Comment Report
External Review (Letter) of

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

Contract No. EP-C-17-023
Task Order 68HE0C18F0790

June 25, 2018

Prepared for:
U.S. Environmental Protection Agency
Office of Research and Development
National Center for Environmental Assessment
Research Triangle Park, NC 27711
Attn: Jeffrey S. Gift, Ph.D.

U.S. Environmental Protection Agency
Office of Research and Development
National Center for Environmental Assessment
26 West Martin Luther King Drive
Cincinnati, OH 45268
Attn: Allen Davis

Prepared by:
Versar, Inc.
6850 Versar Center
Springfield, VA 22151

Peer Reviewers:
Vanja Dukic, Ph.D.
Walter W. Piegorsch, Ph.D.
Wout Slob, Ph.D.

TABLE OF CONTENTS

I. INTRODUCTION

The purpose of this peer review is to provide services to the U.S. Environmental Protection Agency's (EPA) National Center for Environmental Assessment (NCEA), Office of Research and Development (ORD), in organizing and conducting by mail an expert external panel review of the draft NCEA report entitled *Quantitative risk assessment: Developing a complete Bayesian app*roach to Dichotomous Dose-Response Model Averaging. EPA's Benchmark Dose Software (BMDS) is the primary dose-response tool for use in human health risk assessments within the EPA and globally. Currently, there are well over 5,000 registered users across 90 countries. The peer review of this model averaging software is essential to completing BMDS version 3.0 in FY2018, which is a key EPA annual performance measure that has been communicated to the Office of Management and Budget (OMB).

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. The EPA manuscript under review introduces an approach that addresses all of these questions while providing a fully Bayesian model averaging framework; further, in contrast to posterior-sampling methods, EPA approximates 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. EPA develops the method, applies the method to real data, and compares 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.

The purpose of the requested letter review is for EPA to receive written comments from individual experts. Versar has selected three reviewers with expertise in the areas of the use of dose-response models for chemical risk assessment, and familiar with EPA benchmark dose (BMD) methods. In particular, reviewers have experience or expert knowledge of model averaging methods and Bayesian statistics as they apply to dose-response analysis. Reviewers understand maximum *a posteriori* estimation and "profile likelihood" methods.

Peer Reviewers:

Vanja Dukic, Ph.D.
University of Colorado-Boulder

Walter W. Piegorsch, Ph.D.
University of Arizona

Wout Slob, Ph.D.
National Institute for Public Health and Environment (RIVM)
The Netherlands

II. CHARGE TO REVIEWERS

Charge Questions:

1. Are the documentation materials describing the proposed method, including the description of the principles and advantages of BMA in general, accurate and clear?

2. Are the methods described adequate for the derivation of BMDLs that are reasonable for use as points of departure for use in EPA risk assessments? In particular, with respect to:

(a) Use of approximations such as profile posterior density (PPD, a Bayesian analogue of profile likelihood) for model-specific posterior BMD distributions, and Laplace approximation for integrated likelihood (marginal density of the data).

(b) The possibility of having more parameters than dose groups in a given application, for a single model. The EPA states in the report provided that incorporation of prior information for model parameters allows application to data with fewer dose groups than parameters.[1] In current dose-response modeling practice, EPA does not use a model when the number of parameters exceeds the number of dose groups.

i. Do the reviewers agree with EPA that the proposed BMA methodology is reasonable to use, when the number of parameters exceeds the number of dose groups for individual models?
ii. Is additional research suggested, e.g., for some cases that may be problematic?
iii. Related to (i), the total number of parameters combined for all models is expected to frequently exceed the number of dose groups (typically 3-5). The EPA has concluded that if the approach proposed for individual models is reasonable, so are the BMA results. Will the large number of fitted parameters result in "overfitting"?[2]

[1] The EPA is aware of general Bayesian literature which suggests that informative priors can address identifiability issues.

[2] This term is sometimes used to indicate that a model is very flexible resulting in a relatively complicated fit with features that may not be repeatable.

(c) The approach for the derivation of BMD point estimators.[3] The currently-proposed estimator is the weighted average of the maximum a-posteriori point estimates from individual models, weighting by model posterior weights. Do the reviewers suggest one or more alternative or additional BMD point estimators for the model averaging context?

(d) The proposed default model parameter priors defined in the draft manuscript (described as “Prior 1” in Appendix 3 simulation results). The “Diffuse Hill” condition described in Appendix 3 of the draft manuscript support material uses the same prior as the proposed approach, except the hill model’s prior is more diffuse. Can you comment on if this prior is preferable to the proposed “Prior 1” set of priors?

(e) The use of equal model weights (described as “Even” model weighting in Appendix 3 simulation results). As discussed in Section 4 of the draft manuscript, in an effort to account for problematic conditions in the literature, we increased the quantal linear default weighting (“MAQ approach”; described as “QL = 0.5” model weighting in Appendix 3 simulation results), and this resulted in better results with little evidence of deleterious performance for other models. Can you comment on if this prior weighting should be used in place of equal weights?

3. Do you agree with the particular models selected for BMA or do you recommend a different set of models?

4. Was adequate testing of the methods performed? In particular,

(a) What additional steps, if any, are recommended to build confidence in the profile posterior density and Laplace approximations? Are any special situations evident, where these approximations may work particularly poorly?

(b) Is the Monte Carlo testing approach taken for the sensitivity analyses an appropriate tool for evaluating the method?

(c) Are the “templates” adequate or is additional testing recommended in order to evaluate other aspects of study design such as numbers per dose group, or dose spacing.

5. What output other than the BMD, BMDL, posterior weights and plots, would be necessary to provide enough information to users for the purposes of quantitative risk assessment? What fit statistics would be necessary to assess model/method performance/fit (e.g., global goodness-of-fit p-values, scaled residuals, posterior p-value)?

6. Is the USEPA proposal to implement this methodology with default priors reasonable in practice, given the likely user of BMDS who have limited familiarity with Bayesian methods? If yes,

(a) How does this methodology compare with current methods, with regard to likelihood that non-statisticians will use it appropriately and accurately?

(b) What situations may be envisioned where default priors would be over-ridden, and what measures would help to make sure this is done appropriately (if it is needed)?

[3] Possible uses of BMD point estimates include comparative and meta-analyses, common use of the ratio BMD [point estimate]/ BMDL ratio as in indication of the quality of the model results.

7. Additional Recommendations: Are there any additional aspects of software development and testing, or model documentation, or reporting of model results that give you special cause for concern? If so, please describe your concerns and recommendations.

III. INDIVIDUAL REVIEWER COMMENTS

Vanja Dukic, Ph.D.
University of Colorado-Boulder

Peer Review Comments on
Quantitative risk assessment:
Developing a complete Bayesian approach to Dichotomous Dose-Response Model Averaging

Reviewer Vanja Dukic

I. GENERAL IMPRESSIONS

This proposed work is a specific implementation of model averaging in risk assessment with dichotomous parametric dose-response models that rely on the Laplace approximation for evaluation of complex integrals, including the marginal likelihoods. The substantive models going into model averaging are very well thought out, as are the suite of priors considered, and I have no concerns with respect to the scientific conclusions.

