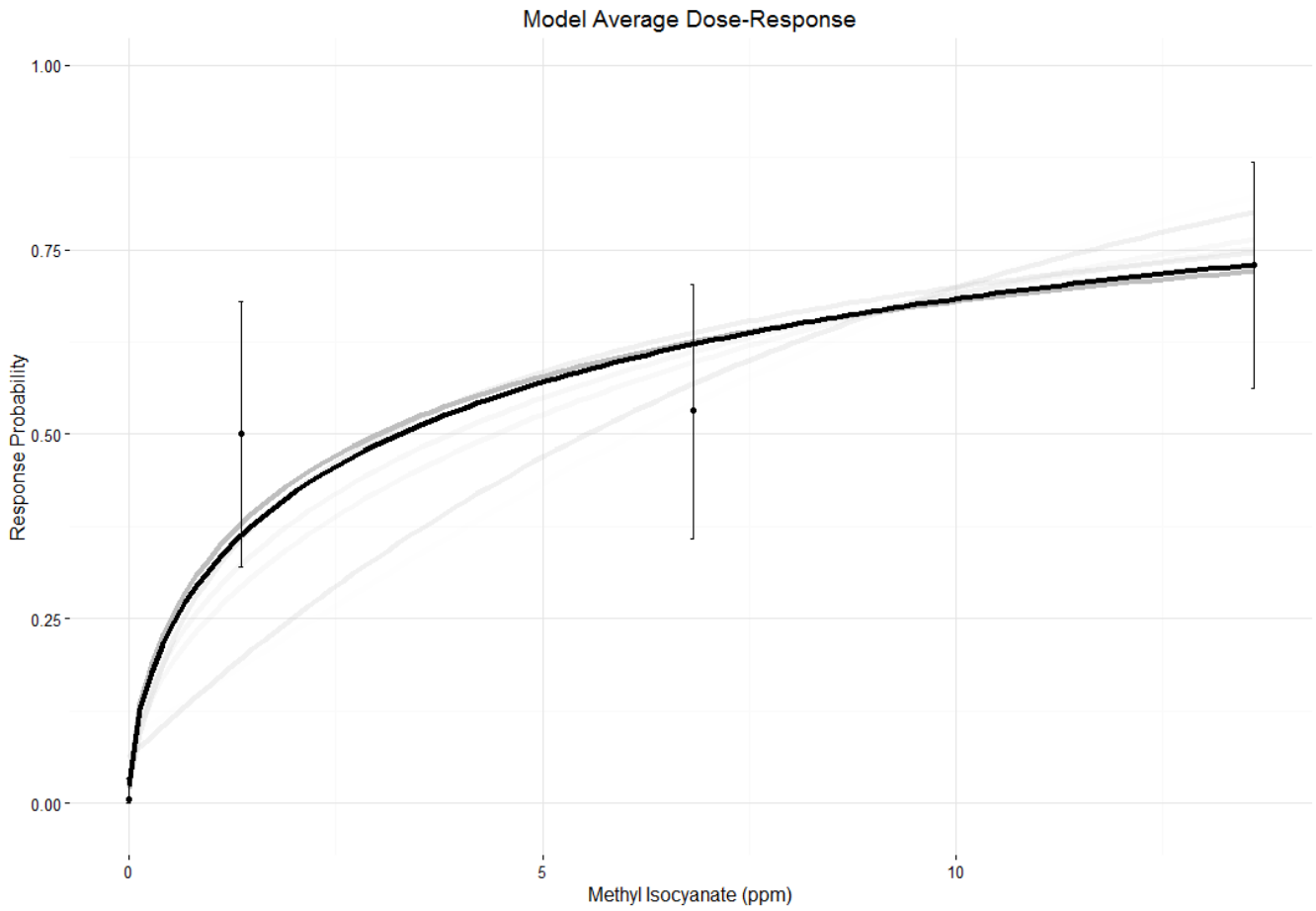

Figure 1: Model average estimate of the dose response function for N-Nitrosomorpholine data. The model average is in black, and the other curves (shades of grey) represent the constituent curves in the model average. The darkness of the grey curves is proportional to the model weight, where darker grey curves receive higher weight.