With respect to statistical and computational issues, section 2 could and should be fleshed out a lot more. There are many omitted details in the statistical domain. PPD is not motivated well, and in fact not even defined. It is not clear at all why it’s needed, and how, if at all, it competes with simple marginalization in these particular parametric models considered.

I know the authors have done a great job assessing computational stability, and while it is impossible to have a numerical routine that will work well in all situations, perhaps they could implement a few more checks. One that comes to mind is well-spaced multiple initial starting points for numerical optimization to see if there are multiple modes in the likelihood/posterior; I am not sure if this has been implemented already, but any evidence of multiple modes should be taken seriously as it affects both MAPs and the Laplace approximation.

II. RESPONSE TO CHARGE QUESTIONS

1. Are the documentation materials describing the proposed method, including the description of the principles and advantages of BMA in general, accurate and clear?

Yes, I have found them to be generally clear in terms of justification and motivation for BMA. The motivation and assumptions behind BMA are clearly stated, and several sensitivity analyses to those assumptions are presented.

I have provided a list of comments at the end of this document that hopefully should provide some guidance for additional clarification (not just to me, but to an average reader/user).

2. Are the methods described adequate for the derivation of BMDLs that are reasonable for use as points of departure for use in EPA risk assessments? In particular, with respect to:

2.a. Use of approximations such as profile posterior density (PPD, a Bayesian analogue of profile likelihood) for model-specific posterior BMD distributions, and Laplace approximation for integrated likelihood (marginal density of the data).

PPD should be more motivated and the procedure fleshed out; this a relatively sparsely used technique in the Bayesian world, where we prefer proper marginalization of parameters to profiling. The authors should provide an additional paragraph or two explaining why PPD is justified, and indeed preferred to (better or faster than) marginalization.

Integrated likelihood is a term generally reserved for nuisance parameter integration. We usually use marginal likelihood within the context of Bayesian model averaging.

Laplace approximation is a well-known technique, so no need for further clarification there.

In general, I found section 2 to be quite a bit on the slim side. The authors should flash it out a bit more, with formulas and details especially for PPD for the posterior. Also, they could provide a lot more details in the supplement.

2.b. The possibility of having more parameters than dose groups in a given application, for a single model. The EPA states in the report provided that incorporation of prior information for model parameters allows application to data with fewer dose groups than parameters[1]. In current dose-response modeling practice, EPA does not use a model when the number of parameters exceeds the number of dose groups.

2.b.i. Do the reviewers agree with EPA that the proposed BMA methodology is reasonable to use, when the number of parameters exceeds the number of dose groups for individual models?

Yes, similar situations are not un-common in the Bayesian modeling world.

2.b.ii. Is additional research suggested, e.g., for some cases that may be problematic?

Perhaps one could think of the extra dose groups as missing data, and this can be formally treated in the Bayesian paradigm. I am not sure if that is currently implemented or not in this approach – the authors should specifically discuss this in the discussion in more detail.

2.b.iii. Related to (i), the total number of parameters combined for all models is expected to frequently exceed the number of dose groups (typically 3-5). The EPA has concluded that if the approach proposed for individual models is reasonable, so are the BMA results. Will the large number of fitted parameters result in "overfitting"?[2]

That is always a possibility. Thus, if the model is going to be used for any policy and decision making, sensitivity analysis (with priors centered at different means, such as in 3 scenarios: pessimistic, optimistic and realistic) is very important.

[1] The EPA is aware of general Bayesian literature which suggests that informative priors can address identifiability issues.

[2] This term is sometimes used to indicate that a model is very flexible resulting in a relatively complicated fit with features that may not be repeatable.

2.c. The approach for the derivation of BMD point estimators.[3] The currently-proposed estimator is the weighted average of the maximum a-posteriori point estimates from individual models, weighting by model posterior weights. Do the reviewers suggest one or more alternative or additional BMD point estimators for the model averaging context?

As it stands, the combination of the MAPs is useful for many reasons (including the way the model is approximated and fitted). The authors could consider using individual posterior means instead, as they are known to have better decision theoretic properties than MAPs. This should also be added to the discussion.

2.d. The proposed default model parameter priors defined in the draft manuscript (described as "Prior 1" in Appendix 3 simulation results). The "Diffuse Hill" condition described in Appendix 3 of the draft manuscript support material uses the same prior as the proposed approach, except the hill model's prior is more diffuse. Can you comment on if this prior is preferable to the proposed "Prior 1" set of priors?

I think all priors should be implemented and the range of results examined. There are situations when one set of priors will be preferred to the other, depending on what the data are, so it is hard to say something general.

2.e. The use of equal model weights (described as "Even" model weighting in Appendix 3 simulation results). As discussed in Section 4 of the draft manuscript, in an effort to account for problematic conditions in the literature, we increased the quantal linear default weighting ("MAQ approach"; described as "QL = 0.5" model weighting in Appendix 3 simulation results), and this resulted in better results with little evidence of deleterious performance for other models. Can you comment on if this prior weighting should be used in place of equal weights?

The weighting is usually up to the modelers, and yes I do agree with the authors. I found the justification reasonably compelling.

3. Do you agree with the particular models selected for BMA or do you recommend a different set of models?

I agree with the ones selected, they reflect the wealth of experience.

4. Was adequate testing of the methods performed? In particular,

4.a. What additional steps, if any, are recommended to build confidence in the profile posterior density and Laplace approximations? Are any special situations evident, where these approximations may work particularly poorly?

In general, the Laplace approximation will fail when there are multiple modes in the posterior. I do not foresee this being the case often, but it could happen. It would be wise to think about this in more

[3] Possible uses of BMD point estimates include comparative and meta-analyses, common use of the ratio BMD [point estimate]/ BMDL ratio as in indication of the quality of the model results.

detail. For example, could multiple starting points be used in numerical optimization algorithms to help test if the same algorithm will converge to different points?

4.b. Is the Monte Carlo testing approach taken for the sensitivity analyses an appropriate tool for evaluating the method?

Yes.

4.c. Are the "templates" adequate or is additional testing recommended in order to evaluate other aspects of study design such as numbers per dose group, or dose spacing.

I would consider them adequate.

5. What output other than the BMD, BMDL, posterior weights and plots, would be necessary to provide enough information to users for the purposes of quantitative risk assessment? What fit statistics would be necessary to assess model/method performance/fit (e.g., global goodness-of-fit p-values, scaled residuals, posterior p-value)?

I would like to see credible bands around the model averages plotted. In addition, I would like to see the averages and their credible bands under different sets of priors, superimposed.

6. Is the USEPA proposal to implement this methodology with default priors reasonable in practice, given the likely user of BMDS who have limited familiarity with Bayesian methods? If yes,

6.a. How does this methodology compare with current methods, with regard to likelihood that non-statisticians will use it appropriately and accurately?

Model averaging is so ubiquitous nowadays (it's related to the majority of winning algorithms in Kaggle competitions for example), that I am optimistic about this.