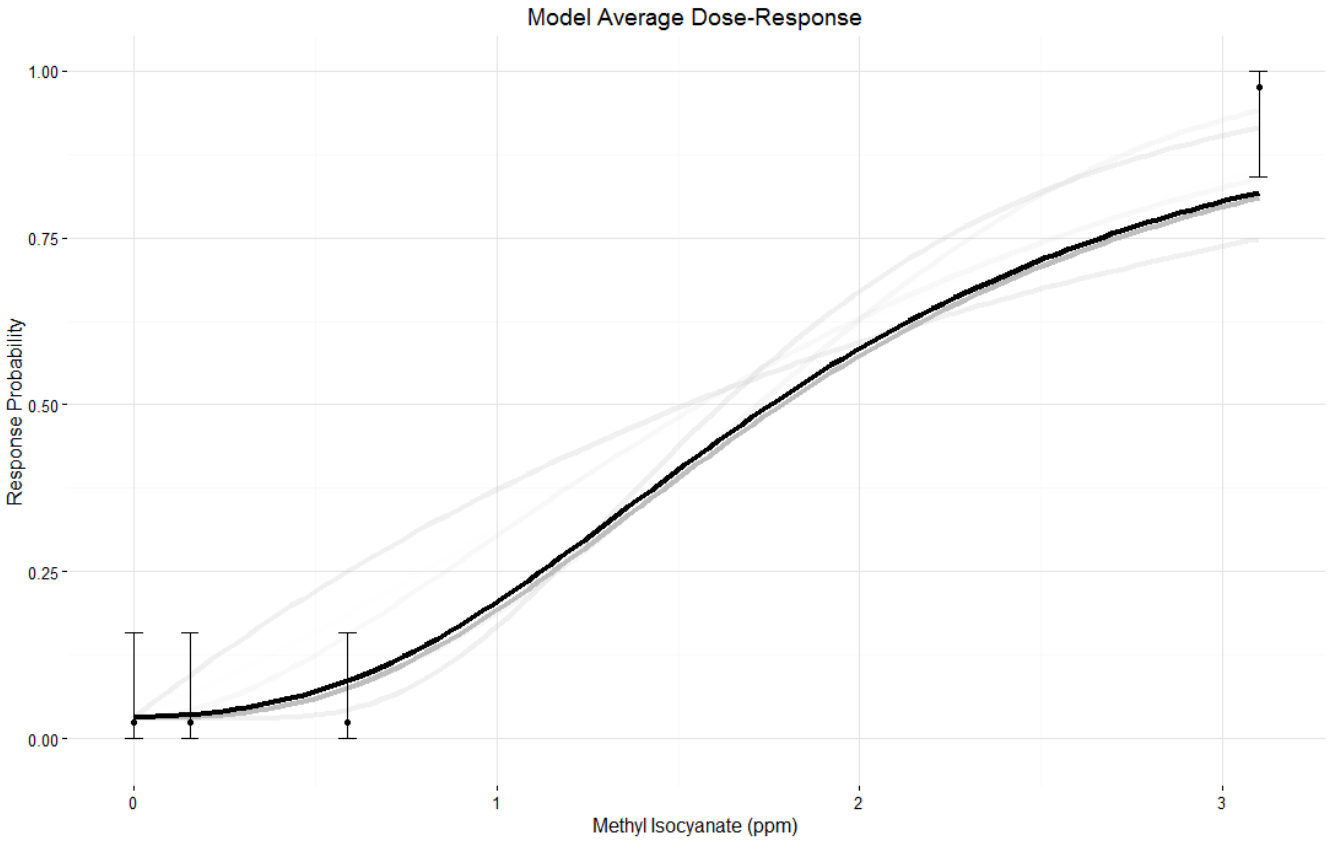

Figure 2: Model average estimate of the dose response function for Methyl Isocyanate dose response data. The model average is in black, and the other curves (shades of grey) represent the constituent curves in the model average. The darkness of the grey curves is proportional to the model weight, where darker grey curves receive higher weight.