6.b. What situations may be envisioned where default priors would be over-ridden, and what measures would help to make sure this is done appropriately (if it is needed)?

This is tricky. Maybe the default multiple prior set sensitivity analysis should always be present in the output. Additional priors can be added, but should always be juxtaposed against the default priors in the sensitivity analysis.

7. Additional Recommendations: Are there any additional aspects of software development and testing, or model documentation, or reporting of model results that give you special cause for concern? If so, please describe your concerns and recommendations.

Please see the additional comments at the end of this document.

III. SPECIFIC OBSERVATIONS

Page	Line	Comment or Question
2	44	The sentence here reads “the comparison of model averaging to other alternative methods such as nonparametric dose response modeling”. I would not consider non-parametric models to be an alternative to model averaging. They could easily be a part of the model set that is being averaged. So I would delete that part. Instead, the last challenge to mention could be the accuracy and feasibility of computation of the marginal likelihood weights required for the averaging.
2	47	“take” should be “make”. I would also add some justification for the need for the priors on models. Something like “given the valuable experiences learned from the past benchmark dose models”...
2	55	delete “are”
		Make it clear in the abstract that you are making the software too (as R package?)
3	1	It is not true that only parametric models can be averaged; current machine learning ensemble averaging methods average all sorts of things, from GAM to GP regressions to trees to networks. How about refocusing the sentence to pertain to benchmark dose modeling and estimating particular parameters rather than prediction? Even simply adding “Parametric” before “Model averaging is a technique...” would work.
3	1	Also, please replace the rest of this sentence with a more precise statement: instead “...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” would you consider writing something like “it estimates the predictor-response relationship as a weighted sum of individual models’ estimates of this relationship, and is one way to take model uncertainty in risk assessment into account.”
3	77/78	Perhaps add a qualifier after “model averaging”, something like “parametric model averaging”

Page	Line	Comment or Question
3	81	It is obvious that the results will (with probability 1) depend on the number and type of models included in model ensemble. Again, please slightly refocus this sentence/paragraph so that it doesn't read like a general intro to model averaging, but rather an intro to the model averaging practice in BMD modeling. I know it ought to be subsumed from the title and the purpose of this article, but it would be better to be precise in the introduction.
4	101	This sentence is a bit awkward: "...weight if models that are more parsimonious do not describe the data well and the data support them." - Could you rephrase it as "...weight if the data support them and the more parsimonious models do not describe the data as well."?
5	121	Can you rephrase "may result in misrepresentation of the true underlying dose-response relationship and significant model uncertainty. Bayesian Model averaging develops a probabilistic framework to incorporates inference from the models considered." as follows: "may result in misrepresentation of the true underlying dose-response relationship and significantly understate model uncertainty. Bayesian Model averaging in contrast develops a probabilistic framework to incorporates inference from all the models considered, while also taking into account the model uncertainty."
5	124	Add "BMDS" in "these same models"
5	128	Use "BMDS" instead of "EPA"?
5	130	These few sentences are incorrect as written currently: "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." – If the ML is truncated at 18, but the prior support is not upper bounded by 18, then in theory Bayes posterior estimates can be above 18 (depending on what the data say) no matter how small that prior probability of 18 and above is. How about adding a qualifier in the last sentence such as: "This will usually result in the Bayesian estimate of this parameter being smaller than its equivalent estimate made using maximum likelihood"?
6	148	You might not want to use both density and distribution interchangeably– I suggest sticking with "density" when using "g" – but do the search/replace in the manuscript whichever term you choose to stick with. Also, need to condition on data D on the left hand side of your eqn 1.
6	151	You may also want to reserve BMD and BMDL to denote the parameter, as you have done thus far in the manuscript. Then, when you want to talk about estimating, refer to BMD and BMDL estimates, eg: "The BMD and BMDL estimates are then computed based on this posterior distribution".
6	153	"the BMDL is taken as the 100*γth percentile for appropriately low confidence level γ" – can you get rid of "confidence" and replace with "probability"? (since it's not an inferential procedure, you don't want to use "credible" either).

Page	Line	Comment or Question
6	153	Same place - Perhaps also rephrase "appropriately low" with "appropriately chosen"?
6	153	Which posterior distribution described in (1)? There are two types there
6	154 and 155	You use the Laplace approximation twice in two successive but different sentences, giving the reader a bit of a double-take. If you use the Laplace approximation for both the marginal likelihood (\pi_k) and for the posterior (and do specify which ones, g_{k} or g_{ma}?), then say so in the same sentence. Or add "also" in the sentence in line 155.
7	169	I would use "marginal likelihood" instead of "integrated likelihood" here (integrated likelihood is generally reserved for nuisance parameter integration).
6-7	156-170	It's a bit odd that profile likelihood is brought into this discussion now. It seems that you are doing simple Laplace approximation and working with marginal densities (integrating out some of the parameters). It is very surprising to just be told about PPDs instead, without any formulas or motivation. Here are some suggestions: 1) How are using "profile posteriors" actually in your approach – are you maximizing the maximized conditional posteriors? Explain, and provide formulas. 2) Why is the profiling needed? The only motivation seems to be is to match up to an existing method (MAPL) – so please need to add more arguments here 3) Perhaps add a whole subsection or at least a paragraph explaining why profiling is needed and why it is better than marginalization (over some of the parameters) in your case. 4) Consider also making a separate section just for "relationship to the existing approaches".
		Suggested notation changes: 1) don't use the same \pi in lines 117-118 and later in the manuscript; maybe use "p" instead of \pi in lines 117-118 2) don't use g for both prior and posterior. Call the priors f or something else, and change that on lines 177, 179, 182, and 183
		Eq 2 and the line below: 1) please put a hat on the inverse Hessian, and say it is evaluated at the MAP \hat \theta 2) Then, swap D and theta in the likelihood l(), so that it corresponds to the definition of the likelihood (parameters given Data). 3) Change on line 178 "the likelihood of the model ** parameters** given the data D." 4) Also, add "evaluated at MAP" after "the prior density for θ_k"
	186	Line 186: since you have MAPs in multiple factors of I_k, is this still technically a "weighted average" of individual MAPs? Or is it just a combination of MAPs, in the absence of a more appropriate name?

Page	Line	Comment or Question
	187	Why is “This is equivalent to the median of the posterior distribution defined in equation (1)” true? Please add a sentence or two providing some guidance to the readers.
	390	“or when there is very little data exist to inform” – typo
	405	“run times depending convergence" – typo
	References	please capitalize proper nouns
		Specify in the Figure captions what the vertical bars represent

Walter W. Piegorsch, Ph.D.
University of Arizona

Peer Review Comments on *Quantitative risk assessment: Developing a complete Bayesian approach to Dichotomous Dose-Response Model Averaging*

Reviewer Walter W. Piegorsch

I. GENERAL IMPRESSIONS

The provided manuscript is well-written. It carefully guides the reader though the complex issues of model adequacy and in particular focuses on a proposed approach using Bayesian model averaging (BMA) for benchmark dose (BMD) and benchmark lower limit (BMDL) calculations. Overall, I strongly support the Agency's desire to advance model averaging (as BMA or as frequentist model averaging, FMA) over current standards in calculating BMDs and especially BMDLs in quantitative risk assessment. I have a few quibbles with some of the suggestions in the manuscript, which will be explicated in my comments below. Past these, however, I support further development of BMA (and FMA) approaches for BMDs and BMDLs, and heavily encourage the Agency to adopt the kinds of methods proposed here into its next version of BMDS.

II. RESPONSE TO CHARGE QUESTIONS

1. Are the documentation materials describing the proposed method, including the description of the principles and advantages of BMA in general, accurate and clear?

Yes, the documented materials appear to be accurate and clear.

2. Are the methods described adequate for the derivation of BMDLs that are reasonable for use as points of departure for use in EPA risk assessments? In particular, with respect to:

> ***2.a. Use of approximations such as profile posterior density (PPD, a Bayesian analogue of profile likelihood) for model-specific posterior BMD distributions, and Laplace approximation for integrated likelihood (marginal density of the data).***
>
> I think use of a PPD as described is a reasonable component of a POD calculation, if the hierarchical model is properly constructed.
>
> ***2.b. The possibility of having more parameters than dose groups in a given application, for a single model. The EPA states in the report provided that incorporation of prior information for model parameters allows application to data with fewer dose groups than parameters[1]. In current dose-response modeling practice, EPA does not use a model when the number of parameters exceeds the number of dose groups.***

[1] The EPA is aware of general Bayesian literature which suggests that informative priors can address identifiability issues.

2.b.i. Do the reviewers agree with EPA that the proposed BMA methodology is reasonable to use, when the number of parameters exceeds the number of dose groups for individual models?

Mathematically the proposal is correct, but I am hesitant to give *carte blanche* approval of the strategy. No matter the model and/or parameters, a design with small numbers of dose groups represents a limited source of information upon which to estimate a dose-response relationship, and especially from this a BMD or BMDL; the recent article by Ringblom et al. (2018) builds on earlier works (e.g., Wignall et al., 2011) to emphasize this underlying concern.

It is nonetheless true, as suggested by footnote 1 herein, that when faced with a limited number of doses Bayesian methods can add information to the model hierarchy via *carefully constructed* prior distributions.

2.b.ii. Is additional research suggested, e.g., for some cases that may be problematic?

Certainly more research is needed to understand how to *carefully construct* priors (cf. 2.b.i above) so that they add pertinent and reasonable information, especially when they are being used to supplement a lack of dose-response information, due, e.g., to small numbers of doses. My (self-admitted) favorite example is my own work with Fang et al. (2015).

2.b.iii. Related to (i), the total number of parameters combined for all models is expected to frequently exceed the number of dose groups (typically 3-5). The EPA has concluded that if the approach proposed for individual models is reasonable, so are the BMA results. Will the large number of fitted parameters result in "overfitting"?[2]

Overfitting is a reasonable concern and one that should not be understated. Only more, careful research into the operating characteristics of the multi-parameter models will give better guidance on how much such overfitting will affect practical outcomes.

2.c. The approach for the derivation of BMD point estimators.[3] The currently-proposed estimator is the weighted average of the maximum a-posteriori point estimates from individual models, weighting by model posterior weights. Do the reviewers suggest one or more alternative or additional BMD point estimators for the model averaging context?

The suggested approach corresponds roughly to previous suggestions throughout the literature (including my own work in Simmons et al., 2013, and Fang et al. 2015), so I am predisposed to argue that these are fairly well-accepted. I suppose one could employ instead some form of frequentist-based weights using information-based quantities along the lines of

[2] This term is sometimes used to indicate that a model is very flexible resulting in a relatively complicated fit with features that may not be repeatable.

[3] Possible uses of BMD point estimates include comparative and meta-analyses, common use of the ratio BMD [point estimate]/ BMDL ratio as in indication of the quality of the model results.

$w_k = \exp(-AIC_k/2)/\sum_k \exp(-AIC_k/2)$ (see, e.g., Piegorsch et al., 2013) or using the BIC in place of the AIC, etc. (Wheeler and Bailer, 2008).

2.d. The proposed default model parameter priors defined in the draft manuscript (described as "Prior 1" in Appendix 3 simulation results). The "Diffuse Hill" condition described in Appendix 3 of the draft manuscript support material uses the same prior as the proposed approach, except the hill model's prior is more diffuse. Can you comment on if this prior is preferable to the proposed "Prior 1" set of priors?

Although I have employed diffuse priors in my own work, I have become more and more wary of highly diffuse priors: they can lead to posteriors that do unexpected things, and more often than not do not represent the sort of "objective" or "non-informative" information for which they are typically employed. My own experience with the Hill prior is limited, however, and so my call here would be for further research into its use for the specific problem of BMA BMD inferences. The simulation study in the draft manuscript is a useful step towards this goal.

2.e. The use of equal model weights (described as "Even" model weighting in Appendix 3 simulation results). As discussed in Section 4 of the draft manuscript, in an effort to account for problematic conditions in the literature, we increased the quantal linear default weighting ("MAQ approach"; described as "QL = 0.5" model weighting in Appendix 3 simulation results), and this resulted in better results with little evidence of deleterious performance for other models. Can you comment on if this prior weighting should be used in place of equal weights?

Perhaps–it seemed to be acceptable in the manuscript, but here again I think more research is needed to study its (and other non-homogeneous weightings') performance in the BMA BMD setting. (Also see Question 3, below.)

3. Do you agree with the particular models selected for BMA or do you recommend a different set of models?

I often like to include the quantal-quadratic (i.e., the Multistage with β_1 set to zero) to try and provide alternative flexibility above and beyond the quantal-linear form. I realize that the Multistage incorporates within itself this particular sub-model, but I find it useful to compare posterior model probabilities from all three versions (along with the Weibull) to gauge how this general structure of dose-response model is doing with the given data set. Of course, when doing so one must keep in mind the warnings about nested models and possible misleading posterior information given by Wheeler & Bailer (2009, Sec. 2).

4. Was adequate testing of the methods performed? In particular,

4.a. What additional steps, if any, are recommended to build confidence in the profile posterior density and Laplace approximations? Are any special situations evident, where these approximations may work particularly poorly?

The work in the current manuscript is a good first step. However, I would encourage further study of these techniques as applied in BMD and BMDL calculations, using more extensive Monte Carlo studies and other forms of environmental dose-response data (see Question 4.b).

4.b. Is the Monte Carlo testing approach taken for the sensitivity analyses an appropriate tool for evaluating the method?

Yes, it is. Of course, further application to a variety of data sets (perhaps using a large, established, public-access database of pertinent dose-response studies, as was discussed during the teleconference call in June 2018) would provide additional guidance on any of the various methods described here.

4.c. Are the "templates" adequate or is additional testing recommended in order to evaluate other aspects of study design such as numbers per dose group, or dose spacing.