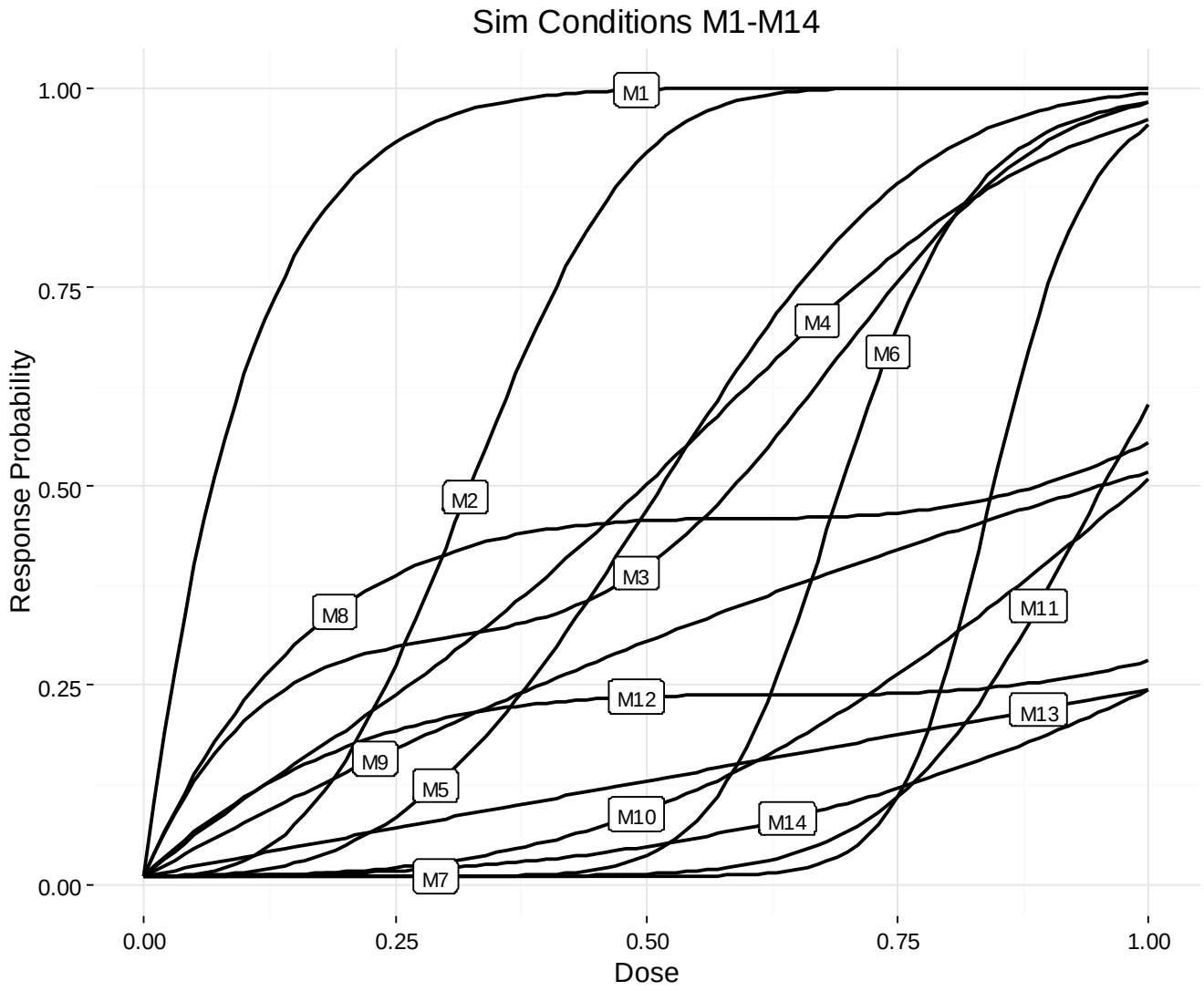

Figure 3: Realized dose-response curves for simulation conditions M1-M14. Simulation conditions were generated using a single parametric model.