The question's suggestion to study in more detail numbers of dose groups and dose spacing is, I think, prescient: as noted in Question 2.b.i above, ongoing research is calling into question use of low numbers of dose groups when BMD estimation is a critical goal of a study (Ringbloom et al., 2018; Wignall et al., 2011).

5. What output other than the BMD, BMDL, posterior weights and plots, would be necessary to provide enough information to users for the purposes of quantitative risk assessment? What fit statistics would be necessary to assess model/method performance/fit (e.g., global goodness-of-fit p-values, scaled residuals, posterior p-value)?

I think the posterior plot should be a fundamental component of any Bayesian output, and I applaud that suggestion. Personally, I would like to see a posterior plot of the BMD (which may be difficult to produce in some settings) for use in visually explicating the amount of uncertainty embedded in the BMD and BMDL posterior calculations. We included this sort of approach in Simmons et al. (2013) and I now wish we had done so in Fang et al. (2015).

6. Is the USEPA proposal to implement this methodology with default priors reasonable in practice, given the likely user of BMDS who have limited familiarity with Bayesian methods? If yes,

6.a. How does this methodology compare with current methods, with regard to likelihood that non-statisticians will use it appropriately and accurately?

A good question, and a tough one to answer. First, I think some form of model averaging is fundamentally necessary when calculation BMDs and BMDLs from data, as we are learning that single-model use, alone or in concert with a model selection effort, can fail miserably to create a useable BMDL (West et al., 2012; Ringblom et al., 2014). Whether this is a BMA as seen herein or an FMA (Piegorsch, et al., 2013) is up to the user.

Following on this, if the user if conversant with Bayesian analysis and posterior interpretation, I think it is worth promulgating a carefully-constructed and well-studied BMA methodology. And even if not – to answer the question – I find myself comfortable with offering the sort of BMA approach seen herein to a lesser-trained community. I expect that waiting for Bayesian methods to spread through a non-statistical user community may take longer than our current life expectancies, unless of course we offer it energetically and with as much background 'educational' material as possible. I am optimistic that misuse of the methods will diminish as the larger scientific body becomes more familiar with the Bayesian paradigm.

6.b. What situations may be envisioned where default priors would be over-ridden, and what measures would help to make sure this is done appropriately (if it is needed)?

The default priors could obviously be overridden when the analyst has strong prior information of her/his own, based on prior experience with use of these models on pertinent data, or from extensive vetting of various priors on, say, a large, established, public-access database of pertinent dose-response studies (as in Question 4.b).

As for measures to ensure this is done appropriately, more research is needed to develop posterior measures that might 'red flag' a poor choice of prior(s) and/or other substandard posterior model fits as part of a standard BMA/BMD software package. I am sorry to report that I do not have useful suggestions towards this goal at present, but I do agree that developments along these lines would be propitious additions to future versions of BMDS that incorporate BMA calculations.

7. Additional Recommendations: Are there any additional aspects of software development and testing, or model documentation, or reporting of model results that give you special cause for concern? If so, please describe your concerns and recommendations.

No further comments.

III. SPECIFIC OBSERVATIONS

Page	Line	Comment or Question
3	70	I suggest including a refr. to West et al. 2012 after "properties".
3	73	Is there a good reference to support "...adopted as standard risk assessment practice"?
4	90	Refer the reader to Table 1 after mentioning the multistage model.
5	122	I suggest including a refr. to West et al. 2012 after "uncertainty".
6	148	It would be useful to include an equation for a "posterior distribution", or perhaps referring the reader to a new Appendix section that briefly explains basic features of the Bayesian prior/likelihood/posterior hierarchy.
6	154	This line confused me. Should "γ" be "1-γ"? And, is the sentence staring with "Model weights..." necessary?
7	172	Add ")"after "(5)".
8	192	"implements"
9	204	"constrained models"
9	208	Is the BMD here based on BMA calculations?
9	212	Re."...indicating the model". Which model is this? Log-logistic? Or are the BMA calculation being called upon?
10	235	"which bounds"
11	248	Delete the double-quote character " at the end of the line.
12	272 & 274	Doesn't this multistage model have 4 parameters?
14	329	Add "(NP)" after "non-parametric".
16	383	"uses the same"
17	390	Appropriate priors can be developed in certain situations: see Fang et al. 2015.
Appdx. 1	3	What is "Hsu (1)"?
Appdx. 1	7	"Hessian"
Appdx. 1	8	Remove the semi-colon and end the sentence in a period.
Appdx. 1	14	To what "expansion" are you referring?
Appdx. 1	-3	...in figure SA1-1

Additional References

Fang, Q., Piegorsch, W. W., Simmons, S. J., Li, X., Chen, C., and Wang, Y. (2015). Bayesian model-averaged benchmark dose analysis via reparameterized quantal-response models. *Biometrics* **71**, 1168-1175.

Piegorsch, W. W., An, L., Wickens, A. A., West, R. W., Peña, E. A., and Wu, W. (2013). Information-theoretic model-averaged benchmark dose analysis in environmental risk assessment. *Environmetrics* **24**, 143-157.

Ringblom, J., Johanson, G., and Öberg, M. (2014). Current modeling practice may lead to falsely high benchmark dose estimates. *Regulatory Toxicology and Pharmacology* **69**, 171-177.

Ringblom, J., Kalantari, F., Gunnar, J., and Öberg, M. (2018). Influence of distribution of animals between dose groups on estimated benchmark dose and animal welfare for continuous effects. *Risk Analysis* **38**, 1143-1153.

Wignall, J. A., Shapiro, A. J., Wright, F. A., Woodruff, T. J., Chiu, W. A., Guyton, K. Z., and Rusyn, I. (2011). Standardizing benchmark dose calculations to improve science-based decisions in human health assessments. *Environmental Health Perspectives* **122**, 499-505.

West, R. W., Piegorsch, W. W., Peña, E. A., An, L., Wu, W., Wickens, A. A., Xiong, H., and Chen, W. (2012). The impact of model uncertainty on benchmark dose estimation. *Environmetrics* **23**, 706-716.

Wheeler, M. W., and Bailer, A. J. (2008). Model averaging software for dichotomous dose response risk estimation. *Journal of Statistical Software* **26**, Art. No. 5.

Wout Slob, Ph.D.
National Institute for Public Health and Environment (RIVM)
The Netherland

Peer Review Comments on *Quantitative risk assessment: Developing a complete Bayesian approach to Dichotomous Dose-Response Model Averaging*

Reviewer Wout Slob

I. GENERAL IMPRESSIONS

The statistical method described here appears to be theoretically sound, although I could not fully judge that based on the current description of the method, which was not entirely clear. This should be improved in a next version in my view. I will therefore leave this aspect to the other reviewers, who may better understand the method despite the unclear description. Assuming that the method is theoretically valid, I think it is a highly valuable method as it combines using prior information with computational speed. The first is of interest as it will be possible in the future to collect informative priors from historical data, and the latter is a great practical advantage, as it will make the use of the BMD approach, in particular with more complicated or composite data, much easier.