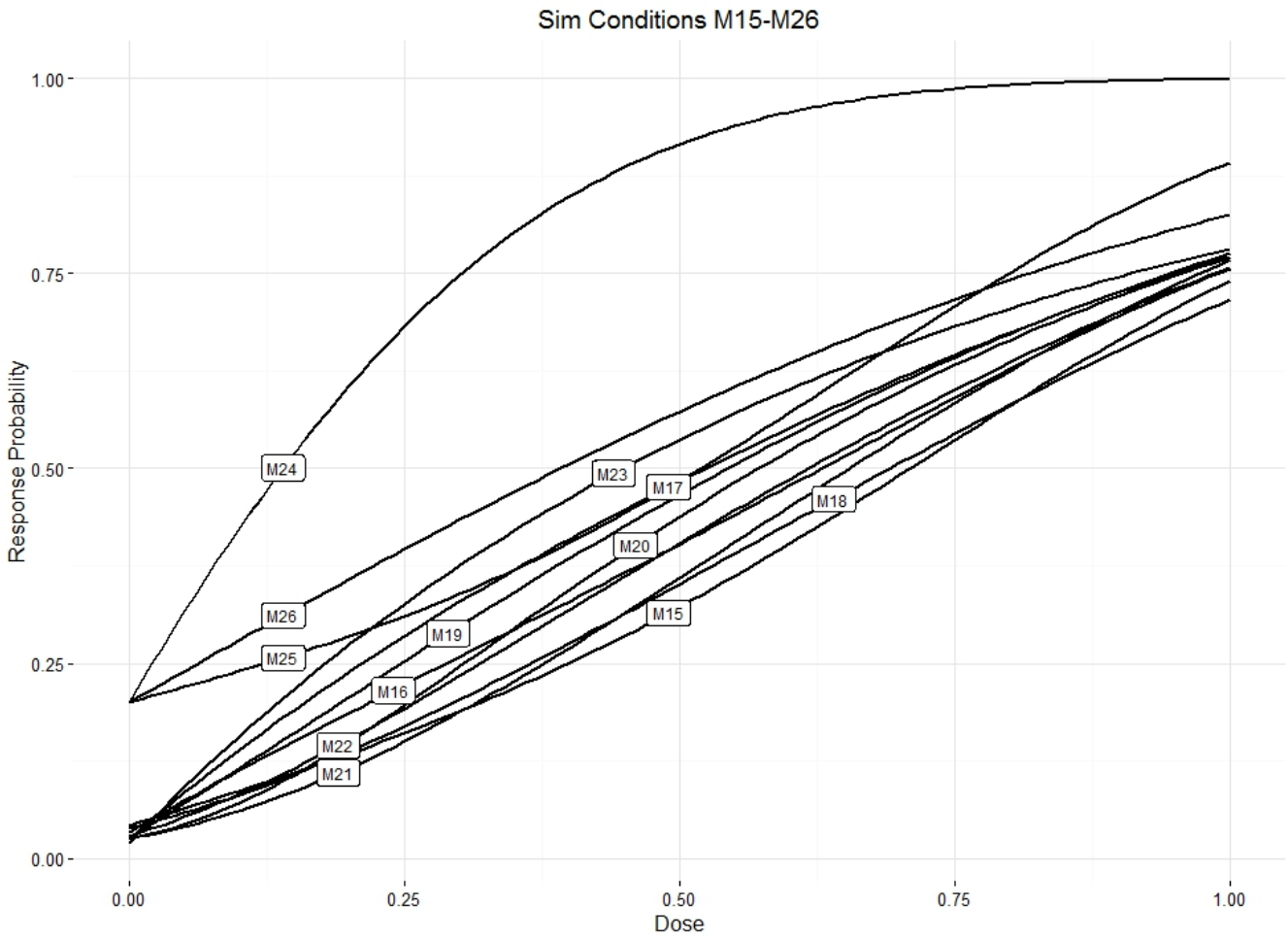

Figure 4: Realized dose-response curves for simulation conditions M15-M26. Simulation conditions M15-M23 were generated using a convex sum of multiple parametric models. Simulation conditions M24 – M26 were generated from a 3 degree multistage parameter to test the performance when a mode is not in the model suite and has a higher background rate.

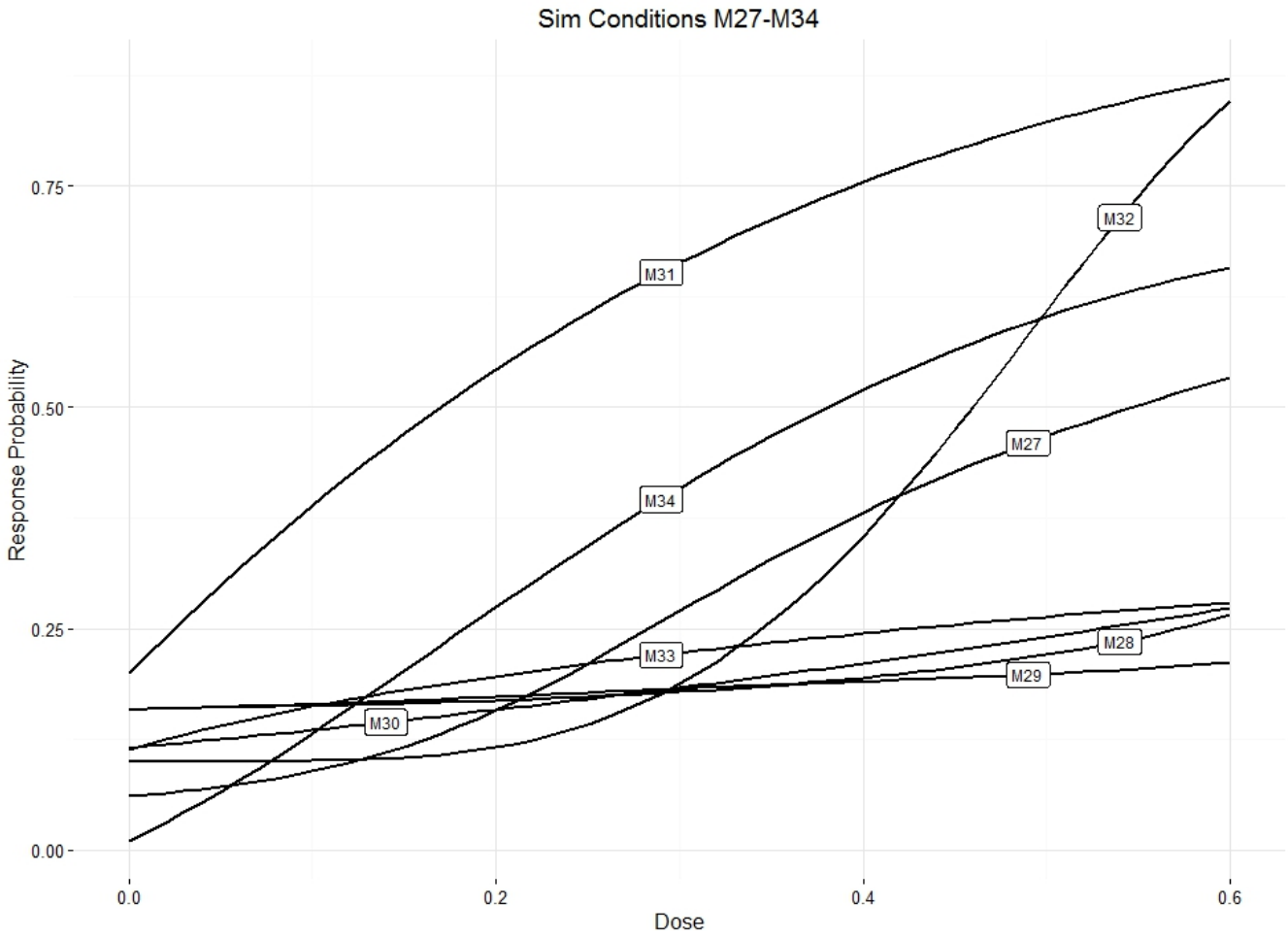

Figure 5: Realized dose-response curves for simulation conditions M27-M34. Simulation conditions were generated using monotone stochastic processes (M27-M32) or were generated from parametric models outside of proposed model averaging approach (M33 and M34).