My recommendation would be to put the paper in another and broader context. Currently, the target question addressed is rather narrow, and formulated from the usual statistical point of view: How well does the proposed method perform relative to other methods, where performance is defined as coverage of the BMDL. I think this is not really the question of interest. I would argue that the use of the approximate method has another justification, and this could be worded in the introduction roughly as follows.

First, it might be stated that BMD experts tend to agree that model-averaging is the appropriate way of doing a BMD analysis. Then, it could be said that a Bayesian approach for MA is favorable over the current ML method. The main reason is that it can use prior information on the model parameters based on historical data, for which we now know that they will result in pretty narrow (= informative) distributions (in particular for the shape parameters). Therefore, it may be expected that using this prior information will result in more precise BMD estimates (smaller CIs), while this (Bayesian) approach will at the same time solve problems associated with poor datasets in many cases (both because of the Bayesian approach as such, and because of the informative priors).

Then, it could be said that both the non-Bayesian and the Bayesian approach of MA is rather time-consuming, which may not be a major problem in doing in an analysis of a single dataset, but in more elaborate analyses (multiple endpoints in a single run, or high-throughput data) this constitutes a practical burden. Then state that this was addressed by developing the proposed method, and that simulation studies will be performed to evaluate if its performance is reasonable, or at least close to other, more computation-intensive methods. Also mention some additional advantages (e.g., more stable in case of poor data, and use of more parameters than doses, although I think this similarly holds for MCMC).

So, briefly, the first important thing is that it can use prior information (just like MCMC) but the great thing about this approximate method is that it is so quick. So, all you need to show is that its performance is acceptable (or at least close to that of MCMC).

The second thing I would like to stress is that the BMDL is of course a key parameter in dose-response assessment, but not the only one. The BMDU is also of great interest, as it provides very important information both to the party that performed the study ("industry") and to the risk manager. The BMDU will tell the former by how much the BMDL might have been higher with a better study than the current one. For instance, a poor study design may result in a very low BMDL but with a BMDU that is orders of magnitude higher. This tells the study director that a better study design might have resulted in a substantially higher BMDL. This may prompt the study director to use a better study design in the future, even when that is more expensive, in particular when a higher BMDL has economic benefits. For the risk manager it is also important to know the BMDU. For instance, without knowing the BMDU a compound with a low BMDL will be easily judged as a potent compound. However, when it is known that the BMDU is very high, this will tell the risk manager that is might not be that potent at all. When the BMDU is not much higher than the BMDL, however, the risk manager can be pretty sure that the compound is indeed potent. For these reasons it is paramount to also know the coverage of the BMDU, and in my view this cannot be omitted from the evaluation of the simulations.

Another general impression is that the paper takes the current state of affairs, where many of the datasets generated do not provide much information, for granted. The position that this paper seems to take is: Let's try to "save" those datasets to the extent possible, by developing a method that results in a reasonable BMDL whatever the quality of the study design or study protocol. The danger is that this will seduce study directors to pay less attention to performing a high quality study, and that the quality of studies will not improve in the future. Now, it might be argued that poor study designs will result in a lower BMDL, but still with the appropriate coverage, as shown by the simulations. The problem however is that the simulations only take random sampling error into account, while in reality (in particular with less meticulous study directors) the response in a particular dose group may be affected by other experimental factors (e.g. confounding factors, or just uncontrolled incidences) that result in nonrandom errors. Those can result in misleading (biased) BMD estimates. With better study designs (more doses, e.g.) such nonrandom errors are easier to detect, and have less impact on the result.

The focus on coverage is understandable from a statistical point of view, but from a RA point of view it is more relevant to know by how much the BMDL from the approximate method would differ from the BMDL from MCMC or ML. For example, suppose that nominal BMDL coverage is 95%, the approximate method results in 90%, while MCMC results in 95%, the question is: How bad is that? Is that reason to reject the approximate method? The answer can only be given by considering the difference in BMDLs. When the BMDLs only differ by 30%, then every risk assessor would (or should) be happy to use the approximate method, given all the other uncertainties in RA. After all, when risk assessors apply 10x10 to the BMDL, they have now idea what the exact coverage of those two assessment factors actually is, let alone what the coverage of the final RfD might be.

Therefore, I think that some other output parameters would be needed in the final tables in the annex, at least BMDU coverage. The BMD/BMDL ratio should either be replaced by the BMDU/BMDL ratio, or just omitted. Comparing BMD precision is of interest when comparing study designs, but not so much among methods. Further, instead of considering the bias in the BMD estimate itself, it would be more interesting to consider the "bias" of the complete BMD confidence interval, by counting the number of times the true BMD falls outside the interval on the left or the right side. Finally, it might be considered to report the median BMDL and median BMDU for each method. This would indicate by how much BMDLs would differ when using one or the other method. As just said, this is a more important criterion from a RA point of view.

A real problem is a resulting coverage of 100%. This is an undetermined result. A resulting coverage of 100% may reflect that all BMDLs are just below the true BMD, or that most of them are miles away. So, a resulting 100% coverage can be related to a pretty good, or to a totally inadequate method (e.g. divide the LOAEL by a million). By the way, this makes the average coverage reported in the tables of the annex rather meaningless.

The simulations are based on a very large number of test conditions, i.e., true DR relationships. However, as the plots show, the variety of shapes is limited. Some curves have two point of inflection, which is not realistic, and these curves should be omitted, as they will give irrelevant information about method performance (we don't want to select a method that performs better in unrealistic situations). Some curves are nearly flat, which is unrealistic as well: we know from real data that the ratio ED50 to BMD10, say, tends to be around a factor of three, while these flat curves would probably show an ED50/BMD10 ratio of various orders of magnitude. These curves should be omitted as well. The remaining curves have similar shapes, but they do differ in one important aspect: the BMD. Given that the applied doses in the study design are fixed, the by far most important difference in test conditions is the fact that the applied doses have different locations relative to the BMD. So, the test conditions in fact represent different study designs (relative dose location) rather than different true models. When this is recognized the interpretation of the simulation results will be totally different. For instance, there is a clear correlation between coverage and value of the true BMD, which is explained by differences in (relative) dose location. Therefore, all the text that relates to interpreting the simulation results needs to be rewritten in my view.

Apart from this issue, it is not possible to see how the authors come to their conclusions formulated in the text based on the data in the tables. The conclusions are simply put on the table, by saying: "as the tables show". Unfortunately, the tables do not show very much to me (in particular those in the annex), apart from the fact that roughly speaking coverages do not really seem to correlate by method very much. It might be that they do, but normal human beings are unlikely to see it. Maybe this becomes (better) visible when the results are somehow transferred into graphics.

Briefly, the simulation results could (and should) be much reduced by omitting unrealistic DR relationships, and by focusing on those that might be realistic. Next, it should be made clear that the different test conditions mainly relate to different study designs, with different dose locations relative to the true BMD, with possibly some impact from differences in tested shapes (which are only mild anyway – which is OK, as they are in real life as well).

II. RESPONSE TO CHARGE QUESTIONS

1. Are the documentation materials describing the proposed method, including the description of the principles and advantages of BMA in general, accurate and clear?

Roughly, yes, but there are small unclarities or missing information at many places. However, an important part that is not clearly described is the approximation of the posterior (model-specific) distribution. I miss a section 2.3 describing the main idea of how this approximation is accomplished (in conceptual rather than in mathematical terms, to the extent possible). Without that explanation very few people will understand the method.

2. Are the methods described adequate for the derivation of BMDLs that are reasonable for use as points of departure for use in EPA risk assessments? In particular, with respect to:

2.a. Use of approximations such as profile posterior density (PPD, a Bayesian analogue of profile likelihood) for model-specific posterior BMD distributions, and Laplace approximation for integrated likelihood (marginal density of the data).

As already mentioned, this can hardly be judged based on the current description of the method.

2.b. The possibility of having more parameters than dose groups in a given application, for a single model. The EPA states in the report provided that incorporation of prior information for model parameters allows application to data with fewer dose groups than parameters[1]. In current dose-response modeling practice, EPA does not use a model when the number of parameters exceeds the number of dose groups.

2.b.i. Do the reviewers agree with EPA that the proposed BMA methodology is reasonable to use, when the number of parameters exceeds the number of dose groups for individual models?

Yes, but this equally holds for the “full” Bayesian approach (MCMC).

2.b.ii. Is additional research suggested, e.g., for some cases that may be problematic?

I already suggested to report some other parameters from the simulations. Further, it would be very helpful to examine the potential impact of realistic priors on the width of the BMD CI, in combination with realistic test conditions, i.e. using models that are realistic (which appear to describe informative dose-response data well; note that the LMS model is often rejected by informative DR data). This would be a strong stimulus to put effort in further building historical databases.

[1] The EPA is aware of general Bayesian literature which suggests that informative priors can address identifiability issues.

2.b.iii. Related to (i), the total number of parameters combined for all models is expected to frequently exceed the number of dose groups (typically 3-5). The EPA has concluded that if the approach proposed for individual models is reasonable, so are the BMA results. Will the large number of fitted parameters result in "overfitting"?[2]

In my view overfitting means that the likelihood does not have a unique optimum. This is a problem in ML theory, but not necessarily in Bayesian fitting. I am not sure if this would pose a problem in the theoretical background of the approximate method, as I do not yet fully understand it.

2.c. The approach for the derivation of BMD point estimators.[3] The currently-proposed estimator is the weighted average of the maximum a-posteriori point estimates from individual models, weighting by model posterior weights. Do the reviewers suggest one or more alternative or additional BMD point estimators for the model averaging context?

The point estimate of the BMD is not of interest to begin with. We don't use it in risk assessment. We only use BMDL as the POD, and the BMDU for the purposes discussed above. As a matter of fact, I think statisticians should educate risk assessors (or other scientists) to forget about point estimates, as most scientists tend to draw conclusions from them that are not justified, or simply wrong. Comparative analyses can only be done based on the complete BMD confidence interval, without that it is not clear if two values differ due to noise or to a real difference. Similarly, for meta-analyses. The use of the BMD/BMDL ratio is a statistical mistake (unless the interval is always symmetric on log-scale, but that is not generally true). Briefly, the point estimate of the BMD is not needed, it only gives rise to wrong conclusions and practices.

2.d. The proposed default model parameter priors defined in the draft manuscript (described as "Prior 1" in Appendix 3 simulation results). The "Diffuse Hill" condition described in Appendix 3 of the draft manuscript support material uses the same prior as the proposed approach, except the hill model's prior is more diffuse. Can you comment on if this prior is preferable to the proposed "Prior 1" set of priors?

Obviously, BMDL coverage will be higher when increasing the variances of the priors. So, as long as the priors are not based on real data, a conservative approach seems obvious. However, as soon as historical data have been sufficiently analyzed, resulting in prior distributions, that information can be used to make more funded choices (or just use the empirical priors themselves).

[2] This term is sometimes used to indicate that a model is very flexible resulting in a relatively complicated fit with features that may not be repeatable.

[3] Possible uses of BMD point estimates include comparative and meta-analyses, common use of the ratio BMD [point estimate]/ BMDL ratio as in indication of the quality of the model results.

2.e. The use of equal model weights (described as "Even" model weighting in Appendix 3 simulation results). As discussed in Section 4 of the draft manuscript, in an effort to account for problematic conditions in the literature, we increased the quantal linear default weighting ("MAQ approach"; described as "QL = 0.5" model weighting in Appendix 3 simulation results), and this resulted in better results with little evidence of deleterious performance for other models. Can you comment on if this prior weighting should be used in place of equal weights?

We are working on analyzing quantal dose-response data, in a similar way as we did for continuous data (Slob and Setzer, 2014), and one of the things that comes out very clearly is that the quantal-linear model is not able to describe informative DR data. So, we know a priori, that its weight should be something like zero or very close to zero. Therefore, the MAQ approach would be very odd, as it contradicts real data. Furthermore, I do not see from the tables that it performs better than the other methods. Anyway, when basing the weights of the models on simulation results, thereby ignoring real prior information on model weights, the very principle of the Bayesian methodology is undermined. When there are specific problems with specific datasets, those should be solved in another way. So, I would not support the MAQ approach.

3. Do you agree with the particular models selected for BMA or do you recommend a different set of models?

Yes. Probably, the LMS model(s) should be omitted completely, given the empirical evidence that they do not describe informative DR data. In particular, the one-stage (quantal linear) model performs really badly. But also the three-stage model is found to rarely fit the data better than the two-stage model, so I think the three-stage (or higher stages) might be omitted as well. Similarly, the probit and logit models are unrealistic models (again, as shown by informative DR data) and should be omitted. So, when following the principles of Bayesian statistics, they should receive a very low prior weight (or zero, maybe).

4. Was adequate testing of the methods performed? In particular,

4.a. What additional steps, if any, are recommended to build confidence in the profile posterior density and Laplace approximations? Are any special situations evident, where these approximations may work particularly poorly?

See my other comments.

4.b. Is the Monte Carlo testing approach taken for the sensitivity analyses an appropriate tool for evaluating the method?

Yes, this is an appropriate method for validating a statistical method (sensitivity analysis is not the appropriate term here).

4.c. Are the "templates" adequate or is additional testing recommended in order to evaluate other aspects of study design such as numbers per dose group, or dose spacing.

It is a general disadvantage of computer simulations that they are only "locally valid". However, when the number of true models is reduced (see above) there would be room to add some simulations with other study designs. For example, one with multiple doses and larger group sizes (e.g. 50), to examine more ideal situations, even though they may not occur very often in practice. However, the goal is to validate the method, not to compare study designs, so an ideal situation and a more realistic situation may suffice.

5. What output other than the BMD, BMDL, posterior weights and plots, would be necessary to provide enough information to users for the purposes of quantitative risk assessment? What fit statistics would be necessary to assess model/method performance/fit (e.g., global goodness-of-fit p-values, scaled residuals, posterior p-value)?

BMDU is crucial in the BMDS output. I recommend *not* to report the BMD, as it will be misused. Reporting the BMDL and BMDU is all that is needed. People might protest against this, but once they are used to it, is will be no problem (actually, many problems will be solved). Posterior weights may be part of the output, although it is not essential. Posterior P-values may be of interest, as a diagnostic tool indicating that there may be a data problem, or a deficiency in the statistical part of the model (e.g., litter effects not taken into account). The guidance should however make clear that it is unlikely that the average model is unable to describe the true dose-response, and that it is most likely flexible enough to describe all true dose-response relationships that may occur.

Scaled residuals should not be reported, as they will be misused as well (for instance, as an argument to drop high doses, which is exactly what should not be done in general).

6. Is the USEPA proposal to implement this methodology with default priors reasonable in practice, given the likely user of BMDS who have limited familiarity with Bayesian methods? If yes,

6.a. How does this methodology compare with current methods, with regard to likelihood that non-statisticians will use it appropriately and accurately?

Non-statisticians cannot do much wrong, as long as the suite of models and all the priors are fixed.

6.b. What situations may be envisioned where default priors would be over-ridden, and what measures would help to make sure this is done appropriately (if it is needed)?

It would be a bad idea to let users choose their own priors. This is a matter that the experts behind the methodology need to find consensus on. For the time being we could decide on implementing either priors 1a or 1b. When we are at the stage that we have a more complete picture of priors derived from historical data, we should try to find consensus on the choice of the priors, and then implement those, as the only option. I would be opposed to the idea that users can have any choice on priors. In the field of risk assessment, that is no option, given the opposing interests that exist. The priors are part of the methodology, and should be fixed in the guidance documents and in the software.

7. Additional Recommendations: Are there any additional aspects of software development and testing, or model documentation, or reporting of model results that give you special cause for concern? If so, please describe your concerns and recommendations.

I probably have made most of my points.

III. SPECIFIC OBSERVATIONS

Page	Line	Comment or Question
		The title should include the word approximate, I think. It may also be helpful to have an acronym for the method, something like approximate Bayesian approach for model averaging (ABAMA - easy to remember).
4	Lines 93-94	Not only “our approach”, also MCMC solves this.
4	Lines 99-100	A prior probability of 1% that the shape parameter is smaller than one is based on an assumption about “supralinearity” (which is a theoretical fallacy). Real data sometimes show that the shape parameter can be smaller than one. The prior used here is not diffuse (relatively non-informative), but gives unwarranted prior information. This will not be revealed by the simulations, because all test conditions assume a “sublinear” true relationship. (However, I noted that M11 is a log-logistic model with shape parameter = 0.7, yet it is “sublinear” in figure 3.There must be an error here. It seems that the abbreviations in the tables and text do not correctly correspond to those with the curves in the figures, while some seem to be missing).
4	Lines 102-104	The implication that sublinear shapes are the (only) ones that are frequently seen in practice is not correct. This is only based on an assumption which involves an error of thinking. In reality, the shape parameter can be smaller than one.
5	Line 133	It would be clearer to state what are limited data, i.e., data with none or only one intermediate response, in this case.
5	Line 134	Sufficient data is not the pivot, it is sufficient information in the data (e.g., at least two intermediate responses).
7	Lines 166-167	This sentence is not clear. By what is the posterior distribution substituted?
8	Lines 185-186	It should be explained how the modes of the individual models translate into the median of the posterior. Or this should follow from the still missing section 2.3, as suggested earlier. Expression (3) Pr reads as probability, while it is a density

Page	Line	Comment or Question
9	Lines 205-206	It is ironic that the first example does not comply with the earlier statement that most dose-response shapes are sublinear. Anyway, it would be interesting to see what happens with the BMDL when the prior does allow that shape parameter to be lower than one, e.g., with probability 30%.
9	Line 209	What does (0.01) mean here? Or is this a typo?
9	Lines 212-213	Yes, this average model describes the data for the very reason that is has infinite slope at dose zero. So, you did not want to allow the individual models to be “supralinear”, but you do allow that their combination - the basis for the BMD estimate - is? That is silly.
11	Line 268	It needs to be made clear what those stochastic processes are, and how the curves were generated.
14	Line 321	Using the expectation of a ratio as a measure of bias is not appropriate. Instead, the expectation of the log(ratio) should be used. For instance, when two simulations result in 1 and 100, while the true BMD is 10, than the method is (so far) unbiased (because the first outcome was a factor of 10 too low, the other a factor of 10 too high). However, the expectation (arithmetic mean) of the two ratios (0.1 and 10) equals 5.05, which indicates a bias. The expectation of the log ratio however is 1, which equals 10 after back-transformation, i.e., indicating no bias, as it should.
14	Lines 329-331	I do not see that the proposed method and the NP method are similar, nor that they often result in > 90% coverage. For M1 coverages are 97.9 and 0%, contradicting both conclusions. Further, near nominal coverage is defined as > 90%. However, a coverage of 100% is not informative, as a coverage of 100% is also achieved by a method that results in extremely low BMDLs in all cases.
14	Lines 332-333	Again, it is unclear how these conclusions are drawn. The results in the tables are not self-evident at all. For example, it is concluded from the simulations that the approximate method performs better than the Shao method. If that correction were justified, then it is rather weird that an approximate method performs better than the method is claims to approximate. This needs explanation.
15	Lines 342-347	Here I am totally lost. Where can I find the results for MAQ in the Annex tables?
16	Line 365	Explain what “the ratio statistic” refers to.
16	Lines 366-367	BMDLs closer to the BMD means that the method results in more precise estimates, not in more stable estimates
17	Line 389	i.e. should be e.g.
24-25	Figures 1 and 2	In black-and-white print it is hard to see the individual curves. Maybe better use line width for indicating goodness of fit.

Page	Line	Comment or Question
26	Fig 3.	The models (or their abbreviations) in figure 3 do not match the models in table SA2-1. The way the models (test conditions) are described is highly confusing.
20-22	Tables 1-3	Please add the number of simulations performed, and nominal coverage. Also check if the abbreviated models correspond to those in Fig. 3.
	Fig. SA1-1	In this figure the mode and median are quite different, while in the approximate method they are equal by definition. Doesn't this imply that some aspect of the approximate method is suboptimal? For example, could it be that the normal approximation of the posterior is not entirely adequate, and that the approximation of then posterior could be improved by a lognormal approximation?
	Tables in annex.	The main problem of these tables is that they are practically unreadable. As many of these models result in very similar curves, it is not useful to report all those results, in particular when any differences in results are mainly driven by different locations of the applied doses relative to the true BMD. So, these tables can be greatly simplified, where the remaining models can be ordered according to the value of the true BMD. The headers of the tables are insufficiently informative. In the tables in the annex, an average is reported over all coverages. However, as soon as the list of coverages include 100%, the average is no longer meaningful. In all cases where the coverage is 100%, it needs to be reported by how much the BMDLs differed from the true BMD. The authors need to think how they can best do that